Engineering Applications of Artificial Intelligence 72 (2018) 228–240

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Using social network analysis and gradient boosting to develop a soccer win–lose prediction model

Yoonjae Cho, Jaewoong Yoon, Sukjun Lee *

Business School, Kwangwoon University, 26 Kwangwoon-gil, Nowon-Gu, Seoul 139-701, South Korea

ARTICLE INFO

Keywords:
Social network analysis
Gradient boosting
Win–lose prediction
Soccer
Machine learning technique

ABSTRACT

We present the conceptual framework of a soccer win–lose prediction system (*SWLPS*) focused on passing distribution data (which is a representative characteristic of soccer) using social network analysis (SNA) and gradient boosting (GB). The general purpose of soccer predictions is to help the field supervisor design a strategy to win subsequent games using the derived information to improve and expand the coaching process. To implement and evaluate the proposed *SWLPS*, actual network indicators and predicted network indicators are generated using passing distribution data and SNA. The win–lose prediction is conducted using the GB machine learning technique. The performance of the *SWLPS* is analyzed through comparison with various machine learning techniques (i.e., support vector machine (SVM), neural network (NN), decision tree (DT), case-based reasoning (CBR), and logistic regression (LR)). The experimental results and analyses demonstrate that the network indicators generated through SNA can represent soccer team performance and that an accurate win–lose prediction system can be developed using GB technique.

1. Introduction

Social network analysis (SNA) has been widely used in various academic fields (e.g., sociology, psychology, anthropology, biology, and medicine). SNA focuses on the structure of ties within a set of social actors, such as persons, groups, organizations, and nations or the products of human activity. Thus, SNA is used to define the structure to be analyzed as a function of the relationship with the participant and to understand the structure of the network based on graph theory, algebra, statistical probability, and simulation (Galaskiewicz and Wasserman, 1993; Wasserman and Faust, 1994).

Nixon (1993) first applied the SNA approach to understand interactions, relationships, and structural changes in sports. This use of SNA was highly acclaimed because of its ability to quantify the attack type and the efficiency of the team. In recent years, many studies have been conducted on the application of SNA for sports, including soccer, basketball, handball, and hockey, in which the individual link structure is a pass. In particular, soccer (one of the most popular sports and the most representative team sport) can be considered a complex and dynamic system that evolves based on the interactions of multiple factors (Gréhaigne et al., 1997). In soccer, the most consequential form of interaction is clearly defined as the pass. In particular, the setting of soccer allows for a direct assessment of the interactions among team members. One of the advantages of investigating soccer is that the boundaries of the teams and the possible interactions of team members are clearly defined (Grund, 2012). In past research on soccer games, it has been considered challenging to predict victory or to quantify the contents of a game since the result of the competition can be changed due to complex factors (e.g., psychological, physiological, and environmental factors) affecting the player (Hughes and Franks, 2004). In addition, since the performance evaluation is determined by the supervisor and the subjective evaluation criteria of the coach, it is necessary to evaluate the nature of the competition through scientific analysis. The general purpose of sports predictions is to help the field supervisor design a strategy to win subsequent games based on the team performance evaluation. In addition, the derived information contributes to improve and expand the coaching process (Gonçalves et al., 2017). Thus, a match result prediction system must be built that considers the characteristics of the sports event and the team's performance.

Here, we propose a conceptual framework for a soccer win–lose prediction system (*SWLPS*) based on SNA and the gradient boosting (GB) machine learning technique; the system is designed for win–lose prediction specifically for Champions League (CL) soccer games. SNA, a method for analyzing organizational performance (Cross et al., 2002), is used to extract network indicators from passing distribution data to

* Corresponding author.
E-mail addresses: yoonjae@kw.ac.kr (Y. Cho), yjw8860@kw.ac.kr (J. Yoon), sjlee@kw.ac.kr (S. Lee).

https://doi.org/10.1016/j.engappai.2018.04.010
Received 13 January 2018; Received in revised form 29 March 2018; Accepted 9 April 2018
Available online 8 May 2018

evaluate how the team's performance affects the likelihood of winning. In the system, network indicators replace data obtained from notational analysis, which has often been used as input in previous research. In fact, passing distribution data can also belong to the category of data obtained via the notational analysis method. However, network indicators are different in that the passing distribution simply means the number of passes, but network indicators measure a team's performance. Thus, network indicators are more suitable than data from notational analysis for prediction of the outcome of a game. Additionally, the system employs GB, which is a family of powerful machine-learning techniques that have had considerable success in a wide range of practical applications (Natekin and Knoll, 2013). In particular, GB performs well in settings in which the number of variables exceeds the number of samples (high-dimensional data) (Lusa, 2017). The number of data points (samples) in this study is insufficient because the game prediction is performed for each season and round. As a result, the system adopting GB yielded excellent performance in game prediction.

We also implement an analysis of variance (ANOVA) analysis to identify the performance of the *SWLPS* through comparison of the performance of various classifiers (i.e., support vector machine (SVM), neural network (NN), decision tree (DT), case-based reasoning (CBR), and logistic regression (LR)) in terms of machine learning techniques for result predictions.

There are no studies related to game result prediction that combine SNA and machine learning. Previous studies in the sports field have usually predicted games using notational analysis data. Among the components of our system that predict the results of future games, there are two important elements that determine the performance of the system. The first element is the generation of appropriate input variables that are able to represent the performance of each team. The second is the compatibility of the classifier that learns the input and predicts the result.

The remainder of this paper is organized as follows. Section 2 briefly introduces the analysis method of sports performance, and Section 3 presents the construction procedure for the *SWLPS*. Section 4 presents the results of an empirical study performed to verify the performance of the *SWLPS*. Finally, conclusions are presented in Section 5.

2. Related work and limitations

Methods for sports performance analysis can be divided into notational analysis, SNA, and result prediction. The details are as follows.

2.1. Notational analysis

Sports performance analysis has been conducted primarily to analyze performance indicators generated through notational analysis, which was first proposed by Charles Reep in 1950 (Pollard, 2002) for football. Notational analysis is a technique for producing a permanent record of the events pertaining to a sporting event and is widely used by sports teams and individuals of various standards (James, 2006). One of the earliest empirical notational analyses on sports was conducted at Reilly and Thomas (1976). They analyzed the number of shots relative to goals scored and considered every move during a soccer game, including the intensity and duration of actions. Thus, they analyzed the computerized notation in terms of evolution. Hughes and Franks (2005) examined the length of passing sequences, the number of passes and team performance using notational analysis, and they found that longer passing sequences produced more goals per possession than shorter passing sequences for successful teams. Lago and Martín (2007) investigated ball possession in soccer and found that determinants affecting possession included the match status (e.g., winning, losing, or drawing), team status (e.g., home team or away team) and style of play. Lago-Ballesteros and Lago-Peñas (2010) analyzed the performance of soccer teams and found specific performance indicators that could be used to discriminate the top teams from the others. They also presented parameters to be used as normative data to collectively design and evaluate practices and competitions to establish peak-performance soccer teams.

2.2. SNA

SNA, which is not a formal theory in sociology but rather a strategy for investigating social structures (Otte and Rousseau, 2002), is a potentially useful method for sports performance analysis. Recently, SNA has received more attention than traditional notational analysis in this area. One of the earliest studies that applied SNA to team sports was conducted by Nixon (1993), who concluded that SNA can provide important insights into the leadership structure of sports teams. However, the few studies that have used SNA in sports settings have focused only on the cognitive or actual interaction between the players during the game (Bourbousson et al., 2010; Cotta et al., 2013; Passos et al., 2011). Grund (2012) studied the issue of within-team network structures and the performance of teams through an analysis of panel data. They found that networks characterized by high intensity and low centralization are indeed associated with better team performance. Pena and Touchette (2012) analyzed the strategy of soccer teams using network theory. They proposed an analysis method to discover the play patterns of each team, including hot spots, potential weaknesses, and the relative importance of each player in games. Cotta et al. (2013) investigated the use of simple graphs and network metrics to analyze the performance and play styles of the Spanish national football team in the World Cup 2010 and explained the results obtained at the complex network level. Fransen et al. (2015) used SNA to provide insight into the leadership structures within sports teams and found that SNA is a valuable tool in this regard. Clemente et al. (2015a) proposed a set of network methods to measure the specific properties of a team and found that network metrics can be a powerful tool to help coaches understand a team's specific properties and support their decision-making to improve the sports training process. Clemente et al. (2015b) analyzed team members' cooperation in basketball using centrality metrics of networks. They found that the specific point guard position is the most prominent position and that SNA is a useful approach to identify the patterns of interactions in basketball.

2.3. Result prediction

In the field of sports, there have been many studies on predicting the outcome of the game and analyzing performance. Koning (2000) and Koning et al. (2003) developed a model that used little prior knowledge and information and was heavily based on pure statistical models, such as ordered probit and Poisson models. Based on the models, they calculated the probability of winning for each team and predicted the most likely winner of a tournament. Rotshtein et al. (2005) developed a model for predicting the result of a football match. In that study, they analyzed the previous results of both teams and tuned fuzzy rules using genetic and neural optimization techniques. Huang and Chang (2010) developed a soccer prediction model using the multilayer perceptron, the backpropagation learning rule and the relative ratio values transformed from game records to be used as input data. The accuracy of the developed model was 76.9%, excluding draws. Halicioğlu (2011) developed a ranking system using the seasonal coefficients of variations of the end of season points and predicted the winner of the league. Snyder (2013) developed a soccer prediction model aimed at betting strategies. They used statistical methods (e.g., the Poisson distribution and multinomial regression), game records (e.g., frequency counts of events and statistical game records), and betting odds offered by various bookmakers and presented the approximate optimal betting strategy for use in simultaneous betting on multiple games with mutually exclusive outcomes.

2.4. Limitations of previous research

These previous studies have suffered from two limitations. First, the studies that have used SNA have analyzed predominantly team performances, strategies, or key players. In addition, they have identified

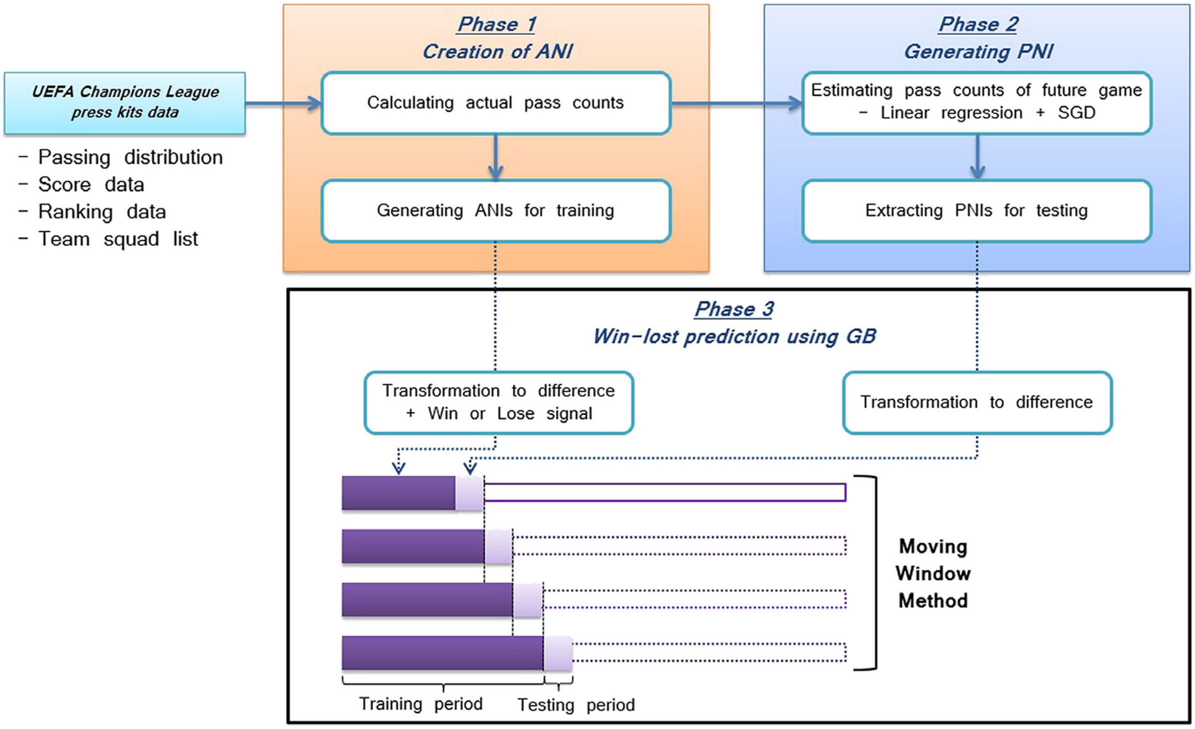

UEFA Champions League press kits data
- Passing distribution
- Score data
- Ranking data
- Team squad list
Phase 1
Creation of ANI
Calculating actual pass counts
Generating ANIs for training
Phase 2
Generating PNI
Estimating pass counts of future game – Linear regression + SGD
Extracting PNIs for testing
Phase 3
Win–lost prediction using GB
Transformation to difference + Win or Lose signal
Transformation to difference
Moving Window Method
Training period
Testing period

Fig. 1. The construction procedure of the proposed *SWLPS*.

relationships between specific network indicators and performance or the play style of a team. However, the benefits of SNA include the ability to evaluate the performance of individuals, groups, or the entire social network (Abbasi and Altmann, 2010). Furthermore, there have been no studies on win–lose prediction involving performance comparison using indicator values extracted through SNA. Second, several studies on result predictions using prior data on each team depended on the raw data obtained through notational analysis. Under this approach, team performance that is not described in the data itself cannot be quantified, and the processing of the data is required to overcome this limitation. Thus, we propose a prediction system to forecast the result of a future game using estimated indicators of social networks as input variables. The original purpose of SNA was to analyze the performance of an organization based on the relationships among its members (Warner et al., 2012; Cioffi-Revilla and O'Brien, 2007). Relationships can be defined in various manners, such as the amount of activity between members of Facebook, the number of communications, or the amount of activity between departments in a company (Kim and Hastak, 2018; Wolf et al., 2009; Griffin and Hauser, 1992). In the same context, a soccer pass can also be defined as a relationship between members of an organization. Therefore, within a game, the passing relationships of each team can be defined as a single network structure, and the performance of each team can be quantified via SNA. Thus, the network indicators generated from the pass data can represent the performance of a team in more detail than data obtained through notational analysis, which reflects only the number of passes between players. To this end, we apply SNA to generate and use indicators based on the pass distribution between players to obtain a win–lose prediction.

3. Development of the SWLPS

This section provides a detailed description of the framework, which includes the three phases of the proposed *SWLPS*. The first phase creates the actual network indicators (ANIs) (specifically, historical pass counts) for training the *SWLPS*. The second phase estimates the pass counts of future games using linear regression and stochastic gradient descent (SGD) and calculates the predicted network indicators (PNIs) with the counts for testing the *SWLPS*. In the third phase, the performance of the *SWLPS* is evaluated using GB and the moving window method (Jang et al., 1993; Lee et al., 2017) with the ANIs and PNIs. Fig. 1 presents the overall structure of the *SWLPS*.

3.1. Phase 1: Creation of ANIs

At this phase, the ANIs $\{I_1, I_2, \dots, I_n\}$ are generated using SNA with historical pass counts and are used as an independent variable in the training period. Note that the dependent variable is the match result (win or lose). SNA, which has been widely studied and associated with team performance (Yang and Tang, 2004), uses network indicators to measure certain group characteristics, such as cohesion, conflict, and leadership. Network indicators are divided into individual, sub-group, and group structure measurements based on the focus of analysis (Wey et al., 2008; Sueur et al., 2011). Individual measurements describe a specific individual's position in the network and the potential effect it has on others. Sub-group measurements describe the presence of subgroups in the network, reflecting the distribution of the ties. Group structure measurements describe aspects of the overall network structure (Wey et al., 2008) and are especially useful for questions about group cohesion (Sueur et al., 2011). In this study, the network indicators of the individual measurements and of the group structure measurements are used to utilize the characteristics of the organization's members as well as the organization. Thus, we use seven network indicators as ANIs, such as the degree centrality (DC), weighted degree centrality (WDC), eigenvector centrality (EC), betweenness centrality (BC), closeness centrality (CC), average path length (APL), and network diameter (ND). Table 1 summarizes the definitions and equations for ANIs used in this study.

3.2. Phase 2: Generating the PNIs

This phase generates the PNIs $\{\hat{I}_1, \hat{I}_2, \dots, \hat{I}_n\}$ as independent variables of the testing period of the *SWLPS*. To obtain the PNIs, the pass

Table 1
Definitions and equations of ANIs.

Network indicators	Definitions	Equations
DC	The DC of a node is defined as the number of edges incident upon that node (Borgatti and Everett, 1997).	$\mathrm{DC}(i) = \sum_j^N x_{ij}$ where x_{ij} is 1 if i is connected to j and 0 otherwise.
WDC	The WDC, an extended version of the DC, is used to the sum the weights when analyzing weighted networks (Barrat et al., 2004; Newman, 2004; Opsahl et al., 2010).	$\mathrm{WDC}(i) = \sum_j^N w_{ij}$ where w_{ij} is defined as the weight between i and j.
EC	The EC, proposed by Bonacich (1972), is defined as the principal eigenvector of the adjacency matrix of a graph. The EC can be thought of as a weighted degree measure in which the centrality of a node is proportional to the sum of the centralities of the nodes it is adjacent to (Borgatti and Everett, 1997).	$\mathrm{EC}(i) = \frac{1}{\lambda} \sum_{j=1}^n A_{ij} \cdot \mathrm{DC}(j)$ where A_{ij} is 1 if there is an edge between i and j and 0 otherwise. λ is a constant.
BC	The BC is defined as the share of times that node i needs node k (whose centrality is being measured) to reach j via the shortest path (Borgatti, 2005).	$\mathrm{BC}(i) = \frac{g_{jk}(i)}{g_{jk}}$ where g_{jk} is the number of binary shortest paths between two nodes, and $g_{jk}(i)$ is the number of those paths that go through i.
CC	The CC, defined by Freeman et al. (1979), is inversely proportional to the total distance from the node of interest to all other nodes in the network.	$\mathrm{CC}(i) = \left[\sum_j^N d(i,j)\right]^{-1} \cdot N$
APL	The APL is the average of the shortest paths between all pairs of nodes in a network. It is often used as a measure of network efficiency (Lee et al., 2006)	$\mathrm{APL} = \frac{1}{2 \cdot N(N-1)} \cdot \sum_{i \neq j} s(i,j)$ where $s(i,j)$ is the shortest distance between i and j.
ND	The ND is the maximum distance between two nodes in a network (Nascimento et al., 2003)	$\mathrm{MAX}(d(i,j)), \forall i$

Notes: i is the focal node, j represents all other nodes, and N is the total number of nodes. $d(i,j)$ is the distance between nodes i and j.

Table 2
Definition of six features affecting the pass counts.

Features	Definitions
Historical passes between players (James, 2006)	Average number of past-game passes between two players.
Position (Scoulding et al., 2004)	Whether two players are of the same position.
Rank (Lago-Ballesteros and Lago-Peñas, 2010)	Whether the rank of team A is higher than the rank of team B.
Historical passes between positions (Scoulding et al., 2004)	Average number of passes in past games between two positions.
Average BC (Pena and Touchette, 2012)	Average BC of both players averaged over past games.
Average pass completion rate between players (Bradley et al., 2014)	Average percentage of passes completed for two specific players averaged over past games.

Table 3
Rounds of the CL.

Rounds	No. of matches
Play-offs	20
Group stage (match day 1–6)	96
Round of 16	16
Quarter-finals	8
Semi-finals	4
Final	1

counts of a future match are predicted using linear regression and SGD. Linear regression is used to predict basis pass counts, and SGD is used to optimize the regression coefficient (weight) for more sophisticated forecasting. Eq. (1) is example of a linear regression equation used in this study.

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{1}$$

$\hat{Y}$ is the predicted pass count (PPC) between two players, and $\{x_1, x_2, \ldots, x_p\}$ are the features that affect the pass counts. Table 2 presents detailed definitions of the features.

β_p includes the weights of the features and is optimized using SGD, which is a type of gradient descent and one of the simplest and most popular stochastic optimization methods (Shamir and Zhang, 2013) and has achieved state of-the-art performance on a variety of machine learning tasks (Chakroun et al., 2017). The advantage of SGD is that each step only relies on a single derivative, and thus, the computational cost is $1/n$ times that of batch gradient descent (Johnson and Zhang, 2013). SGD estimates the gradient on the basis of a single randomly picked example in each iteration and directly optimizes the expected risk since the examples are randomly drawn from the ground truth distribution (Bottou, 2010). Moreover, according to Zhang (2004), compared with traditional perceptron methods, loss minimization-based SGD can be superior (at least in principle). Eq. (2) is the SGD equation used in this study.

$$\beta_{p,t+1} = \beta_{p,t} - \gamma \nabla_\beta Q(z_t, \beta_{p,t}) \tag{2}$$

where γ is the step size controlling the rate of descent, ∇_β is the gradient of β, and $Q(z_t, \beta_{p,t})$ is a $Loss$ (see Eq. (3)) on the randomly picked sample z that consists of the inputs and outputs at t. The SGD updates $\beta_{p,t+1}$ to minimize $Loss$. If $Loss$ converges to zero or if a specified iteration numbers are completed, the optimization process ends.

$$Loss = (predicted\ pass\ counts - actual\ pass\ counts)^2 = (\hat{y} - y)^2 \tag{3}$$

The PPCs are generated using the optimized β_p and are applied to all participating teams in future matches. The PNIs are calculated using the PPC with the same equations as those used with the ANIs.

3.3. Phase 3: Win–lose prediction using GB

At this phase, the ANIs and PNIs generated at phases 1 and 2 are transformed into differences with the opposing team. We use the training and testing data for win–lose predictions. The difference of network indicators between two teams is appropriate for win–lose prediction since this difference is determined by the performance difference between the two teams. Since the prediction is performed separately for each season, the training data are not accumulated, and the learning of GB is limited. Thus, we compare GB with various classifiers on several seasons and sections to determine the stability of the system in additional seasons.

To assess the SWLPS using GB, the following most representative classifiers of machine learning are employed: support vector machine (SVM) (Wang et al., 2013; Guyon et al., 2002; Joachims, 1999; Furey

Table 4
Total of 434 matches for 126 teams in the 2013–2015 CL schedule.

Seasons	No. of participating teams	No. of matches
2013–2014	42 (advance to the finals: 32 teams; preliminary elimination: 10 teams)	145
2014–2015		144[a]
2015–2016		145

[a] Missing final game content in the press kit.

Table 5
The seven ANIs of the play-off match between GNK Dinamo Zagreb and FK Austria Wien in the 2013–2014 season.

(a) GNK Dinamo Zagreb

Players	I_1	I_2	I_3	I_4	I_5	I_6	I_7
1	14.00	53.00	0.33	0.36	0.76	1.29	4.00
2	19.00	66.00	0.12	3.79	0.76	1.29	4.00
3	21.00	105.00	0.50	2.87	0.87	1.29	4.00
4	21.00	74.00	0.22	6.34	0.81	1.29	4.00
5	21.00	85.00	0.36	5.13	0.81	1.29	4.00
6	19.00	57.00	0.13	1.29	0.72	1.29	4.00
7	23.00	109.00	0.44	7.19	0.93	1.29	4.00
8	18.00	42.00	0.12	1.02	0.76	1.29	4.00
9	19.00	79.00	0.30	1.59	0.81	1.29	4.00
10	23.00	80.00	0.29	15.39	0.93	1.29	4.00
11	18.00	51.00	0.16	2.09	0.72	1.29	4.00
12	16.00	48.00	0.14	1.94	0.72	1.29	4.00
13	1.00	1.00	0.00	0.00	0.00	1.29	4.00
14	9.00	18.00	0.06	0.00	0.59	1.29	4.00

(b) FK Austria Wien

Players	I_1	I_2	I_3	I_4	I_5	I_6	I_7
1	9.00	22.00	0.16	0.00	0.00	1.37	4.00
2	15.00	58.00	0.40	1.30	0.00	1.37	4.00
3	12.00	46.00	0.13	0.17	0.00	1.37	4.00
4	18.00	67.00	0.44	4.92	0.00	1.37	4.00
5	15.00	31.00	0.09	0.65	0.00	1.37	4.00
6	19.00	68.00	0.37	5.39	0.00	1.37	4.00
7	21.00	61.00	0.29	9.10	0.00	1.37	4.00
8	17.00	42.00	0.18	4.72	0.00	1.37	4.00
9	20.00	77.00	0.44	9.69	0.00	1.37	4.00
10	18.00	49.00	0.18	14.67	0.00	1.37	4.00
11	15.00	50.00	0.35	2.40	0.00	1.37	4.00
12	4.00	4.00	0.00	0.00	0.00	1.37	4.00
13	5.00	5.00	0.01	0.00	0.00	1.37	4.00
14	0.00	0.00	0.00	0.00	0.00	1.37	4.00

et al., 2000) neural network (NN) (Taormina et al., 2015; Sefeedpari et al., 2016; Gholami et al., 2015; Chen and Chau, 2016), decision tree (DT) (Vlahou et al., 2003; Sugumaran and Ramachandran, 2007; Elson et al., 2004), case-based reasoning (CBR) (Begum et al., 2014; Li and Yeh, 2004; Jo et al., 1997), and logistic regression (LR) (Ohlmacher and Davis, 2003; Swaminathan and Rogers, 1990; Tu, 1996). Each technique is defined as follows.

3.3.1. *Gradient boosting*

Different from bagging, the boosting method sequentially generates base models (Zhang and Haghani, 2015). To establish a connection with the statistical framework, a gradient descent-based formulation of boosting methods was derived (Freund and Schapire, 1997; Friedman et al., 2000; Friedman, 2001). This formulation of boosting methods and the corresponding models were called GB (Natekin and Knoll, 2013). GB, a powerful learning strategy originally designed for classification problems, has been successfully extended to regressions (Persson et al., 2017) and produces competitive, highly robust, interpretable procedures for both regression and classification; it is especially appropriate for mining less-than-clean data (Friedman, 2001). GB is an iterative algorithm that combines simple parameterized functions exhibiting poor performance to produce a highly accurate prediction rule (Guelman, 2012). The prediction accuracy is improved through developing multiple models in sequence by emphasizing the training cases that are challenging to estimate (Zhang and Haghani, 2015).

3.3.2. *Support vector machines*

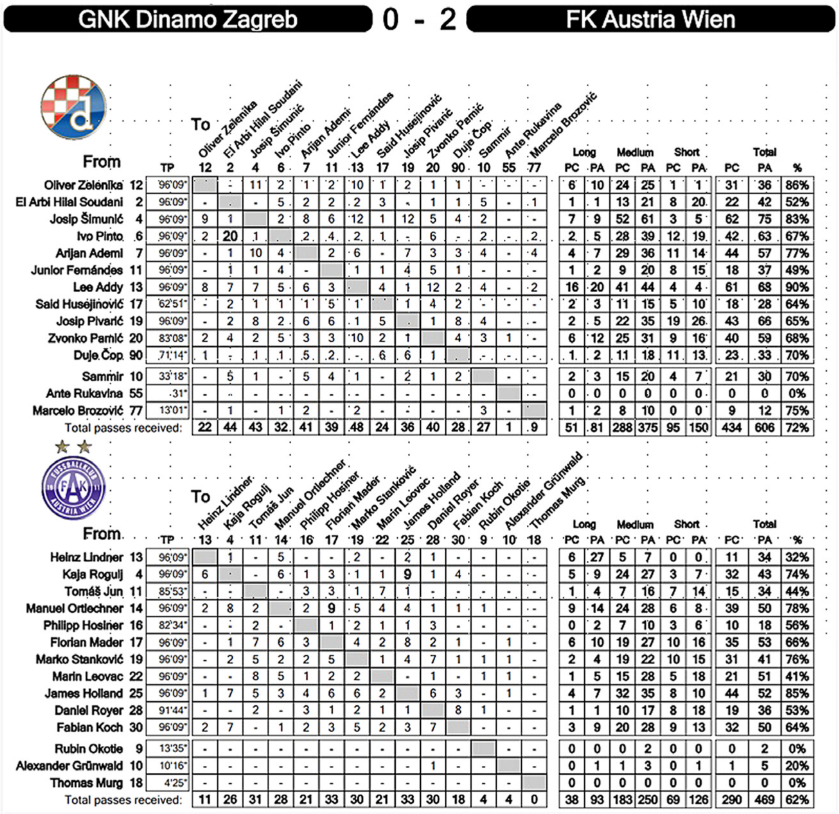
SVMs (Boser et al., 1992; Vapnik and Vapnik, 1998) are known as an excellent tool for classification and regression problems that exhibits favorable generalization performance (Osowski et al., 2004). SVM uses a linear model to separate sample data through some non-linear mapping from the input vectors into a high-dimensional feature space (Wang et al., 2013). The method then finds the linear optimal hyperplane (a decision boundary) to separate the dataset of one class from another. The hyperplane with the maximum margin between the two classes is sought by the SVM classifier (Ma et al., 2018).

3.3.3. *Neural networks*

Neural networks (NNs) are a computer-based technique modeled on the structure and behavior of neurons in the human brain that can be trained to recognize and categorize complex patterns (Bishop, 1995). An artificial NN (ANN) is typically defined as a network composed of a large number of simple processors (neurons) that are massively interconnected, that operate in parallel and that learn from experience (examples) (Specht, 1991). ANNs calculate the weights of each data point within the hidden nodes and keeps them for further use to predict

Table 6
ANIs of the team and results for several matches in the 2013–2014 season.

Matches	Teams	Avg. I_1	Avg. I_2	Avg. I_3	Avg. I_4	Avg. I_5	I_6	I_7	Results
1	GNK Dinamo Zagreb	17.29	62.00	0.23	3.50	0.73	1.29	4.00	Loser
	FK Austria Wien	13.43	41.43	0.22	3.79	0.00	1.37	4.00	Winner
2	PFC Ludogorets 1945	13.57	33.71	0.22	5.57	0.65	1.46	5.00	Loser
	FC Basel 1893	13.43	42.29	0.22	3.64	0.00	1.35	5.00	Winner
3	FC Viktoria Plzen	14.71	48.29	0.22	5.00	0.67	1.41	5.00	Winner
	NK Maribor	15.54	77.54	0.25	3.62	0.04	1.33	6.00	Loser
4	FC Shakhter Karagandy	12.67	33.00	0.26	5.25	0.70	1.48	5.00	Winner
	Celtic FC	17.54	84.92	0.24	3.38	0.80	1.28	6.00	Loser
5	FC Steaua Bucuresti	18.43	60.86	0.22	3.79	0.79	1.29	5.00	Draw
	Legia Warszawa	13.71	39.86	0.24	6.93	0.67	1.53	8.00	
6	Olympique Lyonnais	15.38	56.77	0.24	4.31	0.75	1.36	7.00	Loser
	Real Sociedad de Futbol	17.14	45.86	0.25	4.43	0.76	1.34	5.00	Winner
7	FC Schalke 04	17.57	96.86	0.22	4.29	0.77	1.33	6.00	Draw
	PAOK FC	15.00	43.43	0.23	6.14	0.70	1.47	5.00	



GNK Dinamo Zagreb 0 - 2 FK Austria Wien

To			Oliver Zelenika	El Arbi Hilal Soudani	Josip Šimunić	Ivo Pinto	Arijan Ademi	Junior Fernandes	Lee Addy	Said Husejinović	Josip Pivarić	Zvonko Pamić	Duje Čop	Sammir	Ante Rukavina	Marcelo Brozović	Long		Medium		Short		Total		
From		TP	12	2	4	6	7	11	13	17	19	20	90	10	55	77	PC	PA	PC	PA	PC	PA	PC	PA	%
Oliver Zelenika	12	96'09"		-	11	2	1	2	10	1	2	1	1	-	-	-	6	10	24	25	1	1	31	36	86%
El Arbi Hilal Soudani	2	96'09"	-		-	5	2	2	2	3	-	1	1	5	-	1	1	1	13	21	8	20	22	42	52%
Josip Šimunić	4	96'09"	9	1		2	8	6	12	1	12	5	4	2	-	-	7	9	52	61	3	5	62	75	83%
Ivo Pinto	6	96'09"	2	20	1		2	4	2	1	-	6	-	2	-	2	2	5	28	39	12	19	42	63	67%
Arijan Ademi	7	96'09"	-	1	10	4		2	6	-	7	3	3	4	-	4	4	7	29	36	11	14	44	57	77%
Junior Fernándes	11	96'09"	-	1	1	4	-		1	1	4	5	1	-	-	-	1	2	9	20	8	15	18	37	49%
Lee Addy	13	96'09"	8	7	7	5	6	3		4	1	12	2	4	-	2	16	20	41	44	4	4	61	68	90%
Said Husejinović	17	62'51"	-	2	1	1	1	5	1		1	4	2	-	-	-	2	3	11	15	5	10	18	28	64%
Josip Pivarić	19	96'09"	-	2	8	2	6	6	1	5		1	8	4	-	-	2	5	22	35	19	26	43	66	65%
Zvonko Pamić	20	83'08"	2	4	2	5	3	3	10	2	1		4	3	1	-	6	12	25	31	9	16	40	59	68%
Duje Čop	90	71'14"	1	-	1	1	5	2	-	6	6	1		-	-	-	1	2	11	18	11	13	23	33	70%
Sammir	10	33'18"	-	5	1	-	5	4	1	-	2	1	2		-	-	2	3	15	20	4	7	21	30	70%
Ante Rukavina	55	31"	-	-	-	-	-	-	-	-	-	-	-	-		-	0	0	0	0	0	0	0	0	0%
Marcelo Brozović	77	13'01"	-	1	-	1	2	-	2	-	-	-	-	3	-		1	2	8	10	0	0	9	12	75%
Total passes received:			22	44	43	32	41	39	48	24	36	40	28	27	1	9	51	81	288	375	95	150	434	606	72%

To			Heinz Lindner	Kaja Rogulj	Tomáš Jun	Manuel Ortlechner	Philipp Hosiner	Florian Mader	Marko Stanković	Marin Leovac	James Holland	Daniel Royer	Fabian Koch	Rubin Okotie	Alexander Grünwald	Thomas Murg	Long		Medium		Short		Total		
From		TP	13	4	11	14	16	17	19	22	25	28	30	9	10	18	PC	PA	PC	PA	PC	PA	PC	PA	%
Heinz Lindner	13	96'09"		1	-	5	-	-	2	-	2	1	-	-	-	-	6	27	5	7	0	0	11	34	32%
Kaja Rogulj	4	96'09"	6		-	6	1	3	1	1	9	1	4	-	-	-	5	9	24	27	3	7	32	43	74%
Tomáš Jun	11	85'53"	-	-		-	3	3	1	7	1	-	-	-	-	-	1	4	7	16	7	14	15	34	44%
Manuel Ortlechner	14	96'09"	2	8	2		2	9	5	4	4	1	1	1	-	-	9	14	24	28	6	8	39	50	78%
Philipp Hosiner	16	82'34"	-	-	2	-		1	2	1	1	3	-	-	-	-	0	2	7	10	3	6	10	18	56%
Florian Mader	17	96'09"	-	1	7	6	3		4	2	8	2	1	-	1	-	6	10	19	27	10	16	35	53	66%
Marko Stanković	19	96'09"	-	2	5	2	2	5		1	4	7	1	1	1	-	2	4	19	22	10	15	31	41	76%
Marin Leovac	22	96'09"	-	-	8	5	1	2	2		-	1	-	1	1	-	1	5	15	28	5	18	21	51	41%
James Holland	25	96'09"	1	7	5	3	4	6	6	2		6	3	-	1	-	4	7	32	35	8	10	44	52	85%
Daniel Royer	28	91'44"	-	-	2	-	3	1	2	1	1		8	1	-	-	1	1	10	17	8	18	19	36	53%
Fabian Koch	30	96'09"	2	7	-	1	2	3	5	2	3	7		-	-	-	3	9	20	28	9	13	32	50	64%
Rubin Okotie	9	13'35"	-	-	-	-	-	-	-	-	-	-	-		-	-	0	0	0	2	0	0	0	2	0%
Alexander Grünwald	10	10'16"	-	-	-	-	-	-	-	-	-	1	-	-		-	0	1	1	3	0	1	1	5	20%
Thomas Murg	18	4'25"	-	-	-	-	-	-	-	-	-	-	-	-	-		0	0	0	0	0	0	0	0	0%
Total passes received:			11	26	31	28	21	33	30	21	33	30	18	4	4	0	38	93	183	250	69	126	290	469	62%

(a) PDF format.

Passing distribution																									
Play-off round 1st leg - Wednesday 21 August 2013																									
Stadion Maksimir - Zagreb																									
GNK Dinamo Zagreb									0 - 2						FK Austria Wien										
		To																							
																	Long		Medium		Short			Total	
From		TP	12	2	4	6	7	11	13	17	19	20	90	10	55	77	PC	PA	PC	PA	PC	PA	PC	PA	%
Oliver Zelenika	12	96'09		-	11	2	1	2	10	1	2	1	1	-	-	-	6	10	24	25	1	1	31	36	86
El Arbi Hilal Soudani	2	96'09	-		-	5	2	2	2	3	-	1	1	5	-	1	1	1	13	21	8	20	22	42	52
Josip Simunic	4	96'09	9	1		2	8	6	12	1	12	5	4	2	-	-	7	9	52	61	3	5	62	75	83
Ivo Pinto	6	96'09	2	20	1		2	4	2	1	-	6	-	2	-	2	2	5	28	39	12	19	42	63	67
Arijan Ademi	7	96'09	-	1	10	4		2	6	-	7	3	3	4	-	4	4	7	29	36	11	14	44	57	77
Junlor Femandes	11	96'09	-	1	1	4	-		1	1	4	5	1	-	-	-	1	2	9	20	8	15	18	37	49
Lee Addy	13	96'09	8	7	7	5	6	3		4	1	12	2	4	-	2	16	20	41	44	4	4	61	68	90
Said Husejinovic	17	62'51	-	2	1	1	1	5	1		1	4	2	-	-	-	2	3	11	15	5	10	18	28	64
Josip Pivaric	19	96'09	-	2	8	2	6	6	1	5		1	8	4	-	-	2	5	22	35	19	26	43	66	65
Zvonko Pamic	20	83'08	2	4	2	5	3	3	10	2	1		4	3	1	-	6	12	25	31	9	16	40	59	68
Duje Cop	90	71'14	1	-	1	1	5	2	-	6	6	1		-	-	-	1	2	11	18	11	13	23	33	70
Sammir	10	33'18	-	5	1	-	5	4	1	-	2	1	2		-	-	2	3	15	20	4	7	21	30	70
Ante Rukavina	55	31'	-	-	-	-	-	-	-	-	-	-	-	-		-	0	0	0	0	0	0	0	0	0
Marcelo Brozovic	77	13'01	-	1	-	1	2	-	2	-	-	-	-	3	-		1	2	8	10	0	0	9	12	75
Total passes received:			22	44	43	32	41	39	48	24	36	40	28	27	1	9	51	81	288	375	95	150	434	606	72
		To																							
																	Long		Medium		Short			Total	
From		TP	13	4	11	14	16	17	19	22	25	28	30	9	10	18	PC	PA	PC	PA	PC	PA	PC	PA	%
Helnz Lindner	13	96'09		1	-	5	-	-	2	-	2	1	-	-	-	-	6	27	5	7	0	0	11	34	32
Kaja Rogulj	4	96'09	6		-	6	1	3	1	1	9	1	4	-	-	-	5	9	24	27	3	7	32	43	74
Tomas Jun	11	85'53	-	-		-	3	3	1	7	1	-	-	-	-	-	1	4	7	16	7	14	15	34	44
Manuel Ortlechner	14	96'09	2	8	2		2	9	5	4	4	1	1	1	-	-	9	14	24	28	6	8	39	50	78
Philipp Hosiner	16	82'34	-	-	2	-		1	2	1	1	3	-	-	-	-	0	2	7	10	3	6	10	18	56
Florian Mader	17	96'09	-	1	7	6	3		4	2	8	2	1	-	1	-	6	10	19	27	10	16	35	53	66
Marko Stankovic	19	96'09	-	2	5	2	2	5		1	4	7	1	1	1	-	2	4	19	22	10	15	31	41	76
Marin Leovac	22	96'09	-	-	8	5	1	2	2		-	1	-	1	1	-	1	5	15	28	5	18	21	51	41
James Holland	25	96'09	1	7	5	3	4	6	6	2		6	3	-	1	-	4	7	32	35	8	10	44	52	85
Daniel Royer	28	91'44	-	-	2	-	3	1	2	1	1		8	1	-	-	1	1	10	17	8	18	19	36	53
Fabian Koch	30	96'09	2	7	-	1	2	3	5	2	3	7		-	-	-	3	9	20	28	9	13	32	50	64
Rubin Okotle	9	13'35	-	-	-	-	-	-	-	-	-	-	-		-	-	0	0	0	2	0	0	0	2	0
Alexander Grunwald	10	10'16	-	-	-	-	-	-	-	-	-	1	-	-		-	0	1	1	3	0	1	1	5	20
Thomas Murg	18	4'25	-	-	-	-	-	-	-	-	-	-	-	-	-		0	0	0	0	0	0	0	0	0
Total passes received:			11	26	31	28	21	33	30	21	33	30	18	4	4	0	38	93	183	250	69	126	290	469	62

(b) CSV format.

Fig. 2. Example of (a) PDF format and (b) CSV format for the pass count data generated for matches between GNK Dinamo Zagreb and FK Austria Wien.

Table 7
PPCs between players in the match between Barcelona and Manchester City in the round of 16 during the 2013–2014 season.

(a) Barcelona

Players		PPC	Players		PPC	Players		PPC
From	To		From	To		From	To	
11	10	5.20	16	24	5.90	4	14	5.76
	14	4.98		18	4.31		22	6.26
	22	4.99		3	10.62		16	4.80
	16	7.09		4	5.56		18	4.38
	18	4.34		6	17.35		3	5.68
	6	8.33		9	6.27		6	6.41
10	11	5.33		8	11.12		9	4.88
	14	5.48	24	8	4.39		8	7.15
	22	8.79		16	5.96	6	11	10.96
	16	6.37		18	4.67		10	10.22
	18	4.59		6	5.72		14	6.69
	4	6.46	18	24	4.92		22	14.41
	6	9.80		10	4.91		16	14.43
	9	5.81		14	4.44		18	4.21
	8	7.49		16	4.97		3	8.45
14	24	5.83		1	4.91		4	6.69
	10	5.33		4	4.93		24	6.18
	22	5.11		6	4.98		9	10.46
	16	8.08		8	4.99		8	11.27
	18	3.85	1	14	5.98	9	10	4.80
	1	5.82		22	5.01		22	7.29
	3	10.04		16	5.89		16	5.84
	4	5.97		18	4.48		3	5.73
	6	8.95		3	7.49		4	4.83
	9	4.72		4	4.63		6	7.69
	8	7.86		6	5.04		8	5.75
22	11	5.61		9	4.61	8	11	14.52
	10	10.86		8	5.18		10	6.93
	14	4.64	3	11	6.10		14	5.72
	16	8.84		10	5.23		22	6.59
	18	3.99		14	12.22		16	9.18
	3	7.62		22	6.24		24	3.87
	4	6.97		16	10.18		18	4.35
	6	12.64		18	3.75		1	4.72
	9	7.19		1	6.58		3	6.26
	8	7.03		4	6.85		4	7.68
16	11	5.79		6	9.95		6	11.39
	10	8.72		9	6.82		9	4.97
	14	7.40	4	11	8.07			
	22	8.02		10	6.59			

(continued on next page)

the demand (Sefeedpari et al., 2016). After the training is complete, the weights are fixed and can be used for testing (Raj et al., 2015).

3.3.4. Decision tree

The DT technique is a popular classification method that builds a DT and classifies the given data using this tree (Berson and Smith, 2002; Kim et al., 2001). The DT is a tree in which each non-leaf node denotes a test on an attribute of cases, each branch corresponds to an outcome of the test, and each leaf node denotes a class prediction (Cho et al., 2002). To classify an unknown dataset, the attribute values are tested against the DT. A path is traced from the root to a leaf node that holds the class prediction for the test data. Different from the naive Bayes classifier and

Table 7 *(continued)*

(b) Manchester City FC

Players		PPC	Players		PPC	Players		PPC
From	To		From	To		From	To	
25	10	7.16	22	1	5.90	5	21	7.32
	26	5.68		4	5.27		10	6.24
	21	8.24		9	5.50		1	5.87
	22	5.55		8	6.49		4	5.74
	42	8.52	10	21	5.84		6	4.98
	5	6.92		22	5.23		9	6.09
	4	6.06		42	6.37		8	6.71
	6	6.28		5	5.62	4	25	6.11
	9	5.92		4	5.13		13	4.84
	8	7.63		6	5.23		15	5.59
13	25	7.23		8	6.58		21	5.58
	26	4.98	42	25	8.71		22	4.96
	15	5.49		13	6.60		10	5.84
	21	6.57		15	7.58		42	6.34
	22	4.98		21	8.67		26	4.84
	42	7.34		22	5.99		1	6.09
	9	5.84		26	6.11		5	5.84
15	26	5.05		1	5.40		6	4.83
	21	5.41		5	8.08		9	5.33
	42	6.79		4	5.88		8	5.58
	5	9.72		6	5.00	6	10	5.62
	4	5.19		9	5.86		21	5.96
	9	5.66		8	10.02		22	4.86
21	25	8.01	26	25	6.48		42	5.51
	13	6.26		13	5.21		1	6.72
	15	6.39		21	5.57		4	4.87
	22	6.02		22	5.32		8	5.60
	10	5.67		42	5.61	9	25	6.32
	42	8.98		1	6.46		15	5.55
	5	6.65		4	4.73		21	5.54
	4	5.29	1	10	7.61		22	5.05
	9	5.89		13	5.88		42	6.43
	8	6.01		15	6.14		8	5.42
22	25	6.16		22	5.76	8	25	7.80
	13	5.14		5	5.16		21	5.67
	15	5.27		4	5.65		42	9.87
	21	7.22		9	6.12		5	6.31
	10	5.53	5	25	6.51		4	5.82
	42	6.52		26	6.09		6	5.06
	26	4.89		15	10.88		9	6.17

NN, the output model comprising the DT is interpretable and can easily be converted into classification rules (Wu, 2009).

3.3.5. *Case-based reasoning*

CBR is a rich and knowledge-intensive method for capturing past experiences, enhancing existing problem-solving methods and improving the overall learning capabilities of machines (Schank, 1983). The full set of past experiences encapsulated in individual cases is called the case base. Given a new problem, problems are solved by finding and adapting a matching case and solution; a nearest neighbor matching technique is often used for this purpose (Bose and Mahapatra, 2001; Riesbeck and Shank, 1989).

3.3.6. *Logistic regression*

LR is a regression model used to predict the probability of categorical dependent variables (Ma et al., 2018). This model has been widely used in the conditional monitoring of cutting tools for which the binary classified variables are normality and failure (Chen et al., 2011; Li et al., 2015). LR is used to explain a dependent variable through one or more independent variables (Razanamahandry et al., 2018).

4. Empirical studies

4.1. Data description

The Union of European Football Association's (UEFA) CL consists of play-offs, a group stage, a round of 16, quarter-finals, semi-finals, and a final. Table 3 shows the progression of the CL and the number of matches in each round.

A total of 42 teams participate, and 145 matches are played in each season. In this study, data from 434 CL matches from 2013 to 2015 were used (see Table 4).

The data were obtained from the UEFA press kits. To develop the *SWLPS*, pass counts, squad lists, rankings, and scores of every match were collected for the selected three seasons (2013–2014, 2014–2015, and 2015–2016 seasons). The pass count data included the number of passes completed and the number of attempted passes between all pairs of players. The squad list data included the positions of players, the ranking data included the rankings of each team in the previous season, and the score data included every goal of each team for every match.

Fig. 2 shows an example of pass counts, the number of successful passes between each player and the total number of passes, including the failed passes that were recorded in the press kit. Fig. 2(a) shows the pass data provided in portable document format (PDF) format from the UEFA press kits, and Fig. 2(b) shows the comma separated values (CSV) format data that were parsed using the Python programming language.

4.2. SWLPS experimental results

In phase 1, the ANIs $I_1, I_2, \dots, I_7$ (DC, WDC, EC, BC, CC, APL, and ND) were calculated using actual pass counts by the SNA. Table 5 presents an example of the ANI for the eleven main players and three substitute players in the match between GNK Dinamo Zagreb and FK Austria Wien.

$I_1, I_2, \dots, I_5$ are generated through individual measurements that focus on a specific individual's position in the network and the potential effect this position has on others (Wey et al., 2008). I_6 and I_7 are indicators generated through group structure measurements; these parameters describe the overall network structure and are particularly useful when analyzing questions regarding group cohesion (Sueur et al., 2011). It is necessary to average the performances of individual players since soccer is a team sport and since the result is decided by team performance. Table 6 shows the averaged $I_1, I_2, \dots, I_5$ for the individual players and the result for several matches in the 2013–2014 season.

In phase 2, the PPCs are first estimated to generate PNIs $\hat{I}_1, \hat{I}_2, \dots, \hat{I}_7$ using linear regression and SGD. Table 7 shows the estimated PPCs between players in the match of (a) Barcelona and (b) Manchester City in the round of 16 during the 2013–2014 season.

Table 8 shows the PNIs $\hat{I}_1, \hat{I}_2, \dots, \hat{I}_7$ calculated using the PPCs and SNA for the match between Barcelona and Manchester City in the round of 16 during the 2013–2014 season. Table 9 illustrates the PNIs for each team in the round of 16 during the 2013–2014 season.

In phase 3, the *SWLPS* was evaluated using GB and other techniques (SVM, CBR, NN, DT, and LR). To this end, the ANIs and PNIs for the teams generated in phases 1 and 2 were transformed into differences. Specifically, differences from ANIs were used as training data, and differences from PNIs were used as testing data. Table 10 shows the differences that were transformed from the seven indicators for each team. In the case of soccer matches, the match results are divided into win, draw, and lose. In this study, because the number of draws is not sufficiently large for training, only winning or losing is predicted by eliminating the case of a drawn game.

The performance of the *SWLPS* was improved by applying a moving window method to maximize the number of training data based on the prediction time point over the entire period. Because the teams and members participating in the league for each season are different, the prediction is conducted separately for each season. For this reason, the amount of training data was insufficient (after elimination of draw, the numbers of training data points at the time of the round of 16, which is the initial section of prediction, were 94, 90, and 95 games for the three seasons). The training period are compiled up to the semi-finals stage to learn the team's performance based on the network indicators. Table 11 describes the pairs of training and testing periods for the moving window method.

Table 8
PNIs of the match between Barcelona and Manchester City in the round of 16 during the 2013–2014 season.

(a) FC Barcelona

Players	$\hat{I}_1$	$\hat{I}_2$	$\hat{I}_3$	$\hat{I}_4$	$\hat{I}_5$	$\hat{I}_6$	$\hat{I}_7$
1	13.00	91.31	0.16	0.00	0.67	1.24	2.00
2	19.00	128.91	0.26	1.70	0.80	1.24	2.00
3	21.00	134.85	0.30	4.73	0.92	1.24	2.00
4	20.00	148.10	0.33	1.89	0.86	1.24	2.00
5	23.00	182.68	0.38	6.26	0.92	1.24	2.00
6	9.00	47.44	0.10	0.00	0.60	1.24	2.00
7	19.00	85.97	0.16	4.94	0.75	1.24	2.00
8	13.00	70.35	0.21	0.56	0.80	1.24	2.00
9	18.00	135.78	0.31	2.35	0.86	1.24	2.00
10	20.00	120.53	0.25	1.76	0.86	1.24	2.00
11	23.00	212.21	0.41	6.26	0.92	1.24	2.00
12	16.00	97.67	0.20	0.37	0.71	1.24	2.00
13	22.00	158.42	0.34	7.19	1.00	1.24	2.00

(b) Manchester City FC

Players	$\hat{I}_1$	$\hat{I}_2$	$\hat{I}_3$	$\hat{I}_4$	$\hat{I}_5$	$\hat{I}_6$	$\hat{I}_7$
1	19.00	131.27	0.33	5.44	0.81	1.34	2.00
2	13.00	76.37	0.22	1.16	0.68	1.34	2.00
3	14.00	90.71	0.20	1.92	0.65	1.34	2.00
4	22.00	142.73	0.32	7.78	0.81	1.34	2.00
5	21.00	117.58	0.30	7.64	0.87	1.34	2.00
6	14.00	83.67	0.21	1.74	0.68	1.34	2.00
7	23.00	162.14	0.39	10.79	0.93	1.34	2.00
8	14.00	76.99	0.21	2.22	0.68	1.34	2.00
9	13.00	78.75	0.19	2.09	0.68	1.34	2.00
10	18.00	120.73	0.29	5.28	0.81	1.34	2.00
11	24.00	131.38	0.33	10.17	1.00	1.34	2.00
12	13.00	70.53	0.20	1.36	0.68	1.34	2.00
13	16.00	92.66	0.19	2.25	0.65	1.34	2.00
14	16.00	106.74	0.25	2.17	0.68	1.34	2.00

According to the moving window number, difference values of ANIs and PNIs were used for each training and testing period, and the evaluation was conducted through GB and several classifiers. In the GB, there are four hyper-parameters that need to be tuned: (1) d, the depth of decision trees, which also controls the maximum interaction order of the model; (2) K, the number of iterations, which also corresponds to the numbers of decision trees; (3) α, the learning rate (also called shrinkage), which is usually a small positive value between 0 and 1, where lower values lead to slower fitting, thus requiring the user to increase K; and (4) η, the fraction of data that is used in each iteration (Touzani et al., 2018). In our study, GB was tuned as follows: ($d = 2$, $K = 50$, $\alpha = 0.1$, $\eta = 0.1$). These settings were the same for all seasons and sections.

Based on the experimental results, we determined that GB is the best classifier as far as performance is concerned; Table 12 shows the performance of each classifier.

An ANOVA was performed to evaluate the performance differences between machine learning techniques in terms of result predictions. The test is an extremely important method in exploratory and confirmatory data analysis (Vijayvargiya, 2009). According to the number of factors to be analyzed, ANOVA is classified as one-way ANOVA or two-way ANOVA. In this study, to evaluate the accuracy of the techniques, one-way ANOVA was conducted. The results were analyzed using the SPSS 20.0 software package (SPSS, Inc., Chicago, IL, USA), which

Table 9
Average PNI example of the round of 16 during the 2013–2014 season.

Matches	Teams	Avg. $\hat{I}_1$	Avg. $\hat{I}_2$	Avg. $\hat{I}_3$	Avg. $\hat{I}_4$	Avg. $\hat{I}_5$	$\hat{I}_6$	$\hat{I}_7$
1	FC Barcelona	16.43	82.38	0.25	5.14	0.74	1.40	2
	Manchester City FC	18.00	101.17	0.26	4.07	0.78	1.31	2
2	Manchester United FC	11.86	79.44	0.25	6.71	0.04	1.56	2
	Olympiacos FC	18.29	111.05	0.26	3.86	0.78	1.30	3
3	Borussia Dortmund	14.57	85.90	0.25	6.14	0.69	1.47	2
	FC Zenit	15.00	93.78	0.25	4.57	0.04	1.38	3
4	Arsenal FC	16.92	102.96	0.27	3.54	0.78	1.29	3
	FC Bayern Munchen	14.15	85.74	0.26	5.69	0.70	1.47	2
5	AC Milan	14.00	88.14	0.26	5.15	0.71	1.43	3
	Club Atletico de Madrid	15.29	67.35	0.25	5.71	0.71	1.44	3
6	Bayer 04 Leverkusen	18.15	124.17	0.26	2.92	0.82	1.24	2
	Paris Saint-Germain	16.62	115.88	0.26	3.69	0.78	1.31	2
7	FC_Schalke_04	17.29	108.13	0.25	4.36	0.77	1.34	2
	Real Madrid CF	17.43	89.22	0.25	4.50	0.77	1.35	2
8	Borussia Dortmund	16.71	90.40	0.25	4.93	0.75	1.38	2
	FC Zenit	18.29	107.80	0.26	3.86	0.78	1.30	3
9	FC Barcelona	16.86	95.67	0.26	4.57	0.76	1.35	2
	Manchester City FC	15.43	95.70	0.24	5.64	0.72	1.43	3
10	Manchester United FC	17.14	105.88	0.26	4.43	0.76	1.34	3
	Olympiacos FC	16.29	102.68	0.25	5.29	0.74	1.41	3
11	Chelsea FC	18.00	97.53	0.27	3.00	0.82	1.25	3
	Galatasaray AS	14.29	62.94	0.25	6.43	0.69	1.49	3
12	FC Schalke 04	13.57	76.05	0.24	6.86	0.68	1.53	2
	Real Madrid CF	14.86	83.60	0.25	6.29	0.70	1.48	2
13	AC Milan	19.85	128.19	0.27	2.08	0.86	1.17	3
	Club Atletico de Madrid	18.57	81.82	0.26	3.71	0.80	1.29	3
14	Bayer 04 Leverkusen	19.29	115.11	0.26	3.36	0.82	1.26	3
	Paris Saint-Germain	18.71	92.71	0.26	3.64	0.80	1.28	2

Table 10
Differences of seven indicators between FK Austria Wien and GNK Dinamo Zagreb.

	Avg. $\hat{I}_1$	Avg. $\hat{I}_2$	Avg. $\hat{I}_3$	Avg. $\hat{I}_4$	Avg. $\hat{I}_5$	$\hat{I}_6$	$\hat{I}_7$	Results
FK Austria Wien	13.42	41.42	0.21	3.78	0	1.36	4	Winner
GNK Dinamo Zagreb	17.28	62	0.22	3.5	0.72	1.28	4	Loser
Differences	−3.85	−20.57	−0.01	0.28	0.72	0.078	0	1 (win)

Table 11
Training and testing periods for win–lose prediction based on the moving window method.

Seasons	Moving window no.	Training periods	Testing periods
2013–2014	1	Group stage	Round of 16
	2	Group stage + round of 16	Quarter-finals
	3	Group stage + round of 16 + quarter-finals	Semi-finals
2014–2015	1	Group stage	Round of 16
	2	Group stage + round of 16	Quarter-finals
	3	Group stage + round of 16 + quarter-finals	Semi-finals
2015–2016	1	Group stage	Round of 16
	2	Group stage + round of 16	Quarter-finals
	3	Group stage + round of 16 + quarter-finals	Semi-finals

Table 12
Comparison of classifiers for win–lose prediction.

Techniques	Ranks (Accuracy (%))								
	2013–2014 season			2014–2015 season			2015–2016 season		
	Round of 16	Quarter-final	Semi-final	Round of 16	Quarter-final	Semi-final	Round of 16	Quarter-final	Semi-final
GB	1 (57.14)	1 (83.33)	1 (100.00)	1 (66.67)	1 (83.33)	1 (100.00)	1 (83.33)	1 (83.33)	1 (66.67)
SVM	4 (42.86)	6 (33.33)	3 (66.67)	1 (66.67)	4 (66.67)	2 (66.67)	2 (66.67)	3 (66.67)	4 (33.33)
CBR	1 (57.14)	2 (66.67)	6 (33.33)	3 (50.00)	1 (83.33)	6 (0.00)	4 (58.33)	4 (50.00)	5 (0.00)
NN	4 (42.86)	3 (50.00)	1 (100.00)	3 (50.00)	6 (33.33)	5 (33.33)	4 (58.33)	4 (50.00)	1 (66.67)
DT	3 (50.00)	3 (50.00)	3 (66.67)	3 (50.00)	1 (83.33)	2 (66.67)	6 (50.00)	4 (50.00)	5 (0.00)
LR	6 (28.57)	3 (50.00)	3 (66.67)	3 (50.00)	4 (66.67)	2 (66.67)	2 (66.67)	2 (66.67)	1 (66.67)

is an enormously powerful data analysis package that can perform very complex statistical procedures (Pallant, 2013), as indicated in Table 13. The analysis indicates that the accuracies among classifiers exhibit significant differences. As shown in Table 12, the ranking of the techniques is significant and implies that the GB is the most suitable technique for the *SWLPS*.

Table 13
One-way ANOVA results quantifying the accuracy of the techniques.

	Sum of squares	Degrees of freedom	Mean square	*F*	*P*-value
Between groups	0.672	5	0.134	3.398	0.010
Within groups	1.899	48	0.040		
Total	2.572	53			

Table A.1
Comparison of classifiers for win–draw–lose predicion.

Classifiers	Dataset types	Seasons								
		13-14			14-15			15-16		
		Round of 16	Quarter-final	Semi-final	Round of 16	Quarter-final	Semi-final	Round of 16	Quarter-final	Semi-final
Gradient boosting	Balance dataset	0.69	0.63	0.5	0.56	0.75	0.75	0.38	0.5	0.75
	Imbalance dataset	0.56	0.5	0.5	0.44	0.5	0.5	0.56	0.5	0.5
Neural network	Balance dataset	0.38	0.25	0.75	0.38	0.5	0.75	0.38	0.63	0.5
	Imbalance dataset	0.31	0.38	0.5	0.25	0.25	0.5	0.56	0.5	0.5
Decision Tree	Balance dataset	0.56	0.25	0.5	0.5	0.38	0.5	0.38	0.38	0.25
	Imbalance dataset	0.69	0.38	0.5	0.44	0.38	0.5	0.38	0.38	0.25
Case-based learning	Balance dataset	0.25	0.63	0.25	0.19	0.38	0.25	0.5	0.25	0.5
	Imbalance dataset	0.5	0.38	0.25	0.31	0.5	0.25	0.38	0.38	0.25
Logistic regression	Balance dataset	0.38	0.38	0.5	0.31	0.5	0.5	0.38	0.25	0.25
	Imbalance dataset	0.31	0.38	0.5	0.44	0.5	0.5	0.44	0.5	0.5

Note that SVM is excluded because the SVM of SAS provides only binary classification.

5. Concluding remarks

In this study, we developed an *SWLPS* for soccer match predictions using SNA and GB. As shown through the previous efforts of scholars to predict and explain the outcomes of a sports event (Reed and O'Donoghue, 2005), analysis based on historical records is efficient. The *SWLPS* used ANIs calculated using SNA and match results (win or lose) as training data. The model then generated PPCs using linear regression and SGD, converted the PPCs into PNIs through SNA, and analyzed the PNIs as testing data for predicting match results. In this system, draws were eliminated from the prediction result because the amount of draw data to be used for training was insufficient. Therefore, we tried to perform additional experiments by oversampling the draw case to solve the data imbalance. Oversampling was performed using the SMOTE (synthetic minority over-sampling technique) algorithm, and insufficient draw data were oversampled relative to the number of win (or lose) data. The SMOTE algorithm is an over-sampling approach in which the minority class is over-sampled by creating synthetic examples rather than by over-sampling with replacement (Chawla et al., 2002). More specifically, SMOTE generates new synthetic examples along the lines between it and some randomly selected k nearest neighbors from the minority class (Maciejewski and Stefanowski, 2011). Note that in this system, the test data are not oversampled because the system is confined to predicting future games, and thus, only the training data of draw games are oversampled. In addition, experiments were conducted on the balance dataset, which includes oversampled data, and the imbalance dataset without oversampled data (see Appendix).

Based on the GB of the results presented in the appendix, the accuracy of the match result prediction including the draw exhibited a tendency to increase through oversampling, and the accuracy of GB was generally the best in experiments including the draw. If we simply compare the accuracy between the win–lose and win–draw–lose predictions, then the accuracy of the win–draw–lose prediction was significantly lower than that of the win–lose prediction. However, considering that the correct prediction probability (random guessing) of the win–draw–lose prediction is 33% and that the correct prediction probability (random guessing) of the win–lose prediction is 50%, the accuracy of the win–draw–lose prediction recorded in the appendix is not low.

In an empirical study, the performance of the *SWLPS* was measured by its accuracy, and the *SWLPS* adopted GB as supported by ANOVA. This study contributes to the literature by demonstrating a new match prediction system based on network indicators and demonstrating the usefulness of SNA for sports. Previous existing match result prediction systems used data from notational analysis, including the total number of passes, the number of fouls, the number of corner-kicks, and other such data. Under this condition, managers can be provided with relatively rough information only, such as "the need for a change in the number of cross-passes" or "the need for a change in the number of headers" when using existing game prediction systems for strategy composition. However, the proposed *SWLPS* can provide more specific information, such as the "problem of a specific player" or "the need to change passes between specific players", based on the PNI values generated using the PPC. Although the prediction result does not always match the actual result, this approach can help managers and supervisors consider an alternative perspective when establishing a strategy and developing training processes. Additionally, it is meaningful that the accuracy of the *SWLPS* is satisfactory and that the usefulness of SNA and machine learning in the sports field is revealed.

The main limitation of this study is that drawn games are excluded from the prediction. The number of drawn games was significantly smaller than the number of win or lose cases.

In additional experiments, we attempted to predict win–draw–lose by solving the imbalance of training data through oversampling, but the accuracy of *SWLPS* did not improve significantly because the imbalance in the testing data was not resolved. Moreover, although many other network indicators are available to implement SNA, the proposed system considers only seven network indicators related to individual and network structure measurements because we wanted to calculate team performance based on individual members of the organization and the network structure of the organization. The current work can be extended in future studies. For instance, additional network indicators, such as clustering coefficients and cliquishness, which indicate the statuses of sub-groups, might be considered. Additionally, a method to solve the imbalance of data caused by draw might be considered. If a method to solve the imbalance of testing data, which still exists after oversampling, the accuracy of win–draw–lose prediction might be improved with our framework.

Nomenclature

(SWLPS) soccer win–lose prediction system
(SNA) social network analysis

(GB) gradient boosting
(SVM) support vector machine
(NN) neural network
(DT) decision tree
(CBR) case-based reasoning
(LR) logistic regression
(CL) Champions League
(ANOVA) analysis of variance
(SGD) stochastic gradient descent
(ANIs) actual network indicators
(PNIs) predicted network indicators
(DC) degree centrality
(WDC) weighted degree centrality
(EC) eigenvector centrality
(BC) betweenness centrality
(CC) closeness centrality
(APL) average path length
(ND) network diameter
(PPC) predicted pass count
(UEFA) Union of European Football Association
(PDF) portable document format
(CSV) comma separated values
(SMOTE) synthetic minority oversampling technique

Appendix

See Table A.1.

References

Abbasi, A., Altmann, J., 2010. A social network system for analyzing publication activities of researchers. In: Bastiaens, T.J., Baumöl, U., Krämer, B. (Eds.), On Collective Intelligence. Springer, Berlin, pp. 49–61.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., Vespignani, A., 2004. The architecture of complex weighted networks. Proc. Natl. Acad. Sci. USA 101, 3747–3752.

Begum, S., Barua, S., Ahmed, M.U., 2014. Physiological sensor signals classification for healthcare using sensor data fusion and case-based reasoning. Sensors 14, 11770–11785.

Berson, A., Smith, S.J., 2002. Building Data Mining Applications for CRM. McGraw-Hill, Inc, New York.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. 2, 113–120.

Borgatti, S.P., 2005. Centrality and network flow. Soc. Netw. 27, 55–71.

Borgatti, S.P., Everett, M.G., 1997. Network analysis of 2-mode data. Soc. Netw. 19, 243–269.

Bose, I., Mahapatra, R.K., 2001. Business data mining—a machine learning perspective. Inform. Manag. 39, 211–225.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM, pp. 144–152.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Physica-Verlag HD, pp. 177–186.

Bourbousson, J., Poizat, G., Saury, J., Seve, C., 2010. Team coordination in basketball: Description of the cognitive connections among teammates. J. Appl. Sport Psy. 22, 150–166.

Bradley, P.S., Lago-Peñas, C., Rey, E., 2014. Evaluation of the match performances of substitution players in elite soccer. Int. J. Sports Physiol. Perform. 9, 415–424.

Chakroun, I., Haber, T., Ashby, T.J., 2017. SW-SGD: The sliding window stochastic gradient descent algorithm. Proce. Comput. Sci. 108, 2318–2322.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Chen, X.Y., Chau, K.W., 2016. A hybrid double feedforward neural network for suspended sediment load estimation. Water Res. Manag. 30, 2179–2194.

Chen, B., Chen, X., Li, B., He, Z., Cao, H., Cai, G., 2011. Reliability estimation for cutting tools based on logistic regression model using vibration signals. Mech. Syst. Signal Process. 25, 2526–2537.

Cho, Y.H., Kim, J.K., Kim, S.H., 2002. A personalized recommender system based on web usage mining and decision tree induction. Expert Syst. Appl. 23, 329–342.

Cioffi-Revilla, C., O'Brien, S.P., 2007. Computational analysis in us foreign and defense policy. In: First International Conference on Computational Cultural Dynamics. University of Maryland, College Park, MD.

Clemente, F.M., Couceiro, M.S., Martins, F.M.L., Mendes, R.S., 2015a. Using network metrics in soccer: A macro-analysis. J. Hum. Kinet. 45, 123–134.

Clemente, F.M., Martins, F.M.L., Kalamaras, D., Mendes, R.S., 2015b. Network analysis in basketball: inspecting the prominent players using centrality metrics. J. Phys. Educ. Sport 15, 212.

Cotta, C., Mora, A.M., Merelo, J.J., Merelo-Molina, C., 2013. A network analysis of the 2010 FIFA World Cup champion team play. J. Syst. Sci. Complexity 26, 21–42.

Cross, R., Borgatti, S.P., Parker, A., 2002. Making invisible work visible: Using social network analysis to support strategic collaboration. California Manage. Rev. 44, 25–46.

Elson, J., Tailor, A., Banerjee, S., Salim, R., Hillaby, K., Jurkovic, D., 2004. Expectant management of tubal ectopic pregnancy: prediction of successful outcome using decision tree analysis. Ultras Obstet. Gynecol. 23, 552–556.

Fransen, K., Van Puyenbroeck, S., Loughead, T.M., Vanbeselaere, N., De Cuyper, B., Vande Broek, G.V., Boen, F., 2015. Who takes the lead? Social network analysis as a pioneering tool to investigate shared leadership within sports teams. Soc. Netw. 43, 28–38.

Freeman, L.C., Roeder, D., Mulholland, R.R., 1979. Centrality in social networks: II. Experimental results. Soc. Netw. 2, 119–141.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. 55, 119–139.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 29, 1189–1232.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Statist. 28, 337–407.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906–914.

Galaskiewicz, J., Wasserman, S., 1993. Social network analysis. Sociol. Methods Res. 22, 3–22.

Gholami, V., Chau, K.W., Fadaee, F., Torkaman, J., Ghaffari, A., 2015. Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. J. Hydrol. 529, 1060–1069.

Gonçalves, B., Esteves, P., Folgado, H., Ric, A., Torrents, C., Sampaio, J., 2017. Effects of pitch area-restrictions on tactical behavior, physical, and physiological performances in soccer large-sided games. J. Strength Cond. Res. 31, 2398–2408.

Gréhaigne, J.F., Bouthier, D., David, B., 1997. Dynamic-system analysis of opponent relationships in collective actions in soccer. J. Sports Sci. 15, 137–149.

Griffin, A., Hauser, J.R., 1992. Patterns of communication among marketing, engineering and manufacturing—A comparison between two new product teams. Manag. Sci. 38, 360–373.

Grund, T.U., 2012. Network structure and team performance: the case of English Premier League soccer teams. Soc. Netw. 34, 682–690.

Guelman, L., 2012. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst. Appl. 39, 3659–3667.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learn. 46, 389–422.

Halicioğlu, R.F., 2011. Can we predict the outcome of the international football tournaments?: the case of euro 2000. Doğuş Üniversitesi Dergisi 6, 112–122.

Huang, K.Y., Chang, W.L., 2010. A neural network method for prediction of 2006 world cup football game. In: The 2010 International Joint Conference on Neural Networks, IJCNN. IEEE, pp. 1–8.

Hughes, M., Franks, I.M., 2004. Notational Analysis of Sports, second ed. Routledge, London, pp. 101–105.

Hughes, M., Franks, I., 2005. Analysis of passing sequences, shots and goals in soccer. J. Sports Sci. 23, 509–514.

Pallant, J., 2013. SPSS Survival Manual. McGraw-Hill Education, London, UK.

James, N., 2006. Notational analysis in soccer: past, present and future. Int. J. Perform. Anal. Spor 6, 67–81.

Jang, G.S., Lai, F., Jiang, B.W., Parng, T.M., Chien, L.H., 1993. Intelligent stock trading system with price trend prediction and reversal recognition using dual-module neural networks. Appl. Intell. 3, 225–248.

Jo, H., Han, I., Lee, H., 1997. Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. Expert Syst. Appl. 13, 97–108.

Joachims, T., 1999. Transductive inference for text classification using support vector machines. ICML 99, 200–209.

Johnson, R., Zhang, T., 2013. Accelerating stochastic gradient descent using predictive variance reduction. In: Tesauro, G., Touretzky, D, Leed, T. (Eds.), Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA, pp. 315–323.

Kim, J., Hastak, M., 2018. Social network analysis: Characteristics of online social networks after a disaster. Int. J. Inform. Manag. 38, 86–96.

Kim, J.W., Lee, B.H., Shaw, M.J., Chang, H.L., Nelson, M., 2001. Application of decision-tree induction techniques to personalized advertisements on internet storefronts. Int. J. Electron. Com. 5, 45–62.

Koning, R.H., 2000. Balance in competition in dutch soccer. J. R. Stat. Soc. D 49, 419–431.

Koning, R.H., Koolhaas, M., Renes, G., Ridder, G., 2003. A simulation model for football championships. European J. Oper. Res. 148, 268–276.

Lago, C., Martín, R., 2007. Determinants of possession of the ball in soccer. J. Sports Sci. 25, 969–974.

Lago-Ballesteros, J., Lago-Peñas, C., 2010. Performance in team sports: identifying the keys to success in soccer. J. Hum. Kinet. 25, 85–91.

Lee, S., Enke, D., Kim, Y., 2017. A relative value trading system based on a correlation and rough set analysis for the foreign exchange futures market. Eng. Appl. Artif. Intell. 61, 47–56.

Lee, S.H., Kim, P.J., Jeong, H., 2006. Statistical properties of sampled networks. Phys. Rev. E 73, 016102.

Li, Z., Wang, E., Ou, J., Liu, Z., 2015. Hazard evaluation of coal and gas outbursts in a coal-mine roadway based on logistic regression model. Int. J. Rock Mech. Min. Sci. 80, 185–195.

Li, X., Yeh, A.G., 2004. Multitemporal sar images for monitoring cultivation systems using case-based reasoning. Remote Sens. Environ. 90, 524–534.

Lusa, L., 2017. Gradient boosting for high-dimensional prediction of rare events. Comput. Statist. Data Anal. 113, 19–37.

Ma, H., Wang, Y., Wang, K., 2018. Automatic detection of false positive RFID readings using machine learning algorithms. Expert Syst. Appl. 91, 442–451.

Maciejewski, T., Stefanowski, J., 2011. Local neighbourhood extension of SMOTE for mining imbalanced data. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining, CIDM. IEEE, pp. 104–111.

Nascimento, M.A., Sander, J., Pound, J., 2003. Analysis of SIGMOD's co-authorship graph. SIGMOD Rec. 32, 8–10.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorobot. 7, 21.

Newman, M.E., 2004. Analysis of weighted networks. Phys. Rev. E 70, 056131.

Nixon, H.L., 1993. Social network analysis of sport: emphasizing social structure in sport sociology. Sociol. Sport J. 10, 315–321.

Ohlmacher, G.C., Davis, J.C., 2003. Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. Eng. Geol. 69, 331–343.

Opsahl, T., Agneessens, F., Skvoretz, J., 2010. Node centrality in weighted networks: generalizing degree and shortest paths. Soc. Netw. 32, 245–251.

Osowski, S., Hoai, L.T., Markiewicz, T., 2004. Support vector machine-based expert system for reliable heartbeat recognition. IEEE Trans. Biomed. Eng. 51, 582–589.

Otte, E., Rousseau, R., 2002. Social network analysis: a powerful strategy, also for the information sciences. J. Inform. Sci. 28, 441–453.

Passos, P., Davids, K., Araújo, D., Paz, N., Minguéns, J., Mendes, J., 2011. Networks as a novel tool for studying team ball sports as complex social systems. J. Sci. Medi. Sport 14, 170–176.

Pena, J.L., Touchette, H, 2012. A network theory analysis of football strategies. ArXiv preprint arXiv:1206.6904.

Persson, C., Bacher, P., Shiga, T., Madsen, H., 2017. Multi-site solar power forecasting using gradient boosted regression trees. Sol. Energy 150, 423–436.

Pollard, R., 2002. Charles reep (1904-2002): pioneer of notational and performance analysis in football. J. Sports Sci. 20, 853–855.

Raj, S., Luthra, S., Ray, K.C., 2015. Development of handheld cardiac event monitoring system. IFAC-Papers OnLine 48, 71–76.

Razanamahandry, L.C., Andrianisa, H.A., Karoui, H., Podgorski, J., Yacouba, H., 2018. Prediction model for cyanide soil pollution in artisanal gold mining area by using logistic regression. CATENA 162, 40–50.

Reed, D., O'Donoghue, P., 2005. Development and application of computer-based prediction methods. Int. J. Perform. Anal. Sport 5, 12–28.

Reilly, T.H., Thomas, V., 1976. A motion analysis of work-rate in different positional roles in professional football match-play. J. Hum. Mov. Stud. 2, 87–97.

Riesbeck, C.K., Shank, R.C., 1989. Inside Case-Based Reasoning. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rotshtein, A.P., Posner, M., Rakityanskaya, A.B., 2005. Football predictions based on a fuzzy model with genetic and neural tuning. Cybern. Syst. Anal. 41, 619–630.

Schank, R.C., 1983. Dynamic Memory: A Theory of Reminding and Learning in Computers and People. Cambridge University Press, Cambridge.

Scoulding, A., James, N., Taylor, J., 2004. Passing in the soccer world cup 2002. Int. J. Perform. Anal. Sport 4, 36–41.

Sefeedpari, P., Rafiee, S., Akram, A., Chau, K.W., Pishgar-Komleh, S.H., 2016. Prophesying egg production based on energy consumption using multi-layered adaptive neural fuzzy inference system approach. Comput. Electron. Agric. 131, 10–19.

Shamir, O., Zhang, T., 2013. Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. Proc. Mach. Learn. Res. 28, 71–79.

Snyder, J.A.L., 2013. What Actually Wins Soccer Matches: Prediction of the 2011–2012 Premier League for Fun and Profit (Doctoral dissertation. Master's thesis), University of Washington.

Specht, D.F., 1991. A general regression neural network. IEEE Trans. Neural Netw. 2, 568–576.

Sueur, C., Jacobs, A., Amblard, F., Petit, O., King, A.J., 2011. How can social network analysis improve the study of primate behavior? Am. J. Primatol. 73, 703–719.

Sugumaran, V., Ramachandran, K.I., 2007. Automatic rule learning using decision tree for fuzzy classifier in fault diagnosis of roller bearing. Mech. Syst. Signal Proc. 21, 2237–2247.

Swaminathan, H., Rogers, H.J., 1990. Detecting differential item functioning using logistic regression procedures. J. Educ. Meas. 27, 361–370.

Taormina, R., Chau, K.W., Sivakumar, B., 2015. Neural network river forecasting through baseflow separation and binary-coded swarm optimization. J. Hydrol. 529, 1788–1797.

Touzani, S., Granderson, J., Fernandes, S., 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy Build. 158, 1533–1543.

Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 49, 1225–1231.

Vapnik, V.N., Vapnik, V., 1998. Statistical Learning Theory, Vol. 1. Wiley, New York.

Vijayvargiya, A., 2009. One-way analysis of variance. J. Valid. Technol. 15, 62–64.

Vlahou, A., Schorge, J.O., Gregory, B.W., Coleman, R.L., 2003. Diagnosis of ovarian cancer using decision tree classification of mass spectral data. Biomed. Res. Int. 2003, 308–314.

Wang, W.C., Xu, D.M., Chau, K.W., Chen, S., 2013. Improved annual rainfall-runoff forecasting using PSO-SVM model based on EEMD. J. Hydroinformatics 15, 1377–1390.

Warner, S., Bowers, M.T., Dixon, M.A., 2012. Team dynamics: A social network perspective. J. Sport Manag. 26 (1), 53–66.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications, Vol. 8. Cambridge University Press, Cambridge.

Wey, T., Blumstein, D.T., Shen, W., Jordán, F., 2008. Social network analysis of animal behaviour: a promising tool for the study of sociality. Anim. Behav. 75, 333–344.

Wolf, T., Schroter, A., Damian, D., Nguyen, T., 2009. Predicting build failures using social network analysis on developer communication. In: Proceedings of the 31st International Conference on Software Engineering. IEEE Computer Society, pp. 1–11.

Wu, D., 2009. Supplier selection: a hybrid model using DEA, decision tree and neural network. Expert Syst. Appl. 36, 9105–9112.

Yang, H.L., Tang, J.H., 2004. Team structure and team performance in IS development: a social network perspective. Inform. Manag. 41, 335–349.

Zhang, T., 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the Twenty-First International Conference on Machine Learning. ACM, p. 116.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transp. Res. C Emerg. Technol. 58, 308–324.