Advanced Football Match Winning Probability Prediction: A CNN-BiLSTM_Att Model with Player Compatibility and Dynamic Lineup Analysis

Tao Quan*, Yingling Luo
College of Physical Education and Health Science, Chongqing Normal University, Chongqing, 401331, China

Abstract—**In recent years, with the continuous expansion of the football market, the prediction of football match-winning probabilities has become increasingly important, attracting numerous professionals and institutions to engage in the field of football big data analysis. Pre-match data analysis is crucial for predicting match outcomes and formulating tactical strategies, and all top-level football events rely on professional data analysis teams to help teams gain an advantage. To improve the accuracy of football match winning probability predictions, this study has taken a series of measures: using the Word2Vec model to construct feature vectors to parse the compatibility between players; developing a winning probability prediction model based on LSTM to capture the dynamic changes in team lineups; designing an improved BILSTM_Att winning probability prediction model, which distinguishes the different impacts of players on match outcomes through an attention mechanism; and proposing a CNN-BILSTM_Att winning probability prediction model that combines the local feature extraction capability of CNN with the time series analysis of BILSTM. These research efforts provide more refined data support for football coaching teams and analysts. For the general audience, these in-depth analyses can help them understand the tactical layouts and match developments on the field more deeply, thereby enhancing their viewing experience and understanding of the matches.**

Keywords*—*Football big data; match prediction; feature vector; tactical understanding; match analysis

I. Introduction

Over the past few decades, football has emerged as one of the most popular sports globally, generating significant interest among fans while also attracting the attention of sports scientists, data analysts, and technical researchers [1] [2]. With advancements in technology and the rise of big data, data analysis of football matches has become an expanding research area within competitive sports [3] [4]. Accurate predictions of match winning probabilities provide a scientific basis for team tactics and enhance fans' understanding of the games. Consequently, developing a high-accuracy football match winning probability prediction model holds substantial value for both tactical analysis and audience engagement.

In this context, research on football match analysis technology has progressed. However, the inherent complexity of football matches, characterized by uncertainty and randomness, presents considerable challenges in predicting match outcomes. This study aims to explore and propose innovative data analysis methods to enhance the accuracy of football match-winning probability predictions and facilitate a deeper understanding of tactical layouts and match developments.

Initially, the focus is placed on analyzing player compatibility, as relationships among individual players influence overall team performance. Natural language processing technologies, such as Word2Vec, are employed to generate player feature vectors, which capture potential interactions at the data level and provide foundational features for subsequent winning probability predictions.

Secondly, to account for dynamic factors in football matches, a model based on Long Short-Term Memory (LSTM) networks is introduced to handle time-series data, capturing changes in team lineups and player performance over time. The memory capabilities of LSTM enable the model to learn critical moments during matches and assess the impact of tactical adjustments on outcomes.

Thirdly, to more precisely evaluate the influence of different players on match outcomes, an attention mechanism is integrated into the Bidirectional LSTM (BILSTM) model, resulting in the BILSTM_Att model. The attention mechanism supplies additional information regarding player importance, thereby enhancing prediction accuracy and personalization.

Finally, the study combines the feature extraction capabilities of Convolutional Neural Networks (CNN) with the time-series analysis strengths of BILSTM to propose the CNN-BILSTM_Att model. The CNN layer extracts spatial features of team lineups, while the BILSTM layer processes the evolution of these features over time, with the attention mechanism facilitating the combination of features for more accurate winning probability predictions.

The remainder of the paper is organized as follows: Section I introduces the background and significance of football match winning probability prediction; Section II reviews the literature on prediction methods in competitive sports, developments in natural language processing, and the current state of football match outcome prediction research; Section III discusses data preprocessing, word vector generation, and the development of the CNN-BILSTM_Att winning probability prediction model; Section IV presents a case study to validate the proposed method using a custom experimental dataset; and Section V concludes the paper by summarizing the findings and discussing potential areas for future research.

*Corresponding author

II. LITERATURE REVIEW

This paper will collect works related to existing winning probability predictions to highlight the shortcomings of current research.

A. *Analysis Methods of Winning Probability Prediction in Competitive Sports*

In the field of winning probability prediction research for competitive sports, the progress of deep learning has become a driving force for transformation [5] [6]. The target detection algorithms based on deep convolutional neural networks, as studied by Mukherjee et al. [7], have now become powerful tools for analyzing players and game dynamics. These algorithms learn complex features from massive amounts of data, helping us understand game development and have demonstrated high accuracy on multiple standard datasets.

To meet the needs of real-time analysis in fast-paced sports, researchers like McGarry et al. [8] have been exploring methods to accelerate target detection algorithms, successfully reducing response times by optimizing network architecture and computational efficiency. This is particularly crucial for winning probability prediction, as instant game information can have a significant impact on the outcome.

However, a major challenge in winning probability prediction is the detection of small targets, such as tracking the position of a soccer ball in a football match or monitoring the movements of distant players in a basketball game. Hodge et al. [9] proposed a series of solutions to this challenge, including multi-scale detection methods and the integration of contextual information, which have been proven to improve the recognition rate of small targets.

The use of ensemble learning has provided another breakthrough for winning probability prediction. The work of Chakraborty et al. [10] shows that by combining the results of multiple prediction models, the accuracy and robustness of predictions can be effectively improved. In sports analysis, ensemble methods can synthesize different data sources and algorithms, such as player statistics, team dynamics, historical performance, etc., to form a more comprehensive prediction of winning probabilities.

These technological advances have provided sports analysts, coaches, and betting companies with unprecedented data insights, helping them make more informed and strategic decisions [11]. They also provide sports fans with a richer viewing experience, such as increasing the interactivity and participation of matches through real-time data analysis [12]. As research continues to deepen, we can expect future winning probability prediction methods to become more refined and personalized, possibly using machine learning models to predict individual player performances or simulate match outcomes more accurately. These research outcomes indicate that target detection algorithms are developing in the direction of being more accurate, faster, and more robust.

B. *The Current State of Natural Language Processing Development*

Natural Language Processing (NLP) has become a core branch in the field of artificial intelligence, and its current development status highlights significant progress in understanding and generating human language. Particularly, transformative models such as BERT, proposed by Devlin et al., have revolutionized the ability to deeply comprehend semantic meaning in text by introducing bidirectional training mechanisms. These models have also had a revolutionary impact on a variety of NLP tasks including text classification, question-answering systems, and machine translation [13]. In the specific application scenario of sports competition prediction, the advancement of NLP technology is equally noteworthy.

By utilizing advanced text analysis technologies, researchers have been able to extract key information from live sports commentary, aiding in predicting the pace and outcome of games [14]. Furthermore, the application of sentiment analysis technology has made it possible to identify emotions from fans' social media activity, thereby analyzing the correlation between these emotions and game outcomes. Going further, NLP-based methods are being used to understand the patterns of language use among athletes, with the aim of predicting their performance and mental state, and even potential injury risks [15].

Modern NLP technology is also playing a role in automating the parsing of athletes' injury reports, reducing the workload of analysts in data collection and processing [16]. Additionally, semantic analysis of historical game reports can reveal long-term performance trends of teams and players, providing data support for establishing more accurate prediction models [17].

With the continual evolution of pre-trained models, such as the GPT series' capabilities in generating text, NLP has been used to automatically generate sports news articles, pre-game analyses, and even to simulate interviews with coaches and players, bringing new possibilities to media creation and fan interaction [18]. At the same time, these technologies are also helping to create more intelligent virtual assistants that can provide real-time game updates, statistical data interpretation, and personalized sports consumption advice.

In summary, the application of NLP in the field of sports competition prediction indicates that this technology has not only made significant progress in understanding complex text and language patterns but is also bringing profound changes to sports analysis, fan experience, and media reporting. With further research and the refinement of technology, NLP is expected to have an even greater impact in the field of sports in the future.

C. *Current Status of Research on Football Match Outcome Prediction*

Research on football match outcome prediction involves predicting results by analyzing team and player performance data, match conditions, and other relevant factors. In recent years, this research field has benefited from the rapid development of big data analytics, machine learning, and artificial intelligence technologies, transitioning from traditional statistical methods to more complex and refined predictive models.

With the aid of deep learning technologies, football match outcome prediction models can more effectively process and analyze complex datasets, learn features extracted from historical data, and predict future match outcomes [19]. Some research has adopted structures similar to deep convolutional neural networks, which identify and utilize potential factors affecting match outcomes by learning from a large volume of historical match data [20]. To meet the requirements of real-time prediction, researchers are also striving to improve the computational efficiency of models by optimizing network structures and designing more efficient algorithms, thus reducing the time required for prediction [21]. This enables models to provide predictions rapidly without sacrificing accuracy, satisfying the real-time needs of scenarios such as live match analysis and online betting [22]. In dealing with uncertainties and complex scenarios, such as player absences and weather changes, current research is attempting to incorporate more contextual information and a deeper understanding of the match environment to enhance the robustness of models in the face of varied real game situations [23]. At the same time, football match outcome prediction is focusing on the integrated analysis of multidimensional data, such as players' spatial positioning and team tactical changes. By integrating these multidimensional pieces of information, prediction models can analyze match situations more comprehensively, thus providing more precise predictions [24].

Overall, research on football match outcome prediction has made significant progress in prediction accuracy, processing speed, and adaptation to complex match environments through the adoption of advanced data analysis techniques and algorithmic models. As research continues to deepen and technology continues to develop, this field is expected to achieve higher levels of predictive performance, providing stronger support for football match analysis and related areas.

D. Research Gaps and Future Research Directions

In this research field, we can identify some potential research gaps that may provide directions for future work:

1) Quantification of player psychological states: Current models primarily analyze player performance and compatibility through statistics and machine learning techniques, but players' psychological states also have a significant impact on match outcomes. Integrating players' mentality, stress responses, and other psychological factors into models to predict match outcomes more comprehensively is a research gap that needs to be filled.

2) Comprehensive analysis of environmental factors in the stadium: Although research has considered player compatibility and dynamic performance, factors related to the stadium and environment (such as climate, ground conditions, fan support) are equally important to player performance and match outcomes. Future research could consider incorporating these environmental factors into models to enhance the accuracy and practicality of predictions.

3) In-depth integration of opponent analysis: The tactics of the opposing team, the state of their players, and team compatibility also affect match outcomes, but current models may not fully consider this aspect. Developing models that can analyze the characteristics of both teams and predict the outcomes of their interactions will be a valuable research direction.

The above issues are the key to improving prediction accuracy, and they are also the content of this article's research

III. Data Preprocessing and Word Vectors

A. Collection of Match Data

To ensure the rationality and usability of data, this paper has crawled the historical match records of several strong teams through Whoscored, and obtained the player line-ups used in the matches from these records. The flow diagram of the crawling process is shown in Fig. 1.

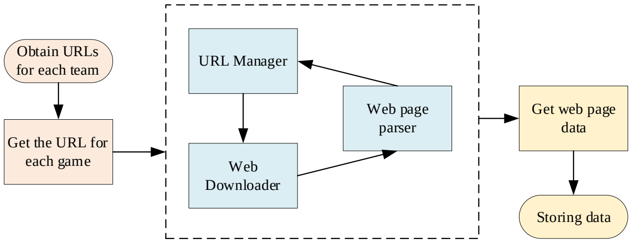

Fig. 1. Match data crawling process.

In the upcoming football events, every coach and fan tries to predict the outcomes of matches by analyzing the line-ups of the teams. Although the performance of the players and the real-time progress of the match are unpredictable and non-generalizable factors, by constructing a reliable pre-match line-up analysis model, we can provide valuable strategic support for the coaching team and also enhance the match experience for fans.

For this analysis, we have already obtained a large amount of line-up data from the database, covering detailed line-up information for 42,596 football matches. To make these data useful, they must first undergo thorough preprocessing. We will filter out information that has potential influence on match results and integrate it to form an efficient match line-up database.

Considering practicality and universality, our model will focus on the following key points:

- Team Strength Balance: Assessing the overall strength of both teams, including the technical statistical data and historical performance of the players.
- Tactical Adaptability: Analyzing the tactical flexibility of each team and how to make corresponding tactical adjustments based on the opponent's line-up.
- Player Condition: Evaluating the current state of key players, including injury conditions and athletic form.
- Historical Match Records: The historical head-to-head records of the two teams are also a factor that cannot be ignored, as it often reveals psychological advantages and disadvantages.

Through the comprehensive analysis of these dimensions, our model aims to provide deeper insights for the upcoming matches, thereby helping coaching teams develop more accurate tactical layouts and bringing richer viewing perspectives to fans.

As the date of the match approaches, we will continue to monitor changes in relevant variables and update our analysis model in a timely manner, ensuring that all predictions are based on the most recent and comprehensive data. We look forward to verifying the accuracy of our model and hope that it will become a powerful tool for predicting football match outcomes.

B. Data Processing

The football match data obtained through web crawling cannot be used directly for lineup analysis; it requires preprocessing of the raw dataset. The main steps of data preprocessing in this article include: data cleaning, data reduction, feature construction, and data annotation.

Since every football player has different positions and technical characteristics, these attributes create special effects that establish interconnections among players. In the match lineup dataset, players are not selected randomly; based on the analysis of the relationships between players, the coaching team selects a set of lineups that can cooperate with each other. In professional matches, the data analysis team will choose a lineup that has strong synergy and can effectively counter the opponent's players.

The interrelationships among players can be summarized as compatibility relationships and counter relationships, both of which jointly influence the outcome of a match. As shown in Fig. 2, the player word vectors in this paper consist of two parts: the first part is the word vectors with compatibility relationships generated by the Word2Vec model, and the second part is the extended vectors generated by the counter relationship algorithm. The composition of the player word vectors is illustrated in the Fig. 2.

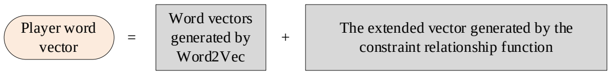
Player word vector
=
Word vectors generated by Word2Vec
+
The extended vector generated by the constraint relationship function

Fig. 2. Composition of word vectors.

1) Acquisition of training corpus for word vectors: This article has crawled the home team lineup selection data from 54,652 professional football matches through Wan Plus eSports, using this data to train the Word2Vec model, and these data exclude the previously obtained 42,596 matches. In professional football matches, the lineups chosen after analysis by the coaching team have good compatibility, and the relationship among players within the lineup is strong. Therefore, choosing the lineup of one side in a professional match as the corpus, the Word2Vec model can generate player word vectors with relatively clear relevance. Some data from the corpus are shown in the table.

In this table, NO. represents the sequence number of the match, and P1, P2, P3, P4, P5 represent five different positions, namely [goalkeeper, defender, midfielder, support, forward]. The feature values of these five characteristics indicate the players selected for the corresponding positions. The CBOW model in Word2Vec can predict a word based on the context of that word in the text. During the training process, word vectors are generated as a byproduct, converting text content into a form of a numerical matrix. While generating word vectors, the CBOW model can also calculate the relevance of these vectors. Therefore, based on this feature of the CBOW model, the names representing players are treated as input text information, and the word vectors generated by the CBOW model become the player word vectors. The five different players chosen by the home team in the corpus can thus be seen as a sentence composed of five different words, and the degree of relevance between these words is the degree of compatibility of the players. The structure of the model used to train the word vectors is shown in the Fig. 3.

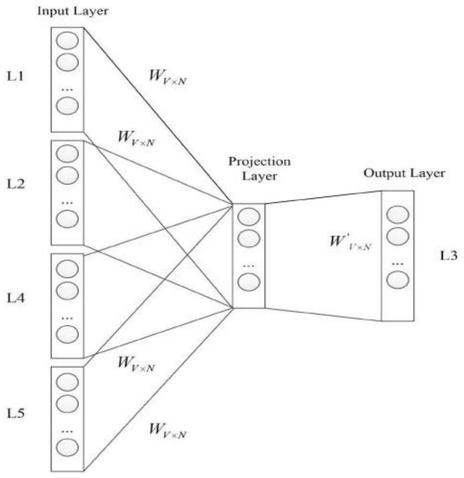

Fig. 3. The model structure for training word vectors.

In this model, the word vector of the target vocabulary is calculated based on the word vectors of the context before and after the target word. In this article, taking the name of the midfielder in the corpus as the target vocabulary (the midfield position usually corresponds to the center of the football formation, similar to the mid-lane position L3 in League of Legends), the preceding vocabulary of the midfielder is the defender (the defender position corresponds to the feature values of features L1 and L2), and the following context is the forward and support (the forward and support positions correspond to the feature values of features L4 and L5). Assuming that the dimension of the word vector is N and the length of the vocabulary is V, the target function of the training model is shown as follows.

$$L = \sum_{L \in V} \log(P(L3 \mid L1, L2, L4, L5))$$

Input layer: First, each player name is encoded using one-hot encoding, for example, player A's one-hot encoding is [1,0,0,0,…]. For convenience of calculation, x_{i-2}、 x_{i-1}、 x_i、 x_{i+1}、 x_{i+2} are used to represent the one-hot encodings of the names of feature players P1, P2, P3, P4, P5, which correspond to the positions [goalkeeper, defender, midfielder, support, forward]. Then all one-hot encodings are multiplied by the input weight matrix, as shown below.

$$I_1 = W_{V\times N} \times x_{i-1}$$
$$I_2 = W_{V\times N} \times x_{i-2}$$
$$I_4 = W_{V\times N} \times x_{i+1}$$
$$I_5 = W_{V\times N} \times x_{i+2}$$

Hidden layer: Average all word vectors to get the hidden vector, as shown in equation

$$\hat{I} = \frac{I_1 + I_2 + I_4 + I_5}{4}$$

Output layer: Multiply the hidden vector of the hidden layer by the output weight, as shown in equation.

$$z = W'_{V\times N}\hat{I}$$

Then use a softmax classifier to obtain the probability that each hero in the hero library is predicted as the current result, as shown in equation.

$$\hat{y} = \mathrm{softmax}(z)$$

In this model training process, the weight from the input layer to the hidden layer is the word vector. The model can calculate the similarity probability between players, and the higher the similarity probability between two players, the stronger their relevance, indicating a strong pairing relationship between them. For example, if a midfielder (such as Messi) often appears with a forward player (such as Suarez), their word vector distance will be very small after generating the word vectors.

2) Acquisition of word vectors: This article uses the Gensim toolkit to import the Word2Vec model. Gensim is an open-source Python toolkit that supports multiple topic modeling algorithms such as TF-IDF, LDA, and Word2Vec. During the process of training word vectors with the Word2Vec model, the model's parameters are set as follows:

The sg parameter is the option to select the training mode. When sg=0, the model uses the CBOW (Continuous Bag of Words) model to train word vectors; when sg=1, the model uses the Skip-Gram model to train word vectors. This article uses the CBOW model to train player word vectors;

The size parameter indicates the dimension of the word vectors produced. In small datasets, this is usually set between 100-200. After multiple experiments, this article selects size=100;

The window parameter indicates the maximum possible distance between the current word and the predicted word. The larger the window parameter is set, the more predicted words need to be enumerated. Since there are five players input at a time in the training corpus, this parameter is set to 5;

The min_count parameter indicates that if a word appears in the entire dataset less than the set value, the word will be directly ignored. Since there is enough competition data in this article and each player has many match records, this parameter is set to 0.

The following are some player vectors after training through the Word2Vec model, for ease of representation, the player names in this article are represented as Player(name):

Player("Messi") = [0.25237, -0.10362, 0.75654, …]

Player("Cristiano Ronaldo") = [0.46895, 0.55863, -0.05965, …]

Player("Neymar") = [-0.75294, -0.63594, 0Sure!]

Compatibility can be calculated through cosine similarity, which reflects the degree of approximation of players in the semantic space. To some extent, this can be mapped to their ability to cooperate in actual games. Based on actual game experience, in football matches, among the players with the highest match compatibility with a key player (such as Messi), most are midfielders or wing-backs who can provide support and assists. These players are able to pass the ball well to Messi and create opportunities for him to score. When Messi receives effective support and has enough space to shoot, his team is more likely to win the game. Therefore, in professional games, Messi appears more frequently with these supporting players, indicating a strong compatibility between them.

Through the analysis of Messi's word vector compatibility, it can be seen that the Word2Vec model used in this paper produces word vectors that well reflect the compatibility between players in the team lineup. Players who often appear with Messi in games have a high degree of compatibility; those who have difficulty appearing with Messi have a lower degree of compatibility. This kind of analysis and discovery provides deep insights into the interaction between players in football matches and team cooperation strategies.

For example, if Messi and a specific midfielder often appear together in actual games and usually can cooperate to lead the team to victory, we can expect that in the word vector space produced by Word2Vec, Messi and this midfielder will have a higher degree of similarity or proximity. Such information is very valuable for the formulation of football strategies, especially when selecting player lineups and devising game tactics.

3) Expansion of word vectors: In football match prediction, not only is there a chemical reaction between teams, but also a suppression relationship. To better reflect the characteristics of the team, this paper uses the suppression relationship to generate a suppression vector to expand the team vector based on statistical data. Thus, the final team vector includes both the chemical reaction between teams and their suppression relationships, making the information contained in the team vector more comprehensive.

The suppression relationship between teams is generally determined based on the experience of fans and professional analysts in actual games. For example, most analyses might find that Team_A is limited by Team_B's tactics when evaluating the match between the two. This kind of suppression relationship is difficult to represent with specific numbers or vectors. This paper studies the win rate data between teams and finds that the suppression relationship can be reflected in the team's win rate.

On football data platforms, such as Opta, WhoScored, or Transfermarkt, one can obtain the overall win rate of each team. It can be seen that in a certain season, Arsenal's overall win rate is 60%, while Chelsea's overall win rate is 55%. The total win rate of these two teams is not much different, and usually, the probability of victory for either team in their matches is close to 50%. However, in actual games, Arsenal may find it difficult to compete with Chelsea, and by analyzing the data, it can be found that Arsenal's actual win rate against Chelsea may only be 40%. Therefore, by comparing the expected win rate and the actual head-to-head win rate of the teams, one can determine whether there is a suppression relationship.

Through the analysis of different teams' win rates and head-to-head win rates, the suppression relationship can be quantified through the following calculation process as shown in Fig. 4, the algorithm flow is as follows:

- Collect historical head-to-head data for Team_A and Team_B, including win rates and head-to-head win rates.
- Calculate the average win rate of the two teams as a reference value for the expected win rate.
- Use the difference between the actual head-to-head win rate and the expected win rate to assess the strength of the suppression relationship.
- Convert this difference into a suppression vector and combine it with the team vector.
- Use the expanded team vector for more accurate match predictions.

By this method, we can more comprehensively understand the competitive relationship between teams and consider more variables when predicting future game outcomes.

First, calculate the global win rate p_i for the teams in the dataset using the following Formula:

$$p_i = \frac{N_i}{M_i}$$

In this formula, i represents any team, N_i represents the number of matches won by the team, and M_i represents the total number of matches the team has participated in. Similarly, the global win rate p_j for the team j that team i is facing can be calculated; By using the global win rates of both teams, the expected win probability when the teams face each other can be calculated using the following formula:

$$E_{ij} = \frac{p_i - p_i \times p_j}{p_i + p_j - 2 \times p_i \times p_j}$$

In this formula, E_{ij}represents the expected win rate when team i faces team j;

Calculate the actual win rate $p'_{i,j}$ of the teams' encounters using the existing match dataset.

$$p'_{i,j} = \frac{n_{ij}}{m_{ij}}$$

In this formula, n_{ij} represents the number of matches won by team i against team j, and m_{ij} represents the total number of matches played between team i and team j;

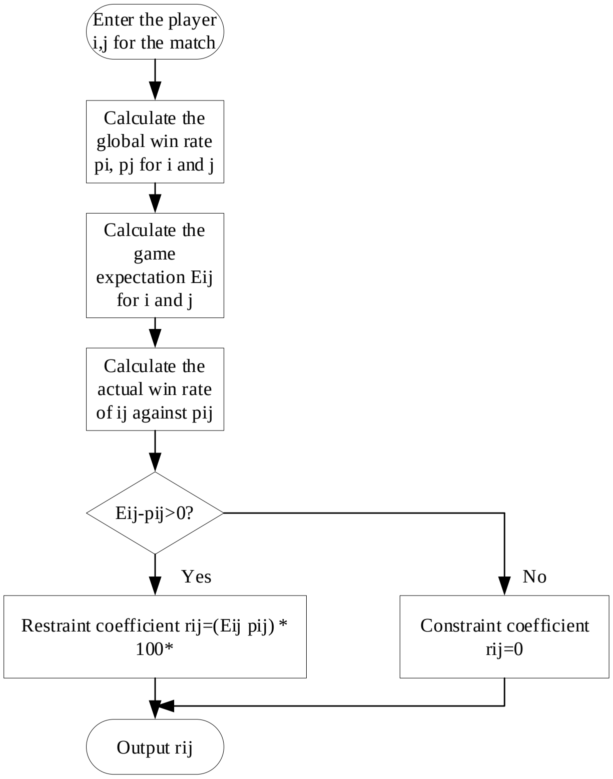
Enter the player i,j for the match
Calculate the global win rate pi, pj for i and j
Calculate the game expectation Eij for i and j
Calculate the actual win rate of ij against pij
Eij-pij>0?
Yes
No
Restraint coefficient rij=(Eij pij) * 100*
Constraint coefficient rij=0
Output rij

Fig. 4. Suppression relationship algorithm flow.

The specific algorithm process is as follows:

By comparing the expected win rate of the encounter between team i and team j with the actual win rate, if the actual win rate is less than the expected win rate, that is, $E_{ij} - p'_{i,j} > 0$,, then it can be considered that team j has a suppressing effect on team i;

After determining the suppression relationship between teams, further quantify the suppression relationship to be input into the prediction model. The win rate difference can be used as the basis for the suppression relationship coefficient. The calculation formula for the suppression relationship coefficient is as follows:

$$r_{ij} = \left(E_{ij} - p'_{i,j}\right) \times 100\%$$

In this formula, r_{ij} represents the coefficient of the suppression relationship between teams. The difference between the expected win rate and the actual win rate represents the suppression relationship index. The larger the value of r_{ij}, the stronger the suppressive effect of team j on team i.

Calculate the suppression relationship coefficient for each team and its opposing teams one by one. If there is no suppression relationship, the suppression relationship coefficient is set to 0. In this way, the suppression relationship vector for each team can be obtained.

In order to facilitate the generation of suppression relationship vectors between teams, an algorithm for generating suppression relationship vectors has been constructed. The logical structure of this algorithm is visualized as follows:

- Collect and process match data.
- Calculate the global win rate of the teams.
- Calculate the expected win rate of the encounters.
- Calculate the actual win rate of the encounters.
- Assess and calculate the suppression relationship coefficient.
- Generate the suppression relationship vector for the teams.

Through these steps, a structured dataset can be obtained, which is used to improve the accuracy of football match outcome predictions.

To facilitate the generation of expanded vectors for hero suppression relationships, this paper has constructed an algorithm for generating expanded vectors. Fig. 5 shows the logical structure of this algorithm.

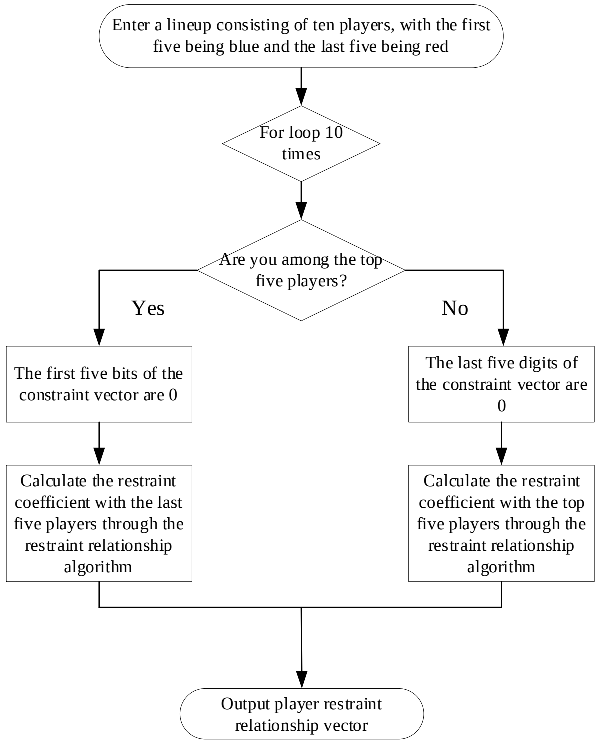
Enter a lineup consisting of ten players, with the first five being blue and the last five being red
For loop 10 times
Are you among the top five players?
Yes
No
The first five bits of the constraint vector are 0
The last five digits of the constraint vector are 0
Calculate the restraint coefficient with the last five players through the restraint relationship algorithm
Calculate the restraint coefficient with the top five players through the restraint relationship algorithm
Output player restraint relationship vector

Fig. 5. The logical structure of the expanded vector generation algorithm.

C. CNN-Enhanced CNN-BILSTM_Att Win Probability Prediction Model

In the field of football match prediction, due to the multitude of factors involved in a game and the complex relationships between them, reliance solely on traditional statistical data and historical records for predictions is limited. For example, the dimension of "head-to-head history" is often affected by various factors, and its degree of influence on the outcome of the match is not easily quantifiable. Current football match prediction methods vary, including those based on odds, subjective predictions based on historical experience, etc., but there is a lack of a unified data mining-based quantitative analysis process to reveal the intrinsic connections between features, which poses obstacles to subsequent match analysis and prediction.

To address this issue, the BILSTM_Att prediction model adopts a Bi-directional Long Short-Term Memory network combined with an attention mechanism to effectively extract the characteristic information of team line-ups. However, the BILSTM_Att model may not fully capture local information when processing input features, and the model fitting process can be slow due to the large number of weight parameters in the hidden layers. To overcome these shortcomings, we propose an improved prediction model: CNN-BILSTM_Att.

This model incorporates the advantages of Convolutional Neural Networks (CNN), which can quickly process data and reduce the dimensionality of features through their pooling layers, thereby effectively reducing the number of weight parameters in the prediction model. This CNN-optimized model not only retains the sensitivity of the BILSTM_Att model to time-series data but also introduces the ability of CNN to extract local features in image and sequence processing, enhancing the model's capability to capture local contextual information.

The improved CNN-BILSTM_Att prediction model can analyze the counter-relations between teams and the dynamics of matches more accurately, thereby improving the accuracy of match outcome predictions. The final model structure and core process can be visually represented, as shown in Fig. 5-8. This will help researchers and analysts better understand the workings of the model and apply it to subsequent match prediction work.

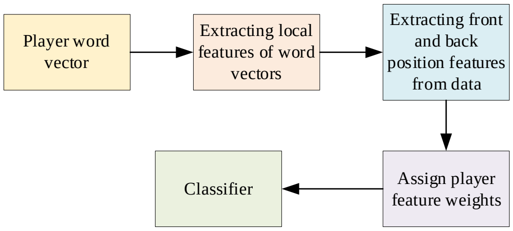
Player word vector
Extracting local features of word vectors
Extracting front and back position features from data
Assign player feature weights
Classifier

Fig. 6. Core process of the prediction model optimized with CNN.

The architecture of the improved model achieves an effective combination of feature extraction and time series analysis, with the following detailed structure:

1) Convolutional Neural Network (CNN) layer: The main task of this layer is to process the input lineup data and extract local features from it. In the context of football data analysis, this is akin to the ability of CNN to extract local pixel features in image processing tasks. In the scenario of football match prediction, the CNN can extract key local patterns and features from player data, positional information, and other relevant statistics, such as the relationships between players and the team's formation structure.

2) Bi-directional Long Short-Term Memory (Bi-directional LSTM) with Attention Mechanism (Att) layer: Data processed by the CNN layer is passed on to the BILSTM_Att layer. The BILSTM can capture the forward and backward dependencies in time-series data, while the attention mechanism helps the model focus on those pieces of information that are most critical to the final prediction. In football match prediction, this means the model can analyze not only the individual characteristics of players but also understand the dynamics of interactions between players and their changing impacts over time.

3) Softmax classifier: Finally, the feature vector output from the BILSTM_Att layer is passed to the softmax classifier. This classifier predicts the possible outcomes of the match, such as home win, draw, or away win, based on the extracted features.

This improved model—referred to as the CNN-BILSTM_Att model—leverages CNN's ability to extract highly relevant local features, and BILSTM's strength in processing time-series data. In this way, the model not only understands the current state of each team's lineup but also considers the potential impact of the evolution of team lineups over time on the match outcome.

The model's architecture can be represented by the following structure diagram:

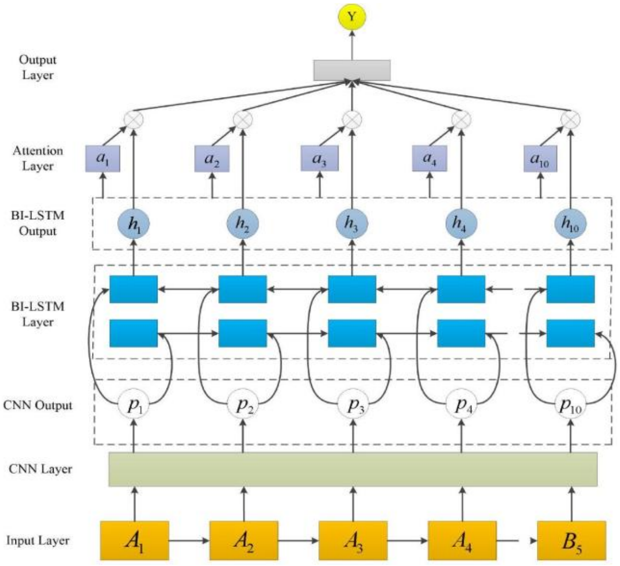

Fig. 7. Structure of the CNN-BILSTM_Att win probability prediction model.

The input and output layers of the CNN-BILSTM_Att model are the same as those of the BILSTM_Att model. In the hidden layers, the CNN-BILSTM_Att model includes an additional CNN layer compared to the BILSTM_Att model. This CNN layer extracts feature vectors from the lineup and then inputs the extracted features into the BILSTM structure. The structure of the CNN model is shown in Fig. 8.

Assuming the team lineup matrix is G, the rows of G represent the vectors of the players in the lineup. Each input lineup consists of 11 players (according to a real football game, each team has 11 field players). The dimensionality of the player vectors is composed of various features such as the technical characteristics, physical data, and tactical role of the players, which is currently set to d. Therefore, G is an 11×d dimensional matrix. $V(W(t))$ represents the d -dimensional feature vector of the t-th player in the lineup matrix G.

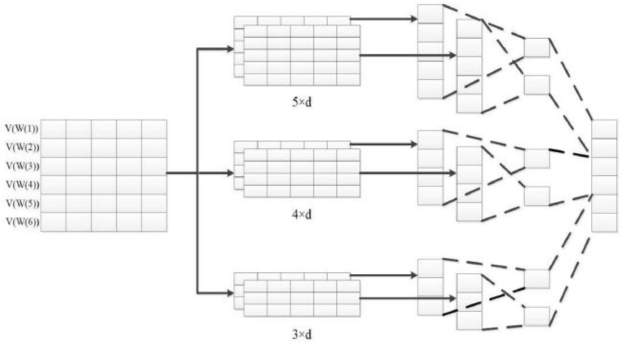

Fig. 8. Convolutional neural network structure.

When the team lineup information is input into the convolutional neural network, the network's convolutional layer uses multiple convolutional kernels of different sizes $l \times d$, to perform convolution on the team lineup matrix G, thus obtaining local feature mappings k_t . The stride of the convolutional kernels is 1. The convolution process is shown as equation:

$$k_t = f(C \cdot G_{t:t+l-1} + b)$$

In equation 5-17, f represents the activation function, and in this model, the Relu function is used as the activation function. C represents the convolutional kernel function, l represents the size of the convolutional window, $G_{t:t+l-1}$ represents the convolutional range of the convolutional kernel, which is from the t-th to the t+l-1-th position of the team lineup matrix G, and b represents the bias term. After convolution, the model can obtain a collection of team lineup feature vectors composed of multiple local features K, with the expression for K shown as equation:

$$K = (k_1, k_2, \dots, k_{m-l+1})$$

In equation, m represents the length of the lineup. In the problem of win probability prediction, each team lineup consists of eleven players, so m=11. After the convolutional layer, a pooling layer can be used to obtain strong features composed of local features. In the CNN-BILSTM_Att model, multiple filters are used to generate the features for each player. The player features after pooling are shown as equation:

$$p = \max(k_1, k_2, \dots, k_{10-l+1})$$

The final representation of the team lineup X^C is derived as equation.

$$X^C = [p_1, p_2, \dots, p_{10}]$$

The model transmits the X^C obtained from the feature extraction of the team lineup data by the convolutional neural network to the bidirectional LSTM. The subsequent data integration and classification steps are the same as those of the BILSTM_Att model.

The improved CNN-BILSTM_Att model first extracts local features, then inputs these features into the bidirectional LSTM network layer and the attention mechanism layer. After integrating features in the attention mechanism layer, it finally uses a softmax classifier for classification. By extracting features from the lineup information through the CNN model, the win probability prediction model can more accurately capture the local information of the lineup, thereby improving the accuracy of predictions.

IV. Case Study

This section will validate the effectiveness of the proposed method using a custom experimental dataset based on a sports stadium scenario.

A. Experimental Environment

The hardware environment for the experiments in this chapter is shown in Table I:

TABLE I. EXPERIMENTAL SOFTWARE AND HARDWARE ENVIRONMENT CHART

Component	Specification
CPU	Intel i7 8700k
GPU	GTX3080
Memory	32G
OS	Ubuntu18.04
CUDA	11.1
Main Frameworks	Pytorch, keras
Main Programming Language	Python3.6

Due to the complexity of the CNN-BILSTM_Att model, Fig. 9 illustrates the training process of this model.

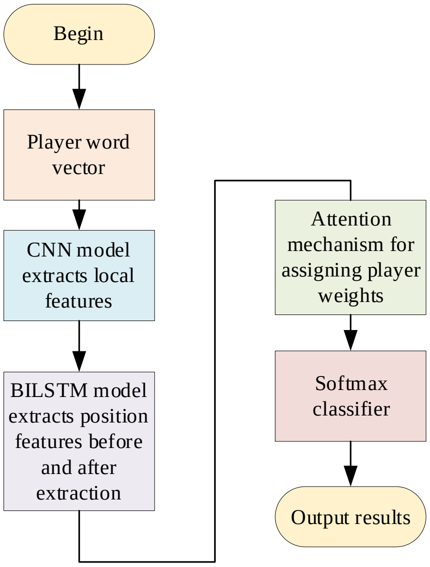
Begin
Player word vector
CNN model extracts local features
BILSTM model extracts position features before and after extraction
Attention mechanism for assigning player weights
Softmax classifier
Output results

Fig. 9. Training process of CNN-BILSTM_Att.

In the training process of the CNN-BILSTM_Att model, the focus is primarily on two aspects: feature extraction and classification. Initially, the CNN model extracts local features of the word vectors representing players. Subsequently, the BiLSTM model is utilized to extract positional features of the players' lineup. Following that, through the attention mechanism, different weights are assigned to players, reflecting their importance in the match. Finally, a softmax classifier is deployed for classification. The parameters of the CNN-BILSTM_Att win probability prediction model are displayed in Table II.

TABLE II. HYPERPARAMETERS OF THE CNN-BILSTM_ATT PREDICTION MODEL

Parameter	Parameter Value
Player Word Vector Dimension	110
BILSTM Layer Dimension	16
Learning Rate	0.01
Output Dimension	1
batch_size	200
Convolutional Kernel Sizes	[2,3,4,5]
Parameter	Parameter Value

The model's output employs the player feature vectors generated in Chapter 3, comprising a 100-dimensional feature vector from players' technical statistical data and game performance, along with a 10-dimensional extension vector (potentially including the player's position, health status, and psychological factors). The BiLSTM layer was optimized to select 16 hidden units. The learning rate and batch_size parameters were also finely tuned to achieve the optimal parameters of 0.01 and 200.

Regarding the CNN layers, this study contrasts convolutional kernels of different sizes to assess their impact on the model's performance. Convolutional kernels with sizes 2, 3, 4, and 5 were selected, and the accuracy of the prediction models with these different sizes on the validation set is shown in Fig. 10. Through these experiments, we aim to determine which convolutional kernel size can most effectively enhance the accuracy of football match outcome predictions.

The deep learning techniques applied to process football match data allow the prediction model to extract valuable features from individual technical statistics of players and the overall tactical execution of the team. This aids in more accurately predicting match outcomes. This approach translates the comprehensive performance of players and teams into a format understandable by the prediction model, thus providing technical support for various types of football analysis.

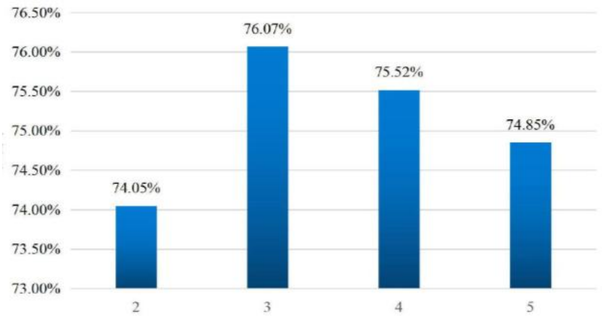

Fig. 10. Accuracy comparison of convolutional kernels of different sizes.

As can be seen from Fig. 10, the accuracy of the four different sizes of convolutional kernels was higher than that of the pure BILSTM_Att model. When the kernel size was 3, the CNN-BILSTM_Att model achieved the highest accuracy in predicting football match outcomes, reaching 76.07%. However, when the kernel size was 2, the model's performance may have been insufficient due to the lack of feature extraction capability. When the kernel size was greater than 3, using larger convolutional kernels meant an increase in the model's weight parameters, which could lead to a decrease in model performance due to issues such as insufficient computational resources.

B. Experimental Results

Fig. 11 shows the accuracy curve of the CNN-BILSTM_Att model for football match win probability prediction over 200 iterations. Overfitting phenomena also occurred during the fitting process of this model. Due to the CNN model's strong feature extraction capability, the CNN-BILSTM_Att model showed a significant improvement in prediction accuracy over the BILSTM_Att model, demonstrating greater expressive power. It can also be seen that the model fitted more quickly, with the CNN-BILSTM_Att model reaching its best performance around the 80th iteration.

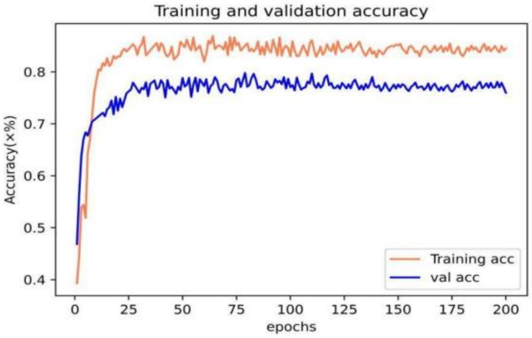

Fig. 11. Fitting curve of the CNN-BILSTM_Att model.

For a more comprehensive comparison, this paper further constructed a CNN-BILSTM_Att model that does not use the extended feature vector as input, namely a CNN-BILSTM_Att model without historical encounter information (No Historical Encounter Relationship CNN-BILSTM_Att, NHR-CNN-BILSTM_Att). This variant differs from the original CNN-BILSTM_Att win probability prediction model only in terms of the input data; it uses only the feature vectors generated from regular team and player statistical data as the model input, in order to assess the impact of historical encounter information on win probability prediction. The model parameters were kept consistent with the complete prediction model. The NHR-CNN-BILSTM_Att model ultimately achieved an accuracy of 74.95%, and the fitting process is shown in Fig. 12.

The results of the NHR-CNN-BILSTM_Att prediction model indicate that the historical encounter information, or the extended vector, has some impact on the final prediction results, suggesting that considering the history of encounters between teams is important when constructing a football match win probability prediction model.

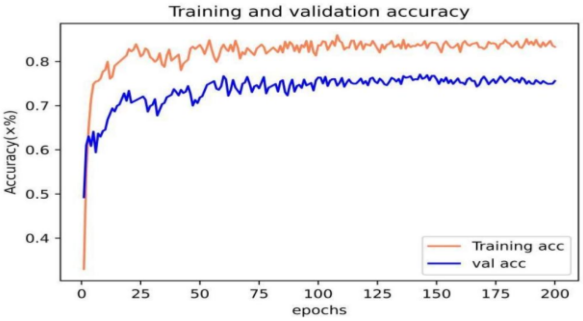

Fig. 12. Fitting curve of the NHR-CNN-BILSTM_Att model.

V. Conclusion

This paper investigates win probability prediction for football matches using deep learning techniques. By analyzing team line-ups and player abilities, the study develops the CNN-BILSTM_Att model, achieving high accuracy in predicting match outcomes. Player statistical data is transformed into numerical vector representations derived from historical performance and ability scores. Initial attempts with three machine learning methods yielded low accuracy, prompting the introduction of an LSTM model to explore key positional information in line-ups. The final models, BILSTM_Att and CNN-BILSTM_Att, incorporate an attention mechanism to weigh player contributions and extract local features effectively. The proposed models enhance win probability predictions by providing valuable insights for coaches, analysts, and fans, thereby enriching the overall viewing experience. Additionally, these methodologies may inspire similar applications in other sports, advancing the field of sports data science.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks.”

References

[1] Zheng, H., Liu, S., Zhang, H., et al. (2024). Visual-triggered contextual guidance for lithium battery disassembly: a multi-modal event knowledge graph approach. Journal of Engineering Design, 1-26.

[2] Fu, T., Li, P., & Liu, S. (2024). An imbalanced small sample slab defect recognition method based on image generation. Journal of Manufacturing Processes, 118, 376-388.

[3] Aoki, R. Y. S., Assuncao, R. M., & Vaz de Melo, P. O. S. (2017). Luck is hard to beat: The difficulty of sports prediction. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1367-1376).

[4] Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. Journal of Artificial Intelligence Research, 73, 1285-1322.

[5] Li, Y., Wang, L., & Li, F. (2021). A data-driven prediction approach for sports team performance and its application to National Basketball Association. Omega, 98, 102123.

[6] Valero, C. S. (2016). Predicting win-loss outcomes in MLB regular season games: A comparative study using data mining methods. International Journal of Computer Science in Sport, 15(2), 91-112.

[7] Mukherjee, S., Huang, Y., Neidhardt, J., et al. (2019). Prior shared success predicts victory in team competitions. Nature Human Behaviour, 3(1), 74-81.

[8] McGarry, T., & Franks, I. M. (1994). A stochastic approach to predicting competition squash match-play. Journal of Sports Sciences, 12(6), 573-584.

[9] Hodge, V. J., Devlin, S., Sephton, N., et al. (2019). Win prediction in multiplayer esports: Live professional match prediction. IEEE Transactions on Games, 13(4), 368-379.

[10] Chakraborty, S., Dey, L., Maity, S., et al. (2024). Prediction of winning team in soccer game: A supervised machine learning-based approach. In Advances on Mathematical Modeling and Optimization with Its Applications (pp. 170-186). CRC Press.

[11] Buhamra, N., Groll, A., & Brunner, S. (2024). Modeling and prediction of tennis matches at Grand Slam tournaments. Journal of Sports Analytics, 10(1), 17-33.

[12] Wen, Z., Liu, J., & Liu, C. (2024). Football momentum analysis based on logistic regression. Frontiers in Computing and Intelligent Systems, 7(2), 60-64.

[13] Shamshiri, A., Ryu, K. R., & Park, J. Y. (2024). Text mining and natural language processing in construction. Automation in Construction, 158, 105200.

[14] Just, J. (2024). Natural language processing for innovation search: Reviewing an emerging non-human innovation intermediary. Technovation, 129, 102883.

[15] Mekkes, N. J., Groot, M., Hoekstra, E., et al. (2024). Identification of clinical disease trajectories in neurodegenerative disorders with natural language processing. Nature Medicine, 1-11.

[16] Gagliardi, G. (2024). Natural language processing techniques for studying language in pathological ageing: A scoping review. International Journal of Language & Communication Disorders, 59(1), 110-122.

[17] Raza, S., Garg, M., Reji, D. J., et al. (2024). Nbias: A natural language processing framework for BIAS identification in text. Expert Systems with Applications, 237, 121542.

[18] Joshi, A., Dabre, R., Kanojia, D., et al. (2024). Natural language processing for dialects of a language: A survey. arXiv preprint arXiv:2401.05632.

[19] Bunker, R., Yeung, C., & Fujii, K. (2024). Machine learning for soccer match result prediction. arXiv preprint arXiv:2403.07669.

[20] Chakraborty, S., Dey, L., Maity, S., et al. (2024). Prediction of winning team in soccer game: A supervised machine learning-based approach. In Advances on Mathematical Modeling and Optimization with Its Applications (pp. 170-186). CRC Press.

[21] Holmes, B., & McHale, I. G. (2024). Forecasting football match results using a player rating based model. International Journal of Forecasting, 40(1), 302-312.

[22] Xiaoyu, F., & Shasha, W. Evaluating the pinnacle of football match key statistics as in-play information for determining match outcomes in Europe's foremost leagues. Social Science Quarterly.

[23] Saribekyan, G., & Yarovoy, N. (2024). Football prediction model based on teams' Elo ratings and scoring indicators.

[24] Gu, C., De Silva, V., & Caine, M. (2024). A machine learning framework for quantifying in-game space-control efficiency in football. Knowledge-Based Systems, 283, 111123.