Prediction of the Playoffs of Baseball Organization League Using the Deep Neural Network

Hee-Chang Han[1], Ju-Hyeok Park[2], Yang-Jae Lee[3], Yoo-Lim Jun[4], Yoo-Jin Moon[5]
[1,2,3,5]Divison of Global Business&Tech., Hankuk University of Foreign Studies, Korea
[4]Dept. of German Lang. and Cultures, Hankuk University of Foreign Studies, Korea
[5] yjmoon@hufs.ac.kr

Abstract : *The paper suggested the Deep Neural Network Model System that utilized the deep neural network to predict Baseball Organization League (BOL) chances of advancing to the playoffs next season. As a research method, the data for each season of BOL were summarized from 1999 data, which were confirmed that among the data for each season, they had a significant effect on the season results such as average score per game, batter On-base Plus Slugging (OPS), and pitcher Walks plus Hits divided by Innings Pitched (WHIP). The model system performed more accurate in the sigmoid, softmax, and binary crossentropy functions than in the linear and tanh functions. As a result of predicting the actual 2022 season, 88.33% accuracy was derived in the neural network model system. It was proved that the season results were good if the value of high-weight variables such as the number of wild pitches and home runs were excellent. This system suggested in this paper is considered to be effective in applying to organizing other baseball teams.*

Keywords - *Deep Neural Network, Artificial Intelligence, Playoff, Prediction, Korea Baseball Organization*

I. INTRODUCTION

Advances in technology are increasing the amount of data consumed in everyday situations. It is said that people generated about 2.5 trillion bytes of data every day, and it was known that 90% of the total data was made in just two years, which occurred a great response in terms of technology development and progress. Despite the vast amount of data generation, data management and analysis in the business have in the traditional way been organized and managed in other parts of the organization. As a result, there is a growing need to break current organizational boundaries and improve integration in the IT and business sectors [1, 2, 3]. These changes can play an important role in driving organizations and businesses to successful transformation. Furthermore, in recent years, global business processes have been rapidly digitalizing, increasing the amount and speed of data [4]. In line with these trends data has emerged as a new profit structure for businesses, and growth and expansion are also key capabilities pursued by companies [5, 6].

Data utilization can also play an important role in the sports field. Just as big data can help companies and customers understanding, improve business efficiency & processes, and empower decision-making, they can help efficiency and decision-making in the sports field. And baseball would be the best sport to use the big data analysis, since baseball and data are good combination for the spectators and box office performance and baseball is a sport that starts with a record and ends with a record. Usually sports records only inform results such as ranking, winning or losing, but baseball records show a "process." The game can be reconstructed from the first inning to the ninth inning even if you don't see the game in person through all the balls thrown by the pitcher and the baseball record written by the batter. However, the recent attention of baseball data is connected to the history of baseball, which has begun to process records that have been accumulated into statistics and to use them as "big data" with the development of advanced technology [7].

Baseball is the sport that provides a lot of data throughout all sports. All actions such as pitchers' pitches, batters' hits, and base runs are recorded. As it is the sport where all actions are data, the outcome may be determined depending on how data are used. Many baseball fans are interested in the data from the previous games and enthusiastic about various Internet communities based on the data. Many studies have been conducted based on data generated in the game since it is a sport that receives a lot of attention from the public and data affects the game. These changes are very different from the past relying on scouts' experiences and intuitive senses. As shown in the movie "Moneyball," if you quickly find items that are most related to victory and utilize data, you can help the team develop.

In this paper, research data were extracted from Statiz, a baseball data recording site, and Kaggle, an artificial intelligence community, to predict the exact probability of each team advancing to the playoffs. In addition, seasonal player data from 1999 to 2021 were extracted from each club's website. And the team's average Runs Batted In (RBI), batter's record, and pitcher's record were digitized through a normalization process, and the artificial neural network was constructed and performed learning by adding the ranking of each season to it. To find the optimal performance to predict the playoff probability, the research proceeded adjusting each variable, the number of hidden layers, and the number of nodes. To measure the performance of the model, it predicted the results of the actual season of 2022. [8, 9, 10]

This paper is organized as follows: Section II describes previous research and introduction of baseball and KBO. Section III describes system architecture of this research including data selection, feature selection, data normalization and the suggested deep neural network model. Section IV provides result analysis of the research model and predictions of information utilization applications. Finally, discussions and contributions are described in Section V.

II. PREVIOUS STUDIES AND PRELIMINARIES

1. Related works

Recently research on winning and losing prediction has not been limited to specific sports, but has been conducted in various fields [2, 7, 11, 12]. But most of the works consist mainly of programs that predict real-time games.

Major League Baseball (MLB) and Korea Baseball Organization (KBO) provide odds of victory for each situation. One of the researches, Suncheon National University research [7], analyzed the effect of pace at the beginning of the game on the victory or defeat of the game by predicting the outcome of each inning in an artificial intelligence environment. In addition to baseball, there were studies on predicting matches based on data after 15 minutes for soccer and League of Legend games [7].

In most cases, the research focused on a specific point in time, and research on the final ranking of the season was relatively insufficient. The most widely known Pythagorean winning rate in baseball [12] was calculated by the ratio of points and total points, starting with the idea of winning if you scored a lot and lost less points, which is considered to be inappropriate as data to prove the overall team ranking. In the other previous studies, On-base Plus Slugging (OPS) and Wins Above Replacement(WAR) were mainly used as data that affected victory, which was also insufficient for accurate analysis. Accordingly, our work focused on making more accurate predictions through the process of securing factors that affected prediction of the game through the deep neural networks. Therefore, based on team stats [13], it was possible to predict the playoff teams during the season and determine whether there was a lot of luck or whether there was room for a rebound.

2. Introduction to Rules of MLB and KBO

In baseball, both teams consisting of nine players (ten in case of designated hitters) play each nine innings of offense and defense. The visiting team goes first and the home team goes second. If three batters or runners are out of the attacking team, the offense will be transferred to the defending team by alternating offense and defense. However, if the home team is leading at the end of the ninth inning after the top of the ninth inning and there is no possibility of a change in the game, the game will end at the top of the ninth inning. When the tie is maintained even after the end of the ninth inning, MLB will play infinite overtime until the game is won, but in the case of the KBO League, if the game continues to be tied until the 12th inning, the game will be terminated and recognized as a draw.

The KBO is a professional baseball league in South Korea. A total of 10 teams will compete, and the season rankings in the league will be calculated in the order of winning rate, and the top five teams will advance to the KBO postseason in the order of winning rate.

People's favorite elements of baseball will be numbers of home runs and strikeouts. However, it was difficult to predict the team's ranking with only two indicators. In order to win in baseball, it is important to simply get on base in offense and to make the opponent stop from getting on base in defense. Therefore, this paper aims to predict the outcome of the game with data such as 'OPS', ' Walks plus Hits divided by Innings Pitched (WHIP)', 'runs', and 'Earned Run Average (ERA)', where it was hypothesized that "OPS" and "WHIP" would have the greatest impact on the victory or defeat of the game.

3. Architecture of the Deep Neural Networks

Deep neural network (DNN) models, suitable approaches for mathematics and human simulation, can be implemented using Keras, Python and etc. [3, 11, 13] The model focuses on an end-to-end approach to develop regression and classification supervised, which are learning algorithms using real business-driven use cases often implemented in Keras, provided by TensorFlow [14, 15].

DNN is an artificial neural network that contains hidden multi-layers; it inherently fuses the process of feature extraction with classification into learning using the fuzzy support vector machine (FSVM) and enables the decision making.

III. THE PROPOSED SCHEME

In this paper, feature selection of neural network structures was performed using ranking data from all teams from 1999 to 2022 to select input variables [16]. And then, the Deep Neural Network model was constructed to predict KBO's playoff teams by entering selected feature values processed by normalization of input data. Finally, an evaluation of the model was conducted utilizing train data of 2022.

1. Big Data Sources

For the dataset, KBO Battling Data and KBO Pitching Data provided by Kaggle were used in the paper [15, 16]. They provided data from 1982 to 2021, which were ‘runs_per_game’, ‘RBI’, ‘batting_average’, ‘OPS’, ‘runs’, ‘hits’, ‘doubles’, ‘triples’, ‘homeruns’, ‘stolen_bases’, ‘caught_stealing’, ‘bases_on_balls’, ‘Grounded into Double Play (GDP)’, ‘Hit By Pitch (HBP)’, ‘sacrifice_hits’, ‘sacrifice_flies’, ‘Intentional Based on Ball (IBB)’, ‘On-Base Percentage (OBP)’, ‘Slugging Average (SLG)’, ‘ERA’, ‘WHIP’, ‘strikeouts’, ‘complete_game’, ‘shutouts’, ‘saves’, ‘P_hits’, ‘P_runs’, ‘earned_runs’, ‘home_runs’, ‘walks’, ‘intentional_walks’, ‘wild_pitches’, and ‘strikeout_walk’.

Among them only data from 1999 to 2021 were used since these data are the final data of the season [8, 9]. The accumulated data such as 'hits' and 'strikeouts' were judged to be inappropriate to be used as they were, and so normalized to have relative values. Since the dataset only provided team stats and did not provide the final ranking, the final ranking was referenced in Statiz [10].

2. Feature Selection

This research was performed in a Python 3.8.5 version of the Google Colaboratory environment and used the Keras and Pandas modules in Tensorflow 2.3.0 version. And, the research decided which variables largely affected the accuracy of predicting the match. To measure the weight of each variable for the suggested model, it built a Single-Layered Neural Network with data such as 'runs_per_game’, ‘RBI’, ‘batting_average’, ‘OPS’, ‘runs’, ‘hits’, ‘doubles’, ‘triples’, ‘homeruns’, ‘stolen_bases’, ‘caught_stealing’, ‘bases_on_balls’, ‘GDP’, and ‘HBP’.

For the feature selection processing, the learning rate used was 0.01, and the optimizer ‘Adam’, and the loss function ‘Binary_crossentropy’, and the Evaluation Function ‘Accuracy.’ The result of this feature selection was arranged in the descending order as illustrated in Table 1.

Table 1. Weights of Variable Data Set

Rank	Variable	Weight
1	OPS	0.7028324
2	runs	0.69473726
3	wild_pitches	0.6411595
4	shutouts	0.5427704
5	GDP	0.52307427
6	home_runs	0.48795867
7	caught_stealing	0.48607063
8	saves	0.37039217
9	P_hits	0.31147838
10	WHIP	0.01213737
11	intentional_walks	-0.22103837
12	SLG	-0.257014
13	ERA	-0.36572212

14	RBI	-0.4309935
15	bases_on_balls	-0.48494875
16	triples	-0.5118612
17	stolen_bases	-0.51379627

As a result of the measurement, elements such as 'OPS', 'runs', 'wild_pitches', 'shutout', and 'GDP' showed relatively large weights. In addition, 'WHIP' which was initially considered an important factor, showed a relatively low weight. 'OPS' was one of the stats that evaluated batters in baseball and was calculated by 'on-base percentage + slugging percentage'. Both factors played an important role in prediction because they were important data in hitting. The "wild_pitches" was recorded when the ball pitched by the pitcher went to a course that the catcher could not catch, and the batter or runner succeeded in advancing further. Since they were data showing the capabilities of pitchers and catchers, it greatly affected the prediction. "Shutout" occurred when one pitcher finished the game without conceding a single point, including an involuntary loss due to a fielder's defensive error. It could be seen as data directly related to victory. "GDP" referred to a defender catching the ball after a batter's hit in a situation where it was no-out or more than one out, and out two runners. It could be seen as data representing good defense. Therefore, predicting the ranking using these five inputs would enable more accurate predictions. Therefore, predicting the ranking using these five inputs would enable more accurate predictions.

Consequently, these five variables were set as independent variables, and whether or not to advance to the playoffs was set as dependent variables.

3. Data Normalization

The variables in baseball data had a high proportion of natural numbers. Since the range of variables is recommended between -1 and 1 for operating the DNN program, the largest number of all variables was set to 1 and the smallest number to 0 so that each value could fit the deviation (X value). The ranking for the year was set to Y value. In order to determine whether to advance to the playoffs, 0 was set from 1st to 5th place of the baseball teams, and the rest to 1.

The suggested model used ‘Gaussian Normalization’ as the normalization method [17, 18]. ‘Gaussian Normalization’ is a method of normalizing x’ = (x – means) / standard deviation instead of the input x. Table 2 showed examples of input data with normalization.

Table 2. Result of Data Normalization

OPS	runs	Wild_P	shutout	GDP
0.775	0.79661	0.04012	0.26315	0.70270
0.741	0.74258	0.05616	0.73684	0.66216
0.759	0.73834	0.07422	0.52631	0.62162
0.755	0.76589	0.10230	0.63157	0.77027
0.711	0.690678	0.04312	0.94736	0.68918
0.736	0.75317	0.05616	0.31578	0.70945
0.722	0.75847	0.05817	0.36842	0.62162
0.674	0.60063	0.05817	0.31578	0.79729
		. . .		

4. Proposed Deep Neural Network Model

The research constructed the deep neural network model for predicting teams advancing to the KBO playoffs using Keras of TensorFlow.

4.1 Softmax vs. Sigmoid for Accuracy Function

The research experimentally compared Softmax with Sigmoid for the Accuracy function setting. Table 3 showed the result of this comparison. Experiments showed that the accuracy difference between Softmax

(0.3666) and Sigmoid (0.8333) was clear. Thus, the research model selected the Sigmoid for the Accuracy function.

Table 3. Accuracy Function Comparison

Accuracy Function	Accuracy
Sigmoid	0.8333
Softmax	0.3666

4.2 Activation Function Comparison

To decide the Hidden-Layer's Activation function, the research experimentally compared ReLU, Softmax and Tanh Function. Table 4 showed the result of this comparison. Experiments showed that among the three activation functions, Softmax had the highest accuracy. Thus, the research used Softmax as the Activation function of the model.

Table 4. Activation Function Comparison

Activation Fuction	Accuracy
ReLU	0.8166
Softmax	0.8500
Tanh	0.3666

4.3 Setting the Loss Function

To establish the Loss Function of the model, the model compared 'MSE (Mean Square Error)', to 'Binary Crossentropy'. The activation functions were all equally set to Adam, and the learning rate was set to 0.1, with the epoch set to 100. Table 5 showed the result of this comparison. We used 'Binary Crossentropy', which had the highest accuracy as Loss function in our model.

Table 5. Loss Function Comparison

Loss Function	Accuracy
MSE	0.8
Binary Crossentropy	0.8333

4.4 Setting the optimizer

To set the Optimizer for the model, the model compared 'Adam', 'Sgd', 'AdaGrad', 'RMSProp', 'AdaMax', and 'Nadam'. The learning rate was set to 0.1, and the epoch was set to 100. Table 6 showed these comparison results. We used 'Adam', which had the highest accuracy and the lowest loss value, as our model's Optimizer.

Table 6. Optimizer Comparison

Optimizer	Accuracy
Adam	0.8666
Sgd	0.6499
AdaGrad	0.6333
RMSProp	0.8166
AdaMax	0.8166
Nadam	0.8166

4.5 Setting the Metrics

To set the Metrics for the model, the model compared 'Accuracy', 'MSE', and 'binary_accuracy'. The learning rate was set to 0.1, and the epoch was set to 100. Table 7 showed these comparison results. We used 'Accuracy', which had the highest accuracy and the lowest loss value, as our model's Metrics.

Table 7. Metrics Comparison

optimizer	Accuracy
Accuracy	0.8333
MSE	0.1244
binary_accuracy	0.8166

IV. EXPERIMENT

As described in the Section III, Sigmoid and Softmax were experimented and compared to set the Accuracy Function. And ReLU, Softmax and Tanh were experimented and compared to set the Activation Function. As the result of the experiments of combined functions the suggested model was confirmed that the accuracy was 0.8833 when Sigmoid, Softmax and Binary Crossentropy were all applied. The model was constructed using the deep neural networks with hidden layers. Based on each team's season-specific records, the batter data(‘RBI’, ‘OPS’, ‘runs’, ‘triples’, ‘stolen_bases’, ‘caught_stealing’, ‘bases_on_balls’, ‘GDP’, ‘SLG’) and the pitcher data(‘ERA’, ‘WHIP’, ‘shutouts’, ‘saves’, ‘P_hits’, ‘home_runs’, ‘intentional_walks’, ‘wild_pitches’)were given to the input floor. The weights were modified through ReLU in the hidden layers.

Through the Sigmoid function, the value of whether a professional baseball game could or couldn’t advance to the playoffs was derived as output. To this end, Binary Crossover was used as the Loss function, and Adam was used as the Optimizer. The architecture set in this study is shown in Fig. 1.

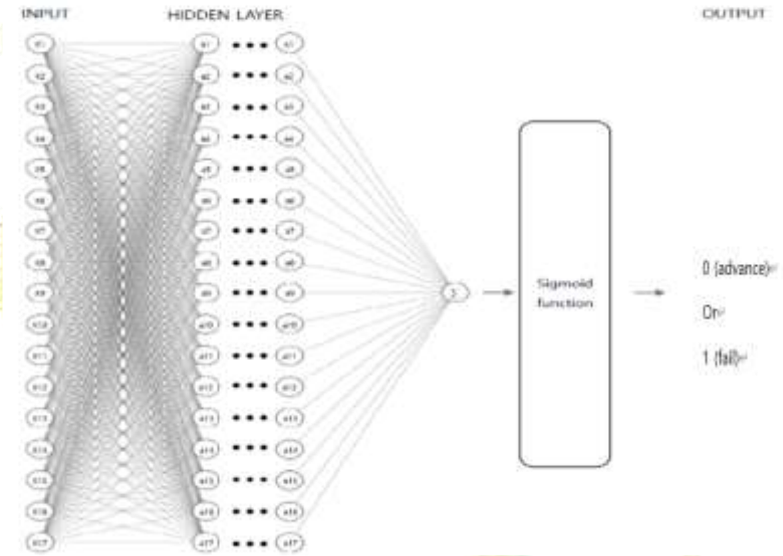

Fig. 1. Architecture of the game prediction

As the result of the execution, the test accuracy using the 2022 data was 0.8833, which succeeded in predicting most ranks. Through the considerably high accuracy, it was possible to conclude that the neural network system suggested was useful in predicting whether a professional baseball game would advance to the playoffs.

V. CONCLUSION

This paper proposed a deep neural network model predicting the playoff advance team of 'KBO'. It utilized various real big data of KBO games and Keras of TensorFlow. The model showed 88.33% accuracy in predicting the '2022 KBO playoffs' without overfitting disadvantages.

Experiments have shown that even ordinary users can use this model to predict the direction of the season by focusing on several factors --- "OPS", "runs", "GDP", "wild_pitches" and "shutouts". It is expected that this model would also be used to predict other baseball league playoff teams abroad.

Theoretically, the model suggested in the research implies that it would be able to obtain better results in identification problems (i.e., precision of Win/Loss prediction at a specific timepoint, degree of open data utilization, coach's strategy, squad's depth and possibility of training index utilization of professional teams) of present various game markets if these DNN analyses are applied through big data including time series data. DNN analyses in this research perform better in prediction than models from traditional machine learning.

In fact, proper DNNs can solve higher-level functions of learning with more complexity and abstraction than shallow neural networks, which provides the probability that these higher-level functions constitute specific objects and scenes. Through the deep functional hierarchy, a suitable DNN can achieve excellent performance in many tasks. Therefore, this research may predict the more exact results of the KBO playoffs through DNN analyses and outperform experts, which suggests that the method presented in this research could be used not only as an effective training indicator but also as complementary means for professional baseball team managers. Also, if the prediction method of baseball utilizing artificial intelligence (AI) and DNN can be socialized, it might induce public excitement about baseball by predicting the possibility of the season results, which contributes to baseball revitalization, prosperity and expansion of the baseball industry.

In addition, professional teams could determine the most important factors that affect their advance to the playoffs by measuring weights. The team could win the game effectively and quickly by practicing what factors affected the win. Accordingly, this method will increase the winning rate of professional teams. This prediction system could also help you for what areas of focus should be placed on recruiting players.

The research limitation exists in collecting real-world data of game applications for the professional environmental data. In this research, data for simple season results were used as learning data. Therefore, the model did not take into account environmental factors such as injuries and scandals that might occur during the season.

The future work to do is as follow. First, research needs to utilize more data to predict the winning team, not just the playoff team. If the research added professional data, it could expect to get more accurate results than previous experiments. Second, the methods proposed in this study should be experimented for the game prediction of other professional baseball leagues.

Acknowledgements

This work was supported by Hankuk University of Foreign Studies Research Fund of 2022.

This paper has been extended the paper titled “Prediction of KBO playoff Using the Deep Neural Network” of the Proceedings of the 2023 Winter Conference in the Korea Society of Computer Information.

REFERENCES

[1] Dong-Hui Sin and Jae-Gil Lee, Trend and Implication of Big Data, *Review of Korean Society for Internet Information, 14(2)*, 2013, 5-17.

[2] Si-Jae No, et al., Win/Loss Prediction of ‘League of Legends’ Utilizing the Deep Neural Network, *Proc. 2021 Winter Conference in the Korea Society of Computer Information*, 2021, 1-4.

[3] Wesley Chai, Mark Labbe, and Craig Stedman, Big Data Analytics, 2021. *https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics*

[4] Mathias Kirchmer, Peter Franz, and Rakesh Gusain, Digitalization of the Process of Process Management, *Proc. Seventh International Symposium on Business Modeling and Software Design*, 2017.

[5] Sangho Kim, A Study on Relationship of BDBA (Big Data Business Analytics) System and Supply Chain Management, *Journal of Korea Research Association of International Commerce, 19(2)*, 2019, 89-107.

[6] Yong Chen, Hong Chen, Anjee Gorkhali, Yang Lu, Liqian Ma, and Ling Li, Big Data Analytics and Big Data Science: A Survey, *Journal of Management Analytics, 3(1)*, 2016, 1-42.

[7] Taehoon Kim, et al., Win/Loss Prediction of KBO Using Inning Data in the Artificial Intelligence Environment, *Proc. 2020 Fall Conferences Online*, 27(2), 2020, 1-3.

[8] Kaggle, Batting Data, 2022. *https://www.kaggle.com/datasets/mattop/baseball-kbo-batting-data-1982-2021*

[9] Kaggle, Pitching Data, 2022. *https://www.kaggle.com/datasets/mattop/korean-baseball-pitching-data-1982-2021*

[10] Statiz, Team Data, 2022. *http://www.statiz.co.kr/league.php?opt=2022*

[11] Fadi Thabtah, Li Zhang, and Neda Abdelhamid, NBA Game Result Prediction Using Feature Analysis and Machine Learning, *Annals of Data Science, 6(1)*, 2019, 103-116.

[12] Jason Rosenfeld, et al., Predicting Overtime with the Pythagorean Formula, *Journal of Quantitative Analysis in Sports, 6(2)*, 2010.

[13] Katy Warr, *Strengthening deep neural networks: Making AI less susceptible to adversarial trickery*, (O'Reilly Media, 2019).

[14] Hyejeong Park, Kyoungha Seok, Juyong Shim and Changha Hwang, *Deep learning from TensorFlow*, (Hanbit Academy Press, 2019).

[15] Jojo Moolayil, *Learn Keras for deep neural networks: A fast-track approach to modern deep learning with Python*, (Apress, 2019).

[16] Heechang Han, et al., Prediction of KBO playoff Using the Deep Neural Network, *Proc. 2023 Winter Conference in the Korea Society of Computer Information*, 2023.

[17] Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep learning*, (MIT Express, 2016).

[18] Jen-Tzung Chien, *Source separation and machine learning*, (Elsevier: Academic Press, 2020).