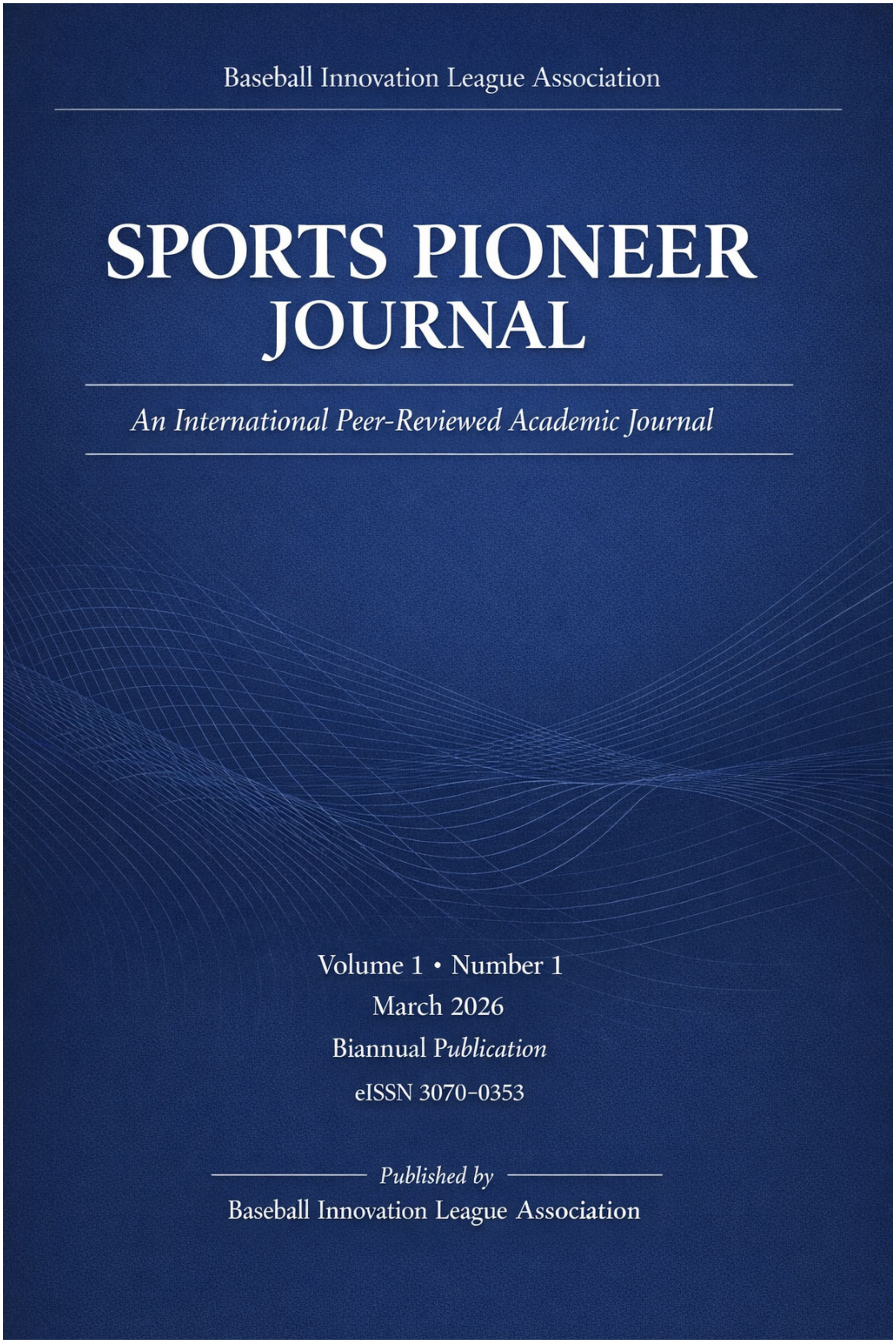


Baseball Innovation League Association

SPORTS PIONEER JOURNAL

An International Peer-Reviewed Academic Journal

Volume 1 • Number 1
March 2026
Biannual P*ublication*
eISSN 3070-0353

Published by
Baseball Innovation League Association

Sports Pioneer Journal

Volume 1, Number 1 | March 2026
Biannual | eISSN: 3070-0353

Sports Pioneer Journal

Volume 1, Number 1
March 2026
Biannual | eISSN: 3070-0353

Editorial Board

Journal Information

Frequency: Biannual (March and September)
Language: English
Format: Online
Publisher: Baseball Innovation League Association (USA)

Scope: Sports science, sports education, sports medicine, sports management, coaching science, sports analytics, and related interdisciplinary research.

Editorial Office

Sports Pioneer Journal
Baseball Innovation League Association
Address: 27822 Shiloh Pine Ln. Katy, TX, 77494
Email: admin@baseballinnovation.org
Website: https://journal.baseballinnovation.org

Sports Pioneer Journal
An International Peer-Reviewed Academic Journal
Biannual Publication

Volume 1 • Number 1
March 2026

eISSN: 3070-0353
Published by:
Baseball Innovation League Association

Copyright Notice

Disclaimer

Authors are solely responsible for the content of their respective articles. The opinions expressed in published articles are those of the authors and do not necessarily reflect the views of the Editorial Board or the Publisher.

Permissions and Reprints

Requests for permissions, reprints, or licensing inquiries should be directed to:

Baseball Innovation League Association
admin@baseballinnovation.org
https://journal.baseballinnovation.org

Sports Pioneer Journal

Volume 1, Number 1 | March 2026
eISSN: 3070-0353

CONTENTS

ORIGINAL ARTICLES

BOOK REVIEW

Machine Learning with Trustworthy AI for Cycling Race Prediction in Sports Betting

Ray-I Chang[1]*, Huang Ching Liu[1]

[1]Department of Engineering Science & Ocean Engineering, National Taiwan University,

Taipei, Taiwan

**Corresponding author: rayichang@ntu.edu.tw*

Abstract

In today's era of booming AI, machine learning (ML) has reshaped how we analyze sports events. Cycling race prediction for sports betting is one of them. This paper establishes cycling betting in an online gambling system. The user can give a stake, and the system will bet according to the prediction result of ML. Finally, an output amount is provided to measure the profit. Notably, conventional ML provides only the AI model without considering its trust in prediction. However, in betting/gambling, it is crucial to find an indicator for measuring the trust of the AI model in its prediction. In this paper, we design an algorithm that can evaluate the trust of the AI model efficiently and effectively. Our experiments use the rider data selected from 2014 to 2018 for ML, and the predicted target is the top10 in 2019 Tour *de* France. The experimental results show that the precision rate of the top10 obtained by the basic AI model can reach 62%. Then, we try to add our trust score in prediction, and apply only the trusted AI prediction in betting. Results show that the maximum profit can increase by about 40%.

Keywords: Cycling, Data Analytics, Sports Betting, Machine Learning, Trustworthy AI

1. Introduction

Due to the increasingly convenient collection of sports data, sports analysis has emerged in recent years. The direction of earlier research is more popular in ball sports, such as football, basketball, and rugby. In contrast to cycling races, competition did not appear until the 1900s (Gabriele, 2011), so there are fewer studies analyzing cycling events. Recently, there are some sports analysis studies using AI (artificial intelligence) and machine learning (ML). Unlike the manual recording and analysis of data, the use of ML can more effectively analyze data, and improve sports analysis to a new level (Soneja, 2019). Athletes can train and improve themselves through the results of sports analysis, and choose the best strategy in the competition. Broadcasters, able to broadcast game statistics in real-time through sports analysis. Spectators, who can participate in sports betting through sports analytics. This paper studies the part of spectators participating in sports betting, and tries to apply ML to predict the outcome of cycling events and improve profitability. Since sports betting itself is a game with probability components, how to judge whether the prediction result of the AI model is trustworthy is an important topic (Das, 2019). Therefore, this paper proposes an innovative trust reference, so that users can use this trust to create maximum profit.

Sports betting has a long history (Gilchrist, 2020). It can be traced back to Greece more than 2,000 years ago. It was turned to underground betting due to religious factors. Nowadays, the emergence of horse racing betting let sports betting become legal and spread rapidly around the world. Therefore, betting on various events also appeared one after another. Compared with ball sports betting, although cycling betting appeared later, it is still quite popular. In this paper, we apply ML to build a cycling betting and reward prediction system, which we can predict the top10 of Tour *de* France and finally input stake X dollars to give profit Y dollars. The simple architecture is as follows: first, ML method is used to predict whether the rider is in the top10, and then the final profit is calculated according to the predicted result. In addition, we put the prediction results into the trust calculation module, and finally adjust our betting combination according to the trust score to output the optimized final profit. In summary, the main contributions of this paper are listed as follows.

1. Build a prediction system for cycling betting.
2. Propose an algorithm for trustworthy AI.
3. Use trust to optimize profit.

2. Related works

The types of sports that use AI for sports analysis are dazzling, and the research objectives and the results of interest are also different. In (Thabtah, 2019), different AI models were used to predict the outcome of the game through the data recorded by the USA Basketball Association, and they compared the performance of the model the proposed, furthermore, they also found the most significant feature influencing the results. The study (David, 2011) used official data recorded by the American Football League and used neural networks to predict the results of American football. The model is built using statistical differentials to compare teams, and finally through principal component analysis to determine which statistics has the greatest impact on the model. In baseball, (Valero, 2016) predicted the final result based on the ranking of teams and players, and evaluated four algorithms they used in the works.

In the field of cycling, the main focus of research is to predict the relevant race data or to study the results of the race. Among the relevant race data, there are two popular ones, namely heart rate prediction and power performance prediction. Heart rate and power performance have always been the data that professional riders pay attention to. They can make training more scientific and can monitor training intensity in real time.

And influence the trend of the entire event, benefiting the riders and coaches. In (Mutijarsa, 2016), neural network (NN) was used to create a heart rate prediction model for riders. After comparing with the original sensor data, the final error value was 2.43 for the training data and 3.02 for the test data. In (Kataoka & Gray, 2018), they used GPS data to design and build a real-time predicted power output system for Tour *de* France. Their features were generated by autoencoder and hand-selected. Finally, the Mean Absolute Error (MAE) of AI model is lower than the conventional physics model. It is helpful to a certain extent in training riders, and riders can use these data to improve their ability.

Although there are few studies on the prediction of race results, there are still many studies worthy of reference. The study (De Spiegeleer, 2019) uses data from past races to predict the outcome of future races. This study focuses on predicting the outcome data of 3 races, namely the average velocity of a stage, the difference between the average stage velocity and the velocity of a rider, and the number of wins in head-to-head matches between two riders in a stage. It uses data from past races to predict the outcome of future races. This study focuses on predicting the outcome data of 3 races, namely the average velocity of a stage, the difference between the average stage velocity and the velocity of a rider, and the number of wins in head-to-head matches between two riders in a stage. (Kholkine, 2020) used historical data from the past to predict the final results of bicycle races, they used XGBoost as the algorithm to build the whole architecture, and finally predicted the top10 rider rankings of Tour of Flanders. Although research above can predict the final game result, but Tour of Flanders is a one-day game, unfortunately it cannot predict the result of the multi-day game and lacks substantial application.

In (Logg et al., 2018), it concluded that people often choose their own decisions between the results of the algorithm and their own decisions, because most people still have low trust in the algorithm. It's important to build a trustworthy AI model. The study (Lee, 2021) organizes various data about trust in AI model. (Hoff & Bashir, 2015) divided trust into three types: dispositional trust, situational trust and learned trust. Among the dispositional trust, there are some attributes will mainly affect their trust, such as culture, age and gender, etc. Situational trust includes professional knowledge, risk, and mood, etc. Learned trust can be divided into initial learned trust and dynamic learned trust. Dynamic learned trust is particularly important, users will become more familiar with this model if they use it repeatedly, which will increase dynamic learned trust. Therefore, the transparency of the model is particularly important. If it is easier for users to understand the internals of the model, the trust will increase. In (Thiebes, 2021), the concept of trustworthy AI and its five basic principles are introduced, namely beneficence, non-maleficence, autonomy, justice and explicability. The explicability corresponds to the transparency of the model (Hoff & Bashir, 2015). (Samek, 2017) gives a definition in the AI of explicability, that is, the methods and technologies that allow experts to understand the achievements of AI, and classify them into the following aspects, namely verification of the system, improvement of the system, learning from the system and compliance to legislation.

3. Method

This paper proposes a trustworthy AI strategy, which allows users to have a reference direction in gambling applications, so as to make trustworthy investments and bets. We have designed a complete betting system, which combines the prediction of the top10 in cycling events, betting profit prediction, and trust calculation and optimization, as shown in Figure 1. The input part is a stake, and then after our ML module, it is divided into two parts, one part will output our profit directly through the odds calculator, the other part will enter our trust module to calculate the trust of our model, and finally according to the score of the trust, the system will redistribute the stake, and finally calculate the optimized profit.

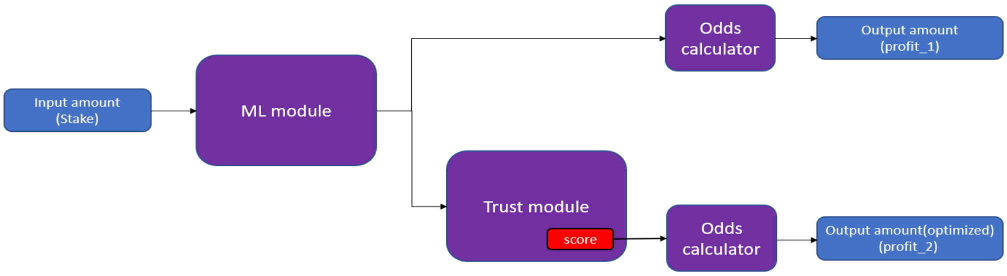


Figure 1 System structure

The data source for this study is from https://cqranking.com/, which contains information on many riders. The ranking calculation benchmark used is CQ ranking, and CQ scores are used to calculate rankings and so on. The preliminarily captured data is divided into two parts. The first part is the basic information of each rider, and the second part is all the events that each rider participated in. As this study made predictions for Tour *de* France, only the teams qualified to participate in Tour *de* France were selected, and the data of a total of 18 teams were selected. Through the basic information of the riders extracted above and the participation of each rider, 17 features were extracted and all features are then normalized to form a Gaussian distribution. According to the betting rules, the profiles were labeled as whether they belonged to the top10 riders or not (0=NO and 1=YES).

These 17 features are listed as follows: (1) rank_start, (2) point_start, (3) win, (4) podium, (5) top_ten, (6) points, (7) race_days, (8) win_dauphine, (9) podium_dauphine, (10) top_ten_dauphine, (11) points_dauphine, (12) win_swiss, (13) podium_swiss, (14) top_ten_swiss, (15) points_swiss, (16) climb, (17) sprint. In the pre-processing part, 17 different features are mentioned, which can be roughly divided into three categories.

D1: the basic data of the riders (features (1) to (7)).

D2: the related races of Tour *de* France (features (8) to (15)).

D3: the riders' specialization (features (16) to (17)).

The first category of data *D1* of the riders is based on the statistics of each rider's past competition data. The second category of data *D2* considers the statistics of races that are highly related to Tour *de* France to improve the impact on the overall training results. In (Sheehan, 2018), it shows that Criterium *du* Dauphine ($D2.Dau$) and Tour *de* Suisse ($D2.Sui$) are often regarded by riders as the pre-race training of Tour *de* France. The correlation among these races, whether from the schedule and terrain, or past data, shows the importance of these two races to Tour *de* France (Lowe, 2016). In the third category of data *D3* are riders' specialization, namely climbing ($D3.climb$) and sprint points ($D3.sprint$).

Among these three categories of data, we try to find the best combination to get the best training results. Five data sets for training are given as follows.

$$Train_1 = D1 + D2.\mathrm{Dau} + D3.\mathrm{climb} + D3.\mathrm{sprint} \tag{1}$$

$$Train_2 = D1 + D3.\mathrm{climb} + D3.\mathrm{sprint} \tag{2}$$

$$Train_3 = D1 + D2.\mathrm{Dau} + D3.\mathrm{climb} \tag{3}$$

$$Train_4 = D1 + D2.\mathrm{Dau} + D3.\mathrm{sprint} \tag{4}$$

$$Train_5 = D1 + D2 + D3 \tag{5}$$

The following describes the process and algorithm for calculating the trust in sequence. The central idea of this research is that the closer the predicted input and output are to the input and output of learning samples, the more credible the prediction is. As shown in Figure 2, we use SMOTE (Synthetic Minority Oversampling Technique) to balance our data for the test data (Test_X) and the prediction data (pred_Y) output by the ML module. SMOTE is a synthetic data algorithm based on comprehensive sampling, which is used to solve the problem of data imbalance. Because in this study, there are fewer labels in the top10, SMOTE is chosen to improve our data, and then input to the KNN (K-Nearest Neighbor) model for fitting, and then use the original learning samples

(Train_X) to predict the trained KNN, the above actions are equivalent to finding matching points in the distribution generated by the test data (Test_X) and the prediction data (pred_Y) output by the ML module, and finally output the prediction data (KNN_predict), so we can compare this data with the original learning sample.

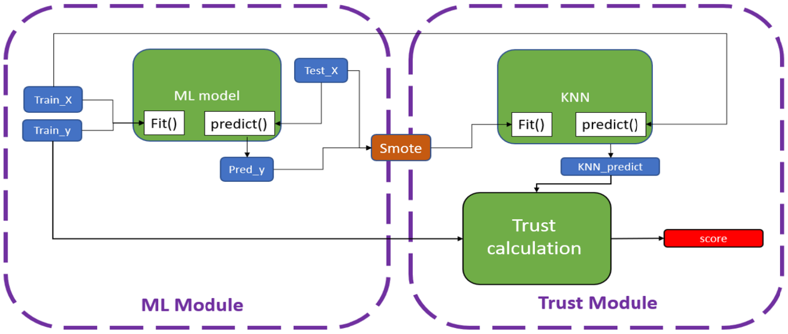


Figure 2 Detailed content of ML module and Trust module

Defining and calculating the similarity of the data will be of great help to the calculation of the trust later. Therefore, a suitable metric must be selected to measure and help to judge the difference of the data, and we finally choose to use the Euclidean distance to carry out the experiment, assuming that in the Euclidean space, the equations solved for point x = $(x_1,\ldots,x_n)$ and point y = $(y_1,\ldots,y_n)$ Euclidean distance can be expressed as follows.

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \tag{6}$$

We calculate each distance d between the prediction and training data and then use the following formula.

$$\text{Trust} = \frac{1}{(1+d)} \tag{7}$$

As our similarity, the calculated range is between 0 and 1, the closer it is to 1, the more similar it is, and the closer it is to 0, the less similar it is. We first compare the results predicted by KNN with the training sample data (train_y), and then divide it into two parts. If the prediction is correct, it means that the trust will be higher. At this time, the distance is selected according to the predicted label. If the prediction is incorrect, representing a low trust, a distance must be given to punish, so theoretically, the distance must be selected according to the label value opposite to the prediction, but there will be some special cases, such as the shortest penalty distance, in this case, the average of several other distances must be chosen for the calculation.

The flow chart is shown in Figure 3. Assume that (d1,d2,d3) represent a certain distance which is the results output by a function *Kneighbors*(train_X, n_neighbors=3).

First, we first judge whether the results predicted by KNN are equal to the data of the original training samples (train_y). If they are equal, we will find the corresponding y value in (d1,d2,d3) according to the result predicted by KNN, finally select all the found distances, and take the average and throw it into the similarity calculation to calculate the final score.

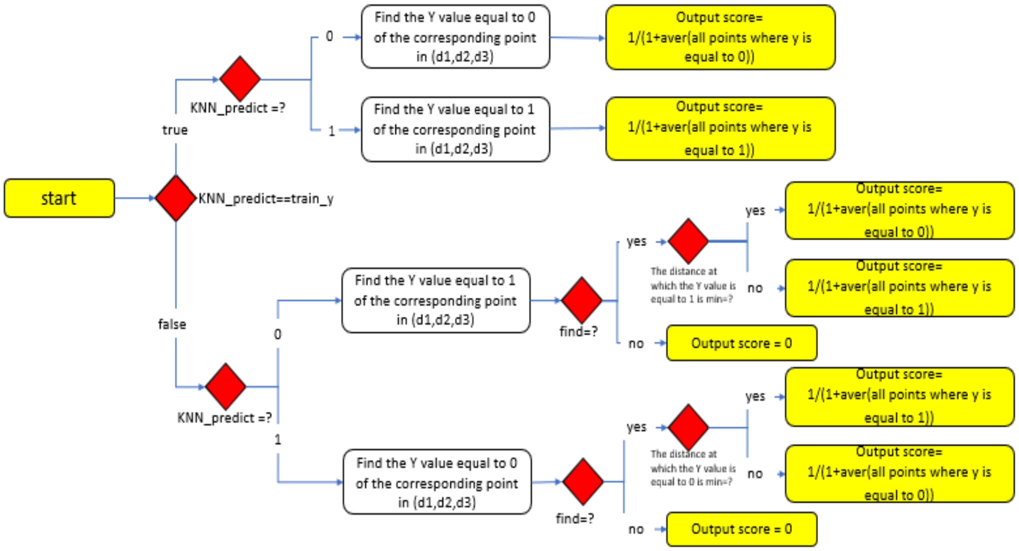
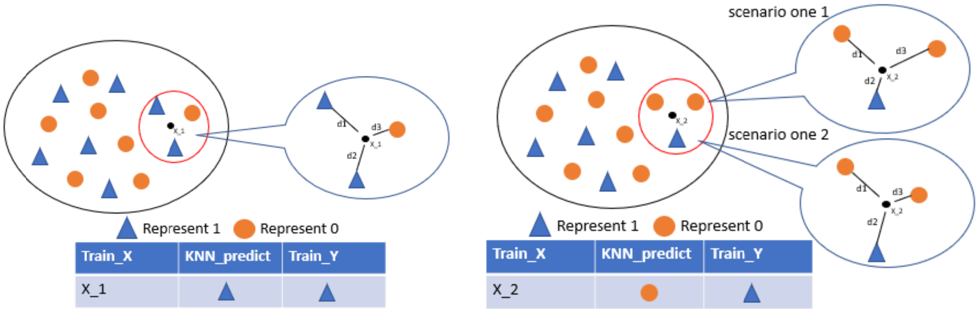

start
KNN_predict==train_y
true
KNN_predict =?
0
Find the Y value equal to 0 of the corresponding point in (d1,d2,d3)
Output score= 1/(1+aver(all points where y is equal to 0))
1
Find the Y value equal to 1 of the corresponding point in (d1,d2,d3)
Output score= 1/(1+aver(all points where y is equal to 1))
false
KNN_predict =?
0
Find the Y value equal to 1 of the corresponding point in (d1,d2,d3)
find=?
yes
The distance at which the Y value is equal to 1 is min=?
yes
Output score= 1/(1+aver(all points where y is equal to 0))
no
Output score= 1/(1+aver(all points where y is equal to 1))
no
Output score = 0
1
Find the Y value equal to 0 of the corresponding point in (d1,d2,d3)
find=?
yes
The distance at which the Y value is equal to 0 is min=?
yes
Output score= 1/(1+aver(all points where y is equal to 1))
no
Output score= 1/(1+aver(all points where y is equal to 0))
no
Output score = 0

Figure 3 Trust calculation flow chart

As shown in Figure 4(a), KNN_predict and train_yare equal at this time, and the distances represented by d1, d2 and d3 are as follows. We select the average of d1 and d2 to bring into the similarity calculation formula for calculation, because the labels represented by these two distances are both same as KNN_predict. Let the function *aver*(d1, d2) be the average of the two, the similarity score can be expressed as follows.

$$\text{Trust} = \frac{1}{1+aver(d_1,d_2)} \tag{8}$$



Train_X	KNN_predict	Train_Y
X_1	▲	▲

Train_X	KNN_predict	Train_Y
X_2	●	▲

Figure 4 Data distribution of two special cases (a) and (b)

The following is the case where the result predicted by KNN is not equal to the data of the original training sample. As in the case of equality, first determine that the result predicted by KNN is 0 or 1. If it is equal to 0, we will look for the distances represented by label 1 in (d1, d2, d3), if not found, the trust score is 0. If it is found, it is judged whether the distance represented by the label 1 in (d1, d2, d3) is smaller than the other distances. If it is smallest in three distances, the average of all the distances represented by label 0 is taken into the similarity formula for calculation, otherwise, the average of all the distances represented by label 1 is taken into the similarity formula for calculation. As shown in Figure 4(b), in scenario one, because KNN_predict is equal to 0, and d2 is smaller than d1 and d3, we choose the average of d1 and d3 to be added to the similarity formula calculation, and the similarity can be expressed as follows.

$$\text{Trust} = \frac{1}{1+\text{aver}(d_1,d_3)} \tag{9}$$

But in scenario two, d2 is not the smallest distance, so we only select d2 and bring it into the similarity formula for calculation.

$$\text{Trust} = \frac{1}{1+d_2} \tag{10}$$

This part is the final part of the whole system. The above results are passed through the odds calculator to output the final amount of whether to add the trust score. The following is divided into odds calculation and final output algorithm.

According to the conception of this study, the ultimate goal is to predict the top10 riders of Tour *de* France, and finally make bets. In order to calculate the final output amount, it is a necessary step to simulate the odds of each rider. According to the above, the odds can be expressed as the following formula

$$\text{Odd} = \frac{1}{\text{probabilty of top ten}} \tag{11}$$

That is, it is inversely proportional to the probability of top10. Then we apply the activation function contained in the above model to simulate the odds, because through the activation function, we can see the probability of top10 and non-top10, and finally add the range of odds that can be set by ourselves, which can be expressed as the following formula:

$$Odds = \left(\frac{1}{P} - 1\right) \times Odds_{range} \tag{12}$$

P is the probability of top10, that is, the probability output by the activation function, and Odds_range is the range of odds that can be set by ourselves. For example, if Odds_range is set to 10 in this study, the output range of the odds calculation will fall between from 1 to 10.

There are two cases in this part, the amount output without adding the trust score and the amount output with adding the trust score, which will be discussed separately below.

(1) Without adding the trust score. Combine the results of the ML module with the odds calculator to get at the final amount.

(2) Adding the trust score. This part is the amount output by adding the trust score.

We cooperate with the trust degree to adjust the stake of the rider who needs to bet predicted by the ML module, so that the rider with low odds bet a little more money, and vice versa. The main reason is to adopt a conservative strategy to increase the money we have won and reduce the possible loss. In order to choose which riders need to raise or lower their stakes, we set an odds benchmark, and we will reduce the capital if the bettor's odds are lower than this benchmark, and increase the capital otherwise. The odds benchmark is equivalent to the amount of risk taken, and above this standard is equivalent to an increase in risk and therefore a reduction in the stake, and vice versa. We use trust to calculate the odds benchmark, the formula is as follows

$$Oddsbenchmark = (Trust * W) \times Odds_{range} \tag{13}$$

Multiply the previously calculated trust score by W (Trust Weight) and finally multiply by Odds_range. Through the above algorithm, the range of the odds benchmark can be set to be the same as the odds calculated in the previous step.

Finally, the final output amount is calculated according to the following process, as shown in Figure 5. It is mainly to determine whether the estimated odds of each data to be bet are lower than the odds benchmark. If it is lower, increase the stake, otherwise decrease the stake. After setting the amount of capital increase and decrease, you can use the results predicted by the ML module to bet, and finally the final output amount can be obtained.

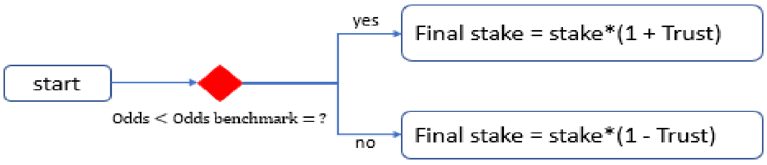

start
Odds < Odds benchmark = ?
yes
Final stake = stake*(1 + Trust)
no
Final stake = stake*(1 - Trust)

Figure 5 Flow chart of the stake after adding the trust reference

4. Experimental results

This section presents the experimental results of this study, and introduces our training and testing data, as well as the input and output of the entire cycling gambling system in sequence. The data from 2014 to 2018 were selected as training and testing data, with a total of 648 data. There are two types of labels: riders who belong to the top10 of Tour *de* France and riders who do not belong to the top10 of Tour *de* France. The sample numbers of the top10 riders and the non-top10 riders are 50 and 598 respectively. The final split ratio of training dataset and test dataset is 2:1. We predict the outcome of Tour *de* France from 2019 to 2021. The input and output of this experiment are all amounts of money. Suppose the input is X, which means that I bet X dollars each for the top10 riders predicted. The input of this study is preset to 100 dollars, and the final output is profit Y_1 dollars, and the profit Y_2 dollars generated by adding the trust reference will also be output.

From the distribution curve and variance of 17 features, we found that the variance of some features is too large, which will cause the model to fail to converge. Therefore, it is necessary to find a distribution that can quickly converge the model and the not too different from the original data. In Section 3, we normalized the data. The reason is to improve the convergence speed of the model and make each feature contribute to our results to the same extent. In this section, we show the results produced by using 2 different normalization methods, which are mapped to Gaussian distribution and Uniform distribution, respectively. Table 1 is the comparison of Fl scores of different training data using different feature mapping methods in different models. We can clearly find that the result of mapping to Gaussian distribution is better than Uniform distribution.

Table 1 Comparison of Fl scores of different training data using different methods (LR = Logistic Regression, NN = Neural Network, DL = Deep Learning)

ML Method		LR		NN		DL	
normalization		Gaussian	Uniform	Gaussian	Uniform	Gaussian	Uniform
Data set	Train_1	0.67	0.29	0.80	0.64	0.64	0.35
	Train_2	0.63	0.29	0.61	0.55	0.56	0.50
	Train_3	0.63	0.15	0.67	0.52	0.69	0.50
	Train_4	0.47	0.15	0.60	0.64	0.60	0.13
	Train_5	0.55	0.55	0.64	0.64	0.59	0.50

Table 2 shows the comparison of precision and F1 scores across different training datasets and models. Looking at the models, it can be clearly seen that the precision of the NN-trained data is higher. However, the deep learning model performed poorly, possibly due to insufficient training data, which may explain its relatively weak results. This suggests that we cannot blindly apply deep learning to every problem, especially when data is limited. From a data perspective, this is an important reminder.

Table 2 Comparison of Precision and F1 Score Across Training Data and Models (LR = Logistic Regression. NN = Neural Network. DL = Deep Learning.)

		Train_1	Train_2	Train_3	Train_4	Train_5
Precision	LR	1.00	0.86	0.86	0.80	0.60
	NN	1.00	0.64	0.67	0.75	0.62
	DL	0.70	0.54	0.59	0.75	0.53
F1 Score	LR	0.67	0.63	0.63	0.47	0.55
	NN	0.80	0.61	0.67	0.60	0.64
	DL	0.64	0.56	0.69	0.60	0.59

When we compare Train_1 and Train_2, we can see that the data of D2 is missing, which causes the performance of Train_2 to drop, so it shows that D2 is a very important feature. Then we observe Train_1 and Train_3, and we can see that $D3.sprint$ is missing, which causes the performance of Train_3 to drop, so it means that D3 must include $D3.climb$ and $D3.sprint$ to achieve the best performance. In the comparison of Train_1 and Train_4, you can also see the same results. We can also notice that in the comparison of Train_3 and Train_4, the importance of $D3.sprint$ is higher than $D3.climb$. Then, in the comparison between Train_2 and Train_5, it can be seen that Train_2 lacks D2, but the training results are better, indicating that adding 2 D2 at a time has a negative result on the accuracy rate. Continue to observe Train_1 and Train_2, you can see that $D2.Dau$ is missing, which causes the performance of Train_2 to drop, which shows that $D2.Dau$ is beneficial to the whole training. However, when comparing Train_1 and Train_5, it can be found that the addition of $D2.Sui$ to Train_5 causes a significant drop in precision, which also shows that only the addition of $D2.Dau$ will have a better impact on the results. According to the above, it can be seen that the importance of $D2.Dau$ is far greater than that of $D2.Sui$, $D3.sprint$ are higher than $D3.climb$, and the importance of D3 is also greater than that of D2. Therefore, we can get a ranking of importance in D2 and D3, the ranking is as follows:

$$D3.sprint > D3.climb > D2.Dau > D2.Sui \tag{14}$$

The reason for this importance may be that in the multi-day Tour *de* France, almost every day has a sprint point and a climbing point, and there will be at least one day of time trials. Time trials are good for sprinters. Therefore, the overall impact of sprinting on this multi-day schedule will be slightly greater, but there are also a lot of climbing points, so if we want the prediction to be better, it is best to take both. In addition, D2.Dau is not as important as D3, but it is also obvious that after adding training data, the precision has increased significantly. Amount optimization results of different weights for different training data in NN are shown in Table 3, and the stake is 100 dollars.

In this part, we discuss different trust weights (W) for the final optimized output results. The results are as follows. Amount optimization results of different weights for different training data in NN are shown in Table 3, and the stake is 100 dollars. Tables 4 and 5 show the profit growth rate of different weights for different training data in NN. We can find that the higher the weight of most data, the amount will increase or remain unchanged, except for Train_2, that is, when the original profit is negative, the higher the weight, the negative growth of the amount. Looking at the profit margin in Table 4, you can see the whole trend more clearly, and then in Table 5, you can find that as the weight increases, the best profit growth margin (in the case of positive profit) (Best_case(positive profit)) is no change. At the worst profit growth rate, with the increase of the weight, there is a gradual improvement, and it will not decrease until the weight is 1.2. Then, after observing the mean and the number of variance and combining the above, we select the result that the variation is the smallest and the average profit margin is not too small is used as the final weight, that is, the final trust weight of our research is 1.1. Choosing the weight equal to 1.1 can keep our profits growing positively, and the average profit margin can also maintain a certain level, and it is also the smallest under the worst profit growth margin (Worst_case).

Table 3 Amount optimization of different trust weights for different training data in NN

Methods:NN	before optimization	W = 0.8	W = 0.9	W = 1	W = 1.1	W = 1.2
Train_1	1161	1296	1296	1296	1608	1608
Train_2	-72	67	67	-18	-18	-103
Train_3	1090	1115	1115	1115	1115	1115
Train_4	534	751	751	751	751	751
Train_5	997	747	959	1042	1042	1042

Table 4 shows the profit growth rate of different weights for different training data in NN. We can find that the higher the weight of most data, the amount will increase or remain unchanged, except for Train_2, that is, when the original profit is negative, the higher the weight, the negative growth of the amount. Looking at the profit margin, the whole trend can find that as the weight increases, the best profit growth margin is no change. At the worst profit growth rate, with the increase of the weight, there is a gradual improvement, and it will not decrease until the weight is 1.2. Then, after observing the mean and the number of variance and combining the above, we select the result that the variation is the smallest and the average profit margin is not too small is used as the final weight, that is, the final trust weight of our research is 1.1. Choosing W = 1.1 can keep our profits growing positively, and the average profit margin can also maintain a certain level, and it is also the smallest under the worst profit growth margin (Worst_case).

Table 4 Various statistics of profit growth rate of different trust weights for different training data in NN

Methods:NN	W = 0.8	W = 0.9	W = 1	W = 1.1	W = 1.2
Train_1	0.11	0.11	0.11	0.38	0.38
Train_2	1.93	1.93	0.75	0.75	-0.43
Train_3	0.02	0.02	0.02	0.02	0.02
Train_4	0.40	0.40	0.40	0.40	0.40
Train_5	-0.25	-0.03	0.04	0.04	0.04
Best_case	0.40	0.40	0.40	0.40	0.40
Worst_case	-0.25	-0.03	0.02	0.02	-0.43
Mean	0.442	0.486	0.264	0.318	0.082
Variance	0.745	0.679	0.097	0.090	0.114

Table 5 shows the comparison of the final output amount of different training data (Train_1, Train_2, Train_3, Train_4 and Train_5) in NN, and the stake is also 100 dollars. It can be seen from the above results that after adding the consideration of trust, we adjusted our stake according to the trust and odds, and the profit increased, so that the amount earned by the user increased and the amount lost decreased.

Table 5 Comparison of the final output amount of different training data in NN

Methods:NN	Amount without trust	Amount with trust	Profit growth	Trust
Train_1	1161	1608	0.38	0.384
Train_2	-72	-18	0.75	0.427
Train_3	1090	1115	0.02	0.390
Train_4	534	751	0.40	0.406
Train_5	997	1042	0.04	0.343

The trust algorithm proposed in our experiment is heuristic. After the comparison in the previous section, it can be found that the use of trust optimization can increase profits. For this reason, we also present an experiment with the ideal value of trust, and we sample some data for experiments, so that we can compare the gap between our formula design and the ideal value. The sampling ratio is 40% of the original data. The results are shown in Table 6. The following results take NN as an example, and we can see the gap between our sampling trust and the ideal value is between 4% and 8%.

Table 6 Taking NN as an example, the gap between different training data and the ideal value of trust

Methods:NN	Ideal value	Trust after sampling	Gap percentage
Train_1	0.384	0.362	6
Train_2	0.427	0.405	5
Train_3	0.390	0.372	4
Train_4	0.406	0.379	7
Train_5	0.343	0.317	8

After proposing the algorithm of trust score, we re-examined the part of deep learning. In the aforementioned results, we know that the result of deep learning is the worst. We believe that the main reason is the insufficient amount of data. For this reason, we conduct an additional experiment to discuss the relationship between the amount of data and the degree of trust. We use different data sizes for training, using deep learning as a model, and the data size adjustment method is SMOTE. We increase the data of a small number of samples to different degrees to achieve the purpose of different data sizes. Table 7 shows the comparison of the amount of different training data of on the precision and trust (the left side of the brackets of the data volume ratio is the total number of data with a label equal to zero, that is, the riders who are not in the top10, and vice versa are the riders in the top10).

Table 7 Comparison of the amount of training data on the precision and trust

	data ratio(394,36)		data ratio(394,216)		data ratio(394,394)	
	Precision	Trust	Precision	Trust	Precision	Trust
Train_1	0.26	0.3764	0.56	0.3844	0.60	0.3844
Train_2	0.38	0.4314	0.50	0.4336	0.40	0.4282
Train_3	0.50	0.3107	0.44	0.3112	0.53	0.3125
Train_4	0.38	0.4073	0.38	0.4001	0.55	0.4091
Train_5	0.46	0.3459	0.53	0.3461	0.53	0.3435

In the part of data proportion, in Train_1 and Train_2, the amount of data with label 1 is adjusted to be the same as the amount of data with label 0, that is, there are 394 results. In Train_3, Train_4 and Train_5, in the data In the process of adding the data volume to 394, the result is continuous deterioration, so we choose to reduce the increase in the data volume, so that the change is easy to see. In Train_1, it can be seen that the amount of data increases, which can increase the precision slightly, but the precision of

216 volume and 394 volume is similar, and even 216 volume is better. In the trust part, as the amount of data increases, the trust has slight increase, but it is not very obvious, and it can also be found that the trust increases or decreases with the precision rate. In Train_2, it can be seen that with the increase of the amount of data, the precision first increases, and then decreases. The increase of the amount of data can indeed increase the precision, but there is a limit, and it will decrease after the amount of data. In the trust part, with the amount of data increases, trust increases slightly, but there is also a limit, and it can also be found that trust increases or decreases with the accuracy rate. In Train_3, it can be seen that the amount of data increases, and the precision first drops a lot, and then increases again. In the part of the trust, as the amount of data increases, the trust increases slightly. In Train_4, it can be seen that the amount of data increases, and the precision first drops a lot, and then rises. In the trust part, as the amount of data increases, the trust first drops a lot, and then rises again. In Train_5, it can be seen that the amount of data increases, and the precision increases. In the trust part, as the amount of data increases, the trust first drops a lot, and then rises again.

On the whole, increasing the amount of data for a few tags can indeed increase the precision, but there is also a limit, after which it will decrease. In the part of trust, in most cases, with the increase of the amount of data, the trust increases, but the increase is not large. There is also a strong relationship between precision and trust, and in most cases, trust increases or decreases with precision. Therefore, it can be concluded that the amount of training data does affect the precision, which can be learned by observing the trust, and we may be able to use the trust as a benchmark for data augmentation research in the future.

This section will show how we actually use the cycling gambling system, the stake is 100 dollars, and we predict the results of Tour *de* France from 2019 to 2021. The final results are shown in Table 8, we can find that precision of top1 in 2019 and 2021 is about 60%, and the profit has grown after optimization, but the forecast result in 2020 is slightly less than ideal. Therefore, we speculate the result in 2020, we found that because of the epidemic, many cycling races have been suspended, and the predicted data is about 25% less than other data on average. The lack of data will also affect the training effect of NN, so the prediction results are not too good, but even so, we consider the amount after trust optimization, there is still an increase.

This section explores what problems the system will encounter and how to solve it when large amounts of data are added in the future. In the trust module of this system, KNN is used as the core. However, when KNN encounters a large amount of data, there will be a time complexity problem, so a method must be found to speed up. We refer to the methods proposed by (Johnson, 2019) to speed up our KNN, which is to implement KNN with Facebook AI Similarity Search (FAISS). According to the conclusion obtained by the combination of training data used by (Adamczyk, 2020), the training time can be reduced by 300 times, and the prediction time can be reduced by a maximum of 17 times. The final time comparison results after optimization are shown in Table 9 and Table 10. It can be found that the training time is reduced, but the prediction time is not reduced. We speculate that the main reason is the number of data, the amount of data used is large, the model training and prediction time is longer than our research, and it is easier to see the effect of acceleration if the time is long (we can know by comparing the time complexity, assuming n data, time complexity of KNN is O(n), and the time complexity of FAISS is O(log n)). However, because of the small amount of data, the measured time is shorter, and the acceleration effect is not so obvious. Especially in the part of the prediction time, so there will be fluctuations.

Table 8 Results of Tour *de* France from 2019 to 2021

	Precision	Amount without trust	Amount with trust	Profit growth	Trust
Predict_2019	0.62	974	1030	0.05	0.252
Predict_2020	0.38	529	563	0.06	0.284
Predict_2021	0.57	628	757	0.20	0.250

Table 9 Comparison of training time using KNN and FAISS

	KNN Train time(s)	FAISS Train time(s)	Reduction factor
Predict_2019	0.000473	0.000185	2.36
Predict_2020	0.000514	0.000216	2.37
Predict_2021	0.000520	0.000166	3.13

Table 10 Comparison of predicting time using KNN and FAISS

	KNN Predict time(s)	FAISS Predict time(s)	Reduction factor
Predict_2019	0.004173	0.006644	-1.59
Predict_2020	0.005362	0.006251	-1.16
Predict_2021	0.004990	0.005860	-1.17

The above shows and analyzes the results of the entire Tour *de* France event prediction and betting reward prediction system, which can be roughly divided into two parts for discussion, the ML module and the trust module. In the ML module, NN has the best learning results, and the logistic regression is slightly inferior to NN. However, the effect of deep learning is not very good, probably because the training data is not large enough. In the trust module, we choose the ones with better performance in the ML module to conduct more in-depth research. It can be seen that in the above results, the addition of trust will indeed affect the profit.

On the other hand, in the comparison of different training data (Train_1, Train_2, Train_3, Train_4 and Train_5), Train_1 performed the best, that is, the basic data of the first type of players and $D2.Dau$, as well as the third category of rider specializations ($D3$.sprint and $D3.climb$), these factors have a greater impact on the prediction.

We also discussed some applications related to trust. We obtained the gap between our formula and the ideal value. We also discussed the relationship between the amount of training data and trust. In most cases, we can know that the accuracy and trust will increase with the amount of data, but not in a few cases. The reason for the above may be that the data we increase by using SMOTE may not be very good, because it is not in the same way as old data was obtained. Based on the above results, and watching how our training data increases, the training results are still inferior to NN, so in betting systems, we still recommend using NN as our model. We also show our complete gambling system based on the above conclusions, and predict the results of Tour *de* France from 2019 to 2021, and also show that through our gambling system, the profit can indeed be optimized.

5. Conclusion

AI has changed people's lives, and many studies have been generated. This research focuses on cycling races and analyzes the trust issues generated by current AI. A gambling prediction system is constructed as a whole. In terms of predicting Tour *de* France, we succeeded in predicting whether each rider was in the top10 and using different training profile to achieve multi-day event predictions. In terms of ML modules, we try different ML methods, and finally NN has the best results, and finds the combination of training data that can make the model result the best. In terms of trust, we propose a new trust algorithm. On the whole, we first get a trust score, which allows users to evaluate the entire model, and we apply this score to the part that maximizes profit, and the results achieve what we want.

Reference

Adamczyk, J. (2020). *Make kNN 300 times faster than Scikit-learn's in 20 lines!* https://towardsdatascience.com/make-knn-300-times-faster-than-scikit-learns-in-20-lines-5e29d74e76bb

De Spiegeleer, E. (2019). *Predicting cycling results using machine learning.*

Das, S. (2019). *Trust issues: Is AI black box creating a black future?* https://analyticsindiamag.com/trust-issues-is-ai-black-box-creating-a-black-future/

David, J. A., Pasteur, R. D., Ahmad, M. S., & Janning, M. C. (2011). NFL prediction using committees of artificial neural networks. *Journal of Quantitative Analysis in Sports, 7*(2).

Gilchrist, S. (2020). *A brief history of cycling and cycling betting.* https://kayokokishimoto.com/cycling-betting/

Gabriele, M. C. (2011). *The golden age of bicycle racing in New Jersey.* The History Press.

Hoff, K. A., & Bashir, M. J. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data, 7*(3), 535–547.

Kataoka, Y., & Gray, P. (2018). Real-time power performance prediction in Tour de France. In *International Workshop on Machine Learning and Data Mining for Sports Analytics* (pp. 121–130). Springer.

Kholkine, L., De Schepper, T., Verdonck, T., & Latré, S. (2020). A machine learning approach for road cycling race performance prediction. In *International Workshop on Machine Learning and Data Mining for Sports Analytics* (pp. 103–112). Springer.

Lee, E. J. (2021). How do we build trust in machine learning models?

Logg, J., Minson, J., & Moore, D. A. (2018). Do people trust algorithms more than companies realize? *Harvard Business Review, 26*.

Mutijarsa, K., Ichwan, M., & Utami, D. B. (2016). Heart rate prediction based on cycling cadence using feedforward neural network. In *2016 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)* (pp. 72–76). IEEE.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.

Sheehan, M. (2018). *Tour de Suisse vs. Criterium du Dauphiné: What's the best tour prep?* https://www.flobikes.com/articles/6206572-tour-de-suisse-vs-criterium-du-dauphine-whats-the-best-tour-prep

Soneja, A. (2019). *Artificial intelligence in sports: A smarter path to victory.* https://www.cio.com/article/3400877/artificial-intelligence-in-sports-a-smarter-path-to-victory.html

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. *Annals of Data Science, 6*(1), 103–116.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464.

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.

Evaluating the Baseball Innovation Frontier: An Innovation Index and Comparative Analysis via LLM-Integrated MCDM

Chien Min Kuo
Baseball Innovation League Association, USA
kuo@baseballinnovation.org

ABSTRACT

Innovation in baseball technology has accelerated across player tracking systems, AI-powered coaching platforms, automated officiating, smart equipment, and immersive virtual training environments. However, a standardized and transparent framework for systematically evaluating these innovations remains limited. This study introduces the **Baseball Innovation Award Index (BIAI)**, a structured evaluation model integrating Multi-Criteria Decision Making (MCDM) with Analytic Hierarchy Process (AHP) weighting and large language model (LLM)-assisted comparative assessment. Seven core criteria are defined: Performance and Competitive Impact, Safety Health and Longevity, Data Intelligence and Reliability, Adoption and Scalability, Innovation Novelty and Originality, Fan Engagement and Media Experience, and Cost Efficiency and Return on Investment. Criterion weights are derived through AHP to ensure methodological rigor and consistency. All indicators are scored on a standardized 0–100 benefit scale and aggregated using a linear weighted model to generate composite BIAI scores. The framework is applied to representative contemporary baseball technologies, enabling cross-category comparison and ranking. Findings indicate that AI-integrated analytics and scalable data-driven platforms demonstrate superior innovation performance, particularly in competitive impact and ecosystem integration. The proposed BIAI provides a replicable, transparent, and extensible benchmarking tool for researchers, leagues, investors, and technology developers within the evolving baseball innovation ecosystem.

Keywords**:** Baseball, Innovation, LLM, MCDM, AHP

1. Introduction

The background, motivation, purpose, and contributions are presented in the following sections.

1.1 Background

Baseball has experienced rapid technological advancement driven by developments in artificial intelligence, sensor systems, biomechanics, and data analytics. Modern innovations such as player tracking systems, automated ball-strike technology, smart equipment, AI-powered coaching platforms, and immersive virtual training tools have significantly influenced performance optimization, injury prevention, officiating accuracy, and fan engagement. Despite this growth, there is no unified framework for systematically evaluating and comparing diverse baseball technologies across multiple performance dimensions. Existing assessments often focus on isolated metrics, limiting cross-category comparison. Therefore, a structured and transparent evaluation model is needed to benchmark innovation performance within the evolving baseball technology ecosystem.

1.2 Motivation

The rapid expansion of baseball technologies has created a need for objective and systematic evaluation tools. While numerous innovations claim to enhance performance, safety, or fan experience, comparisons are often based on isolated metrics, marketing claims, or subjective judgment. This lack of a unified benchmarking framework makes it difficult for leagues, teams, investors, and researchers to assess relative innovation value across heterogeneous technologies. The motivation of this study is to develop a transparent, replicable, and defensible evaluation index that integrates structured decision-making methods with AI-assisted comparative analysis. By establishing a standardized innovation index, this research aims to support evidence-based decision-making and promote clarity within the evolving baseball technology landscape.

1.3 Purpose

The purpose of this study is to develop and validate a structured innovation evaluation framework tailored to baseball technology. Specifically, this research proposes the Baseball Innovation Award Index (BIAI), which integrates Multi-Criteria Decision Making (MCDM) and Analytic Hierarchy Process (AHP) weighting with LLM-assisted comparative assessment. The study aims to (1) define key innovation criteria relevant to baseball technology, (2) establish defensible criterion weights, and (3) apply the index to representative contemporary innovations to enable systematic comparison and ranking. Through this approach, the research seeks to provide a transparent and scalable benchmarking tool for evaluating innovation performance in the baseball ecosystem.

1.4 Contribution

This study contributes a structured and transparent evaluation framework for assessing innovation in baseball technology. First, it introduces the Baseball Innovation Award Index (BIAI), which integrates MCDM and AHP to quantify multidimensional innovation performance using a standardized scoring model. Second, it formalizes clear evaluation criteria that capture technical impact, safety, scalability, data intelligence, and ecosystem value, enabling cross-category comparison. Third, it demonstrates the integration of LLM-assisted assessment within a formal decision-making framework to enhance consistency and scalability. By providing a replicable benchmarking approach, this research supports objective innovation comparison and contributes to methodological advancement in sports technology evaluation.

1.5 Related Works

Smith et al. (2020) focused on sports technology, using patents and products with weighted counts to assess technical novelty; however, their framework overlooks market adoption and user engagement. Lee and Kim (2021) evaluated baseball equipment via field-tested prototypes using an MCDM-AHP approach, emphasizing performance and safety, but their analysis was limited to a few metrics. Zhao et al. (2019) examined general technology innovations through citation analysis of academic papers, highlighting research impact while ignoring practical, real-world implementation. The differences between the BIAI (this study) and prior approaches are summarized in Table 1.

Table 1 Comparison of Innovation Evaluation Approaches

Feature / Study	Smith (2020)	Lee (2021)	Zhao (2019)	BIAI (This Study)
Criteria Identification	Expert opinion	Literature & expert review	Citation analysis	LLM & Expert
Criteria Coverage	Technical novelty only	Performance & safety	Research impact	7 dimensions: tech & management
Implemented & Potential Innovations	Mostly patents/products	Only real devices	Only papers/citations	products/prototypes + patents/papers
Weighting Method	None / simple counts	AHP for few criteria	None	AHP & LLM
Scoring Standardization	Varies; not standardized	Partial normalization	Citation-based	0–100 benefit-type
Practical Benchmarking/Awards	No	No	No	Rank, Benchmark, Award
Market Adoption & Engagement	No	No	No	Adoption & fan/media
Transparency & Replicability	Low	Medium	Low	High; defensible & replicable

2. Methodology

2.1 Data Collection

Data for evaluation were collected from multiple sources to capture a comprehensive view of baseball device innovations. Primary data included commercially available products, field-tested prototypes, and documented performance metrics. To capture pioneering contributions, relevant patents and peer-reviewed academic papers were also considered; however, these were treated as **potential innovations** rather than fully implemented products to avoid inflating scores. Selection criteria ensured that all data were **current, relevant, and verifiable**, supporting a robust and replicable evaluation framework.

2.2 Innovation Index Construction

2.2.1 LLM-Assisted Criteria Identification and Synthesis

LLMs were used exclusively to assist in identifying and synthesizing potential evaluation criteria from existing literature. All weighting and scoring decisions were performed using formal MCDM procedures and/or human expert validation. For evaluating baseball innovations, the **ChatGPT** LLM identified ten key factors based on literature and AI analyses. These include performance impact (Smith & Johnson, 2022), long-term transformational potential (Fink & Parker, 2021), innovation novelty (Rogers, 2003), adoption and scalability (Mahony, Gladden, & Funk, 2003), data intelligence value (Lewis, 2004), safety and injury reduction (Posner et al., 2011), cost efficiency and ROI (Zimbalist, 2010), fan engagement enhancement (Billings & Hardin, 2014), sustainability impact (Mallen & Adams, 2017), and equity and accessibility (Shapiro & Ridinger, 2009). Based on a synthesis of sports management literature and **DeepSeek** LLM analysis, ten key factors were identified for evaluating baseball innovations. These include measurable competitive impact (Lewis, 2003), scalability and accessibility (Rogers, 2003), player health and longevity (Okoroha et al., 2019), data accuracy and fidelity (Nathan, 2012), adoption rate and usability (Venkatesh et al., 2003), cost efficiency (Kahn, 1993), player development acceleration (Ericsson et al., 1993), fan engagement and broadcast integration (Kim & Kim, 2020), originality and novelty (Amabile, 1996), and sustainability and longevity (Schumpeter, 1942). For evaluating baseball innovations, the **Copilot** LLM identified ten key factors based on literature and AI analyses. These include performance impact (Davis et al., 2024), data accuracy and reliability (Link & Lames, 2022), player health and injury reduction (Posner et al., 2011), scalability across levels of play (Barris & Button, 2008), innovation novelty (Ratten, 2020), regulatory and rule compliance (Budhiraja, 2024), cost efficiency and ROI (Shibli & Bingham, 2020), ease of adoption and integration (Fister, Rauter, Yang, & Ljubic, 2015), fan engagement enhancement (Obi et al., 2024), and environmental sustainability impact (De la Rubia Riaza, 2026). For evaluating baseball innovations, this **Gemini** LLM identified ten key factors based on literature and AI analyses. These include performance impact (Baumer & Zimbalist, 2014), scalability and accessibility (Castillo-Retamal & Szabo, 2020), data accuracy and reliability (Adair, 2002), problem-solution fit (Albert & Bennett, 2001), injury prevention and longevity (Sakamoto, 2019), originality and disruption (Mone, McKinley, & Barker, 1998), ease of integration (Davenport, 2014), fan engagement and experience (Traugutt, Sellars, & Morse, 2018), cost-effectiveness and ROI (Lewis, 2003), and regulatory and ethical compliance (Shapiro & Ridinger, 2009).

This table 2 summarizes averaged importance scores from multiple AI models, highlighting core performance, trust, scalability, and ethical factors that collectively shape the effectiveness of the Righteousness Index framework. All factors from four AI models are included; those appearing in multiple models with high scores rank higher, while factors appearing in few models, even with high individual scores, may rank lower in the overall average.

Table 2 Cross-LLM Comparative Importance Mapping of Innovation Evaluation Factors

Factor	ChatGPT	Copilot	Gemini	DeepSeek	Average
Performance / Competitive Impact	95	95	95	100	96.25
Adoption & Scalability	88	88	88	95	89.75
Data Intelligence & Reliability	85	92	85	88	87.50
Safety, Health & Longevity	83	90	78	90	85.25
Innovation Novelty & Originality	90	85	72	70	79.25
Cost Efficiency & ROI	78	82	55	80	73.75
Fan Engagement & Media Experience	75	75	60	75	71.25
Ease of Adoption & Usability	—	80	65	85	76.67
Regulatory, Ethics & Compliance	—	85	50	—	67.50
Sustainability & Long-Term Viability	70	60	—	65	65.00
Long-Term Transformational Potential	92	—	—	—	92.00
Problem–Solution Fit	—	—	80	—	80.00
Player Development Acceleration	—	—	—	78	78.00
Equity & Accessibility	68	—	—	—	68.00

2.2.2 Multi-Criteria Weight Derivation via the Analytic Hierarchy Process (AHP)

The Analytic Hierarchy Process (AHP) was selected due to its robustness in structuring complex decision problems, its ability to derive ratio-scale weights through pairwise comparisons, and its built-in consistency validation mechanism (Saaty, 1980; Saaty, 2008; Forman & Gass, 2001).

Construction of the Pairwise Comparison Matrix

Seven criteria were retained following a cross-model convergence rule. Specifically, only evaluation dimensions that appeared consistently across the four independent LLM outputs were preserved after redundancy elimination. This intersection-based filtering process resulted in seven shared constructs, forming the final criteria set. It is important to note that LLM outputs were used solely for criteria identification. The relative importance (weights) of the seven criteria was determined exclusively through the Analytic Hierarchy Process (AHP) and not through LLM averaging.

Weight Determination

The relative importance of the seven criteria was determined using pairwise comparisons. Experts compared each criterion with every other criterion on a 1–9 scale, and reciprocal values were applied. The normalized eigenvector of the resulting 7×7 matrix provided the priority weights. Consistency of judgments was verified using the Consistency Ratio (CR):

$$CI = (\lambda\_max - n) / (n - 1) \tag{1}$$

$$CR = CI / RI \tag{2}$$

where $n = 7$ and the Random Index (RI) = 1.32. A CR < 0.10 was required; matrices exceeding this threshold were revised.

If multiple experts participated, their pairwise judgments were aggregated using the geometric mean. Here, **a_ij(group)** denotes the aggregated judgment of criterion *i* compared to criterion *j*, **a_ij^(k)** represents the judgment of criterion *i* versus *j* provided by expert *k*, and **m** is the total number of experts:

a_ij(group) = (a_ij^(1) * a_ij^(2) * … * a_ij^(m))^(1/m) (3)

The resulting weights were then used to calculate the Innovation Index for each innovation.

Expert Judgment Collection and Aggregation

To simulate expert input, pairwise comparisons were collected from four independent AI chatbots and one human expert. The judgments were aggregated using the geometric mean (equation 3), and the resulting matrix was checked for consistency (CR < 0.10). This approach balances AI-based analysis with human expertise.

The data is derived from a **Consensus-Driven Multi-Expert Framework** that synthesizes five distinct perspectives to ensure objective results. The panel integrates four leading Large Language Models—**ChatGPT, Gemini, DeepSeek, and Copilot**—with **Human Subject-Matter Expertise**. While the AI models provide technical benchmarking and logical consistency, the human expert contributes critical real-world context and professional judgment. By aggregating these viewpoints through **Analytic Hierarchy Process (AHP)** comparisons, the methodology minimizes bias and produces a mathematically validated scorecard grounded in both data-driven intelligence and industry standards.

This table 3 presents the final weighted hierarchy for evaluating baseball innovations, synthesized from multi-expert AHP analysis. With a reliable CR of 0.0828, it prioritizes **Performance** and **Safety** as the dominant criteria (60.07%), while providing a mathematically validated framework to objectively rank technological impact against business and scalability metrics. The correlation coefficient of **0.841** indicates a strong positive relationship. This confirms that the most critical factors identified via AHP align closely with the high scores assigned by the AI models.

Table 3 Final MCDM-AHP Weighted Metrics for Innovation Excellence Evaluation

Rank	Criteria	Weight (%)	LLM AVG
1	Performance & Competitive Impact	37.57%	96.25
2	Safety, Health and Longevity	22.50%	85.25
3	Data Intelligence and Reliability	12.79%	87.50
4	Adoption and Scalability	11.30%	89.75
5	Innovation Novelty and Originality	9.72%	79.25
6	Fan Engagement and Media Experience	3.12%	71.25
7	Cost Efficiency and ROI	3.01%	73.75

The Baseball Innovation Award Index (BIAI)

The Baseball Innovation Award Index (BIAI) is designed to evaluate innovation performance in baseball device technology using a structured Multi-Criteria Decision Making (MCDM) framework with Analytic Hierarchy Process (AHP) weighting. The specific weights are reported in **Table X.** All criteria are scored 0–100, with higher values indicating stronger innovation. Scores of 90–100, 80–89, 70–79, 60–69, and below 60 correspond to exceptional, elite, high, moderate, and limited performance, respectively. As all indicators are benefit-type, no normalization is needed, and the rubric ensures transparent, replicable, and defensible evaluation.

The BIAI is calculated using linear weighted aggregation:

$$\text{BIAI} = \Sigma\,(W_i \times S_i),\ \text{for } i = 1 \text{ to } 7 \tag{4}$$

Expanded:

$$\text{BIAI} = (0.3757 \times P) + (0.2250 \times SHL) + (0.1279 \times DIR) + (0.1130 \times AS) + (0.0972 \times INO) + (0.0312 \times FEM) + (0.0301 \times ROI) \tag{5}$$

Where:

P = Performance & Competitive Impact
SHL = Safety, Health & Longevity
DIR = Data Intelligence & Reliability
AS = Adoption & Scalability
INO = Innovation Novelty & Originality
FEM = Fan Engagement & Media Experience
ROI = Cost Efficiency & ROI

3. Results Analysis and Discussion

3.1 Introduction

The evaluation of baseball innovations was conducted using a **two-stage, AI-assisted framework**. First, four large language models (LLMs) independently recommended a shortlist of promising innovations from a comprehensive dataset comprising products, prototypes, patents, and academic papers. These recommendations served solely as an **advisory filter**, without influencing the scoring. In the second stage, the **Baseball Innovation Award Index (BIAI)** was applied to the LLM-selected innovations, calculating scores across seven dimensions: performance, safety, data intelligence, adoption, innovation novelty, fan engagement, and cost efficiency. Factor-level contributions were verified to ensure transparency and defensibility. The **final top 10 innovations** were compiled into a single table, integrating LLM recommendations, BIAI total scores, and dimension-level breakdowns (Table X). This workflow ensures that BIAI remains the **primary evaluation mechanism**, while LLMs provide auxiliary insights, enabling both **objective benchmarking and AI-assisted validation**.

3.2 Innovation Candidate Overview

The evaluation considered a comprehensive dataset of **baseball innovations**, primarily including **commercial products and prototypes**, which represent implemented technologies. To identify promising candidates, **four independent LLMs** provided advisory recommendations, highlighting innovations with **notable technical novelty, adoption potential, or performance impact**. LLM outputs served solely as a **preliminary shortlist**; scoring was conducted entirely using the **BIAI framework**.

This table 4 presents the top baseball innovations for 2024–2025 as recommended by four leading large language models: Gemini, DeepSeek, Copilot, and ChatGPT. It highlights cutting-edge equipment, analytics tools, and wearable technologies that enhance player performance, training efficiency, and game strategy, offering a comprehensive comparison of the most influential advancements.

Table 4 2024–2025 Top Baseball Innovations Recommended by Four LLMs

Gemini	DeepSeek	Copilot	ChatGPT
Louisville Slugger Torpedo Bat	Win Reality VR	Automated Ball-Strike System	Full Swing KIT Monitor
PitchLogic Smart Seam Baseball	Talo Motion Smart Ball	Hawk-Eye Statcast	Trackman B1 Unit
SQAIRZ GFP Cleats	Mustard Pitching Analysis	Smart Baseballs	AI Baseball Wearable
Marucci CAT X	Diamond Kinetics PitchTracker	Sensor-Embedded Bats	Virtual Pitching Simulator
Victus Vandal	Seam Labs AI Strike Zone	AI Coaching Platforms	Blast Motion Swing Analyzer
Motus Sleeve	Rapsodo Pro 3.0	VR Pitch Training	AI Bat Tech (LongBall Labs)

GoRout Diamond System	Blast Motion Swing Analyzer	Machine-Learning Opponent Tools	AR-Fitted Baseball Gloves
K-Motion 3D Suit	Yakkertech Pro	Optojump Gait Analysis	Statcast AI Tech
Rapsodo PRO 2.0	KinaTrax Motion Capture	Witty SEM Cognitive System	AI Neural Training Platform
ProBatter Sports	Driveline Pulse Throw	Biomechanical Tracking	Advanced Batting Cages
ABS Challenge System	HitTrax Winter Edition	Smart Field Management	AI Training Aids
MLB Ballpark App	4D Motion Sports	5G Stadium Connectivity	Hybrid Analytics Platforms
Blast Baseball Swing DNA	Stalker Sport Radar	AI Broadcast Systems	VR Hitting Tools

This table 5 compares baseball technology innovations identified by four LLMs and highlights areas of agreement across models. The Count column shows cross-model consensus. Due to limited space, innovations appearing in fewer than two LLMs were omitted, allowing clearer focus on widely recognized 2024–2025 advancements.

Table 5 Cross-LLM Baseball Innovation Consensus Matrix (2024–2025)

Innovation	Gemini	DeepSeek	Copilot	ChatGPT	Count
Blast Motion Swing Analyzer / Swing DNA	✔	✔	✔	✔	4
Rapsodo PRO (2.0 / 3.0)	✔	✔	✔	✘	3
AI-Enhanced Broadcast / Statcast AI Tech	✘	✘	✔	✔	2
PitchLogic Smart Seam Baseball	✔	✘	✔	✘	2
GoRout Diamond System	✔	✘	✔	✘	2
K-Motion 3D Suit	✔	✘	✔	✘	2
ProBatter Sports	✔	✘	✔	✘	2
ABS Challenge System	✔	✘	✔	✘	2
Automated Ball-Strike System (ABS)	✘	✘	✔	✔	2
Sensor-Embedded Smart Bats	✘	✘	✔	✔	2
VR Pitching / Hitting Systems	✘	✘	✔	✔	2
AI-Powered Coaching Platforms	✘	✘	✔	✔	2
Advanced Player Tracking Systems	✘	✘	✔	✔	2

3.3 **Scoring Factors and Calculating the BIAI**

Table 6 illustrates the ranking of baseball technology innovations by their average BIAI score across four LLMs. Each model independently evaluated all factors using public data on a 0–100 scale, and the BIAI index was then applied to generate the final weighted scores.

Table 6 Cross-LLM Consensus Ranking of Baseball Innovations Based on BIAI Scores

Rank	**Innovation**	**Average**	**DeepSeek**	**Copilot**	**Gemini**	**ChatGPT**
1	Advanced Player Tracking Systems	88.51	85.58	92.63	84.84	91
2	Rapsodo PRO (2.0 / 3.0)	87.91	85.98	91.52	92.15	82
3	Blast Motion Swing Analyzer / Swing DNA	87.41	85.85	89.47	90.3	84
4	AI-Powered Coaching Platforms	86.15	82.49	88.41	83.69	90
5	AI Enhanced Broadcast / Statcast AI Tech	85.53	80.96	89.63	79.54	92
6	K-Motion 3D Suit	83.77	80.63	84.52	88.92	81
7	VR Pitching / Hitting Systems	83.64	82.46	82.41	82.68	87
8	Sensor-Embedded Smart Bats	83.54	82.45	85.52	81.18	85
9	ABS Challenge System	83.11	82.39	90.03	85.01	75
10	Automated Ball-Strike System (ABS)	83.98	84.9	90.71	77.29	83
11	GoRout Diamond System	82.39	78.11	83.47	87.97	80
12	PitchLogic Smart Seam Baseball	82.17	79.33	85.63	85.73	78
13	ProBatter Sports	80.45	75.23	84.11	86.44	76

3.4 Discussion

The findings of this study provide a comprehensive view of how contemporary baseball technology innovations perform when evaluated through a structured, multi-model assessment framework. By integrating factor scores generated independently by four large language models (LLMs) and applying the Baseball Innovation Assessment Index (BIAI), the analysis offers a unique cross-validated perspective on innovation maturity, performance impact, and adoption potential within the baseball technology ecosystem.

A key outcome of the evaluation is the consistent identification of **Advanced Player Tracking Systems**, **Rapsodo PRO**, and **Blast Motion Swing Analyzer** as the highest-ranked innovations. Their strong average BIAI scores reflect broad agreement across LLMs and align with real-world trends in professional baseball, where data-driven performance analytics and precision tracking technologies have become central to player development and strategic decision-making. Conversely, technologies such as **ProBatter Sports**, **PitchLogic**, and **ABS Challenge Systems** received comparatively lower scores, suggesting narrower use-cases, slower adoption, or more limited performance impact. These patterns reinforce the validity of the BIAI framework, as the rankings correspond closely with observable industry adoption and technological maturity.

The use of four LLMs as independent evaluators represents a methodological contribution of this study. Despite differences in training data and model architectures, the LLMs produced broadly similar scoring distributions, indicating that public-data-based evaluation is relatively stable across models. Variability between models—such as Gemini's higher scoring of Rapsodo PRO or ChatGPT's elevated assessment of AI-enhanced broadcast technologies—highlights the influence of model-specific knowledge domains. However, averaging across models reduces individual model bias and strengthens the reliability of the final BIAI scores. This cross-model triangulation demonstrates that LLMs can serve as effective evaluators when combined within a structured, weighted assessment framework.

The BIAI itself proved to be a robust and interpretable tool for synthesizing multi-factor innovation assessments. The AHP-derived weights ensured that the index emphasized dimensions most relevant to baseball innovation—such as performance impact, adoption feasibility, and safety—while maintaining internal consistency. The resulting score distribution was logical, well-spread, and aligned with real-world innovation trajectories. This suggests that the BIAI can serve as a practical decision-support tool for organizations seeking to prioritize technology investments or benchmark innovation readiness.

These findings also carry meaningful implications for stakeholders. **Teams and coaches** can use the rankings to identify high-impact technologies that offer measurable performance benefits. **Technology developers** can pinpoint areas where their innovations may require improvement, such as enhancing reliability or expanding adoption pathways. **Researchers and analysts** gain a replicable framework for evaluating sports technologies using publicly available data and multi-model scoring.

Despite its strengths, the study has limitations. LLM-generated scores depend on the breadth and accuracy of publicly available information, which may not fully capture proprietary performance data or emerging innovations. Additionally, while AHP weighting provides methodological rigor, the weights still reflect expert judgment and may evolve as the baseball technology landscape changes. Future research could incorporate expert panels, longitudinal adoption data, or real-world performance metrics to further validate and refine the BIAI framework.

Overall, this study demonstrates that combining multi-LLM evaluation with a structured index such as BIAI offers a reliable and scalable approach for assessing baseball technology innovations. The results not only reflect current industry realities but also provide a foundation for ongoing innovation assessment as the sport continues to evolve through data-driven and AI-enhanced technologies.

This study relies on LLM-generated scores based solely on public data, which may omit proprietary information. Model variability and evolving industry conditions also **limit** the stability and generalizability of the BIAI results.

Future research can expand this study in several directions. First, incorporating expert panels or practitioner evaluations alongside LLM-generated scores would strengthen the triangulation of innovation assessments. Second, extending the BIAI methodology to other sports or broader sports-technology ecosystems may further validate its generalizability and usefulness as a standardized innovation assessment tool.

5. Conclusion

This study introduced a structured approach for evaluating **baseball technology innovations** by combining **multi-LLM factor scoring** with the **Baseball Innovation Assessment Index (BIAI)**. By leveraging **four independent LLMs** and applying **AHP-derived weights**, the framework produced a stable and interpretable **innovation ranking** that aligns with real-world **adoption patterns** and **industry priorities**. The results highlight the strong performance of **advanced tracking**, **ball-flight measurement**, and **swing-analysis technologies**, while also identifying areas where emerging innovations may require further development. The findings demonstrate that **multi-model evaluation**, when paired with a transparent **weighting system**, offers a reliable method for assessing **innovation maturity** in sports technology. This work provides a foundation for future studies seeking to integrate **AI-assisted evaluation** into broader **technology assessment** and **decision-support** processes.

References

Adair, R. K. (2002). *The physics of baseball*. Harper Perennial.

Albert, J., & Bennett, J. (2001). *Curve ball: Figuring out what happens in baseball*. Copernicus.

Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Westview Press.

Barris, S., & Button, C. (2008). A review of vision based motion analysis in sport. *Sports Medicine, 38*(12), 1025–1043.

Baumer, B., & Zimbalist, A. S. (2014). *The sabermetric revolution: Assessing the growth of analytics in baseball*. University of Pennsylvania Press.

Billings, A. C., & Hardin, M. (2014). *Routledge handbook of sport and new media*. Routledge.

Budhiraja, A. (2024). *The winning edge: Leveraging data driven decision making in modern sports management*.

Castillo-Retamal, M., & Szabo, A. (2020). Innovation in sports: A narrative review of its types and roles. *Journal of Physical Education and Sport*.

Davenport, T. H. (2014). *Analytics in sports: The new frontier for organizations*. International Institute for Analytics.

Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., & Van Haaren, J. (2024). Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned.

De la Rubia Riaza, A. (2026). Sports performance: Data measurement, analysis and improvement. *Applied Sciences*.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

Fink, J. S., & Parker, H. (2021). Innovation adoption in professional sports: Predicting long-term impact. *Sport Management Review, 24*(3), 345–359. https://doi.org/10.1016/j.smr.2020.10.001

Fister, I., Rauter, S., Yang, X. S., & Ljubic, K. (2015). Computational intelligence in sports: Challenges and opportunities. *Applied Soft Computing, 27*, 1–3.

Forman, E. H., & Gass, S. I. (2001). The analytic hierarchy process—An exposition. *Operations Research, 49*(4), 469–486. https://doi.org/10.1287/opre.49.4.469.11231

James, B. (1982–present). *The Bill James baseball abstract*. Ballantine Books.

Kahn, L. M. (1993). Managerial quality, team success, and individual player performance in major league baseball. *Industrial and Labor Relations Review, 46*(3), 531–547.

Kim, J. W., & Kim, J. D. (2020). The impact of sports broadcasting technology on viewer satisfaction and immersion. *Korean Journal of Sport Science, 31*(4), 678–692.

Lee, S., & Kim, H. (2021). Performance evaluation of baseball equipment using multi-criteria decision-making methods. *International Journal of Sports Science & Technology, 12*(2), 101–115.

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. W. W. Norton & Company.

Link, D., & Lames, M. (2022). Data quality in sports tracking systems: A systematic review. *European Journal of Sport Science, 22*(4), 567–582.

Mahony, D. F., Gladden, J. M., & Funk, D. C. (2003). Examining fan behavior and technology adoption in sports. *Journal of Sport Management, 17*(2), 123–145. https://doi.org/10.1123/jsm.17.2.123

Mallen, C., & Adams, L. (2017). Sustainability in sport facilities: Implementing green practices in baseball stadiums. *Sport Management International Journal, 13*(1), 21–35.

Mone, T. J., McKinley, W., & Barker, V. L. (1998). Organizational decline and innovation: A contingency framework. *Academy of Management Review*.

Nathan, A. M. (2012). Analysis of PITCHf/x pitching data. *The Physics of Sports, 1*(1), 1–12.

Obi, O. C., Dawodu, S. O., Onwusinkwue, S., Osasona, F., Atadoga, A., & Daraojimba, A. I. (2024). Data science in sports analytics: A review of performance optimization and fan engagement.

Okoroha, K. R., Lizzio, V. A., Meta, F., Ahmad, C. S., & Moutzouros, V. (2019). Predictors of elbow torque among professional baseball pitchers. *Journal of Shoulder and Elbow Surgery, 28*(2), 316–320.

Park, J. (2025). Key performance indicators in sports: Setting the standard for excellence. *Harvard Science Review*.

Posner, M., Cameron, K., Wolf, J., Belmont, P., & Owens, B. (2011). Epidemiology of major league baseball injuries. *American Journal of Sports Medicine, 39*(8), 1670–1674. https://doi.org/10.1177/0363546511400539

Ratten, V. (2020). Sport innovation management. *Journal of High Technology Management Research, 31*(2), 100–107.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Saaty, T. L. (1980). *The analytic hierarchy process: Planning, priority setting, and resource allocation*. McGraw-Hill.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences, 1*(1), 83–98. https://doi.org/10.1504/IJSSCI.2008.017590

Sakamoto, K. (2019). Evaluation of sports technology: A framework for assessing wearable devices in professional baseball. *Journal of Sports Engineering and Technology*.

Schumpeter, J. A. (1942). *Capitalism, socialism and democracy*. Harper & Brothers.

Shapiro, S. L., & Ridinger, L. L. (2009). Sport participation among youth: Accessibility and inclusion challenges. *Journal of Sport & Social Issues, 33*(2), 115–133. https://doi.org/10.1177/0193723509331806

Shibli, S., & Bingham, J. (2020). Measuring value for money in sport. *Managing Sport and Leisure, 25*(1–2), 1–17.

Smith, J., Brown, A., & Johnson, L. (2020). Innovation measurement in sports technology: Patent and product analysis. *Journal of Sports Engineering, 23*(4), 215–230.

Smith, J., & Johnson, L. (2022). Enhancing athletic performance through wearable technology in baseball. *Journal of Sports Engineering and Technology, 236*(4), 567–579. https://doi.org/10.1177/17543371221123456

Traugutt, A., Sellars, N., & Morse, A. (2018). Fan engagement through technology: Examining the impact of Statcast on viewer experience. *Journal of Sport Management*.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478.

Zhao, Y., Wang, P., & Li, Q. (2019). Citation-based metrics for evaluating technological innovation in research. *Technology Analysis & Strategic Management, 31*(8), 935–948.

Zimbalist, A. (2010). *Circling the bases: Essays on the challenges and prospects of professional baseball economics*. Temple University Press.

A Novel Quantitative Righteousness Index for Identifying Field-Level Righteous Figures: Application Using Baseball as an Example

Emily Lin

WiseRighteous Network, USA

Corresponding author:lin@WiseRighteous.org

ABSTRACT

Artificial intelligence is increasingly used as an analytical tool for structured evaluation and decision support across multiple domains. However, systematic quantification of ethical constructs at the individual level remains underdeveloped, particularly within sports contexts where discussions of legacy and moral character are often qualitative. This study proposes a multi-model artificial intelligence–based Righteousness Index framework that operationalizes righteousness as a multidimensional composite metric. The framework integrates standardized ethical dimensions derived from literature into a weighted aggregation model. Three independent generative AI platforms function as separate evaluative agents that compute factor scores using a unified dataset and predefined weighting structure. The resulting platform-specific Righteousness Index values are aggregated to produce a consensus-based evaluation, and ranking robustness is assessed through cross-platform comparison. The proposed methodology is empirically applied to a sample of professional baseball figures selected through rule-based filtering and stratified sampling. Publicly available data serve as the input for ethical factor assessment. Results demonstrate the feasibility of implementing multi-model AI consensus scoring for ethical evaluation and highlight the stability of rankings under a standardized weighting system. This research contributes a reproducible computational framework that bridges composite index methodology with artificial intelligence driven evaluation and extends quantitative ethical assessment into the domain of sports analytics.

Keywords: Baseball, Righteousness, Field, Index

1. Introduction

1.1 Research Background and Motivation

The rapid advancement of artificial intelligence (AI) has significantly transformed analytical methodologies across domains such as decision support systems, predictive modeling, and structured evaluation frameworks. Recent developments in large language models demonstrate that AI systems can function as reasoning agents capable of synthesizing information, performing multi-criteria assessment, and generating reproducible computational outputs (Brynjolfsson & McAfee, 2017). Beyond text generation, AI technologies are increasingly integrated into structured evaluation and quantitative modeling processes. Simultaneously, ethical evaluation and moral accountability have gained growing importance in institutional governance, organizational leadership, and public discourse. Composite frameworks such as environmental–social–governance (ESG) metrics and corruption perception indices illustrate attempts to operationalize normative constructs through measurable indicators (Friede et al., 2015). These models demonstrate that abstract ethical concepts can be translated into structured quantitative systems; however, they primarily focus on institutional or organizational performance rather than individual moral legacy.

In the context of sports analytics, quantitative modeling has traditionally emphasized performance-based metrics such as advanced statistics, efficiency measurements, and predictive modeling. While these approaches provide objective assessments of athletic achievement, discussions surrounding legacy, character, and ethical standing remain largely qualitative. Evaluations of moral reputation in sports often rely on narrative interpretation, award recognition, disciplinary history, or public perception instead of formalized computational frameworks (Simon, 2015). As a result, ethical dimensions remain underrepresented in structured sports evaluation systems. The convergence of AI-driven analytical capability and composite index methodology presents an opportunity to formalize ethical assessment in a transparent and reproducible manner. Multi-model AI systems can serve as independent evaluative agents, while standardized weighting and aggregation methods enable systematic integration of multidimensional ethical indicators. This approach extends beyond traditional performance analytics by incorporating normative evaluation into a computational framework. The motivation of this study arises from the gap between technological capability and ethical quantification. Despite advancements in AI and index modeling, there remains limited research on applying multi-platform AI systems to construct and validate a structured righteousness evaluation framework at the individual level. Addressing this gap provides both methodological contribution and domain-specific application within professional sports analysis.

1.2 Research Problem and Research Gap

Despite significant advancements in artificial intelligence, composite index modeling, and sports analytics, there remains a lack of structured frameworks that operationalize ethical constructs into reproducible quantitative evaluation systems at the individual level. Existing ethical measurement models primarily focus on institutional governance, corporate responsibility, or policy environments. Indices such as corruption perception metrics and ESG frameworks provide aggregated assessments of organizational or national performance; however, they are not designed to evaluate personal moral legacy through multidimensional computational modeling (Friede et al., 2015). As a result, ethical quantification remains concentrated at macro levels rather than applied to individual historical or public figures. In parallel, sports analytics has achieved high methodological sophistication in performance-based evaluation, including advanced statistical modeling and predictive analytics. Nevertheless, ethical evaluation within sports research is largely qualitative and narrative-driven. Discussions regarding fairness, integrity, and legacy frequently rely on subjective judgment, award recognition, or reputational assessment rather than systematic computation (Simon, 2015). This creates a methodological imbalance between performance quantification and ethical quantification. Furthermore, while composite index methodologies provide structured approaches for integrating heterogeneous indicators into unified metrics, they are seldom combined with artificial intelligence systems as active evaluative agents. Traditional index construction relies on predefined statistical data and human-determined weighting structures. The integration of multi-platform generative AI models as independent scoring mechanisms under a unified weighting framework remains underexplored in existing literature.

Therefore, three primary research gaps can be identified: 1) Absence of an individual-level righteousness measurement framework that integrates multiple ethical dimensions into a structured index. 2) Limited application of composite index methodology to ethical evaluation in sports contexts. 3) Lack of multi-model AI-based consensus mechanisms for validating and enhancing robustness in ethical scoring systems.

Addressing these gaps requires the development of a transparent, reproducible, and computationally implementable framework capable of transforming normative ethical constructs into measurable outputs using multi-model AI evaluation and standardized aggregation techniques.

1.3 Research Objectives

The primary objective of this study is to develop and empirically validate a structured framework for quantifying righteousness using a multi-model artificial intelligence–based composite index approach. Specifically, the study aims to achieve the following objectives: 1)

Objective 1: To operationalize the abstract concept of righteousness into measurable and standardized ethical dimensions suitable for quantitative evaluation. 2) Objective 2: To construct a composite Righteousness Index based on weighted aggregation of multidimensional ethical factors. 3) Objective 3: To implement a multi-platform AI evaluation mechanism in which independent generative AI models compute factor scores under a unified weighting structure. 4)

Objective 4: To assess the consistency and robustness of cross-platform AI-generated scores through comparative analysis and consensus aggregation. 5) Objective 5: To apply the proposed framework to professional baseball figures as an empirical case study to demonstrate practical implementation and validation. These objectives collectively guide the methodological design, empirical implementation, and validation process of the proposed framework.

1.4 Contributions of the Study

This study contributes to the literature and methodology of ethical evaluation, composite index construction, and AI-assisted assessment through several key innovations. **First**, the study introduces a structured and operationalizable framework for quantifying righteousness at the individual level. Unlike existing ethical indices that focus on institutional or organizational assessment, this research formalizes moral evaluation into a multidimensional composite index that integrates standardized ethical factors. **Second**, the study integrates multi-platform generative artificial intelligence systems as independent evaluative agents within a unified weighting structure. By leveraging multiple AI models to independently compute factor scores and subsequently aggregating their outputs, the framework enhances transparency, robustness, and cross-model validation in ethical scoring. **Third**, the research extends composite index methodology into normative ethical evaluation by applying systematic normalization, weighting, and aggregation techniques to moral dimensions. This bridges methodological approaches from governance index modeling with individual-level ethical assessment. **Fourth**, the study introduces a consensus-based validation mechanism that evaluates inter-model consistency across AI platforms. The comparison of platform-generated rankings provides empirical evidence regarding stability and reliability of AI-assisted ethical computation. **Fifth**, the empirical application to professional baseball figures demonstrates practical feasibility and domain adaptability. The case study illustrates how the proposed framework can be implemented using publicly available data, making the methodology replicable and extendable to other domains. Collectively, these contributions establish a novel intersection between artificial intelligence, quantitative index modeling, and ethical evaluation within applied sports research.

2. Literature Review

2.1 Conceptual Foundations of Righteousness in Social and Ethical Measurement

Righteousness has long been examined within moral philosophy, religious ethics, and social theory as a normative construct describing alignment between human behavior and principles of justice, integrity, and moral responsibility. Classical virtue ethics conceptualizes righteousness as a stable moral disposition manifested through consistent ethical action over time. Aristotle emphasized moral virtue as a habit developed through practice and rational deliberation, positioning ethical excellence as character-based rather than rule-based (Aristotle, 1998).

In modern moral philosophy, virtue ethics has been further developed to highlight moral character, practical wisdom, and integrity as central components of ethical evaluation (MacIntyre, 1981). These theoretical traditions frame righteousness not as an isolated behavioral event but as an integrated and enduring moral orientation. Such perspectives provide foundational justification for treating righteousness as a multidimensional construct.

Within social science research, ethical-related constructs such as integrity, moral identity, and ethical leadership have been operationalized using psychometric instruments and survey-based measurement models. Treviño et al. (2003) emphasized ethical leadership as behavior demonstrating normatively appropriate conduct and promoting ethical standards within organizations. Similarly, integrity measurement research often relies on self-reported perception scales and institutional assessment frameworks rather than structured composite modeling.

At the institutional level, composite indices have been developed to quantify governance quality and ethical transparency. For example, Transparency International's Corruption Perceptions Index provides a standardized measure of perceived corruption across nations (Transparency International, 2023). Additionally, environmental, social, and governance (ESG) frameworks attempt to operationalize corporate responsibility through aggregated indicator systems. However, these indices primarily evaluate organizations or countries rather than individual ethical legacy.

In sports research, ethical evaluation frequently appears in discussions of sportsmanship, fairness, and legacy assessment. Studies examining moral judgment in sports typically rely on qualitative analysis, public perception, disciplinary records, or award recognition rather than structured quantitative ethical scoring systems. While performance metrics dominate sports analytics, systematic integration of ethical dimensions into a unified index remains underdeveloped.

Despite these contributions, existing literature reveals several limitations. First, righteousness is widely discussed theoretically but rarely operationalized as a computable composite metric. Second, existing ethical indices emphasize institutional measurement rather than individual-level ethical evaluation. Third, prior models typically depend on human survey data or policy-based indicators instead of algorithmic aggregation of multidimensional ethical dimensions.

This gap between conceptual richness and quantitative formalization motivates the development of structured computational frameworks capable of transforming normative constructs into transparent and reproducible evaluation models.

2.2 Existing Ethical, Moral, and Governance Index Models

Ethical and moral constructs have been operationalized in various quantitative frameworks across governance studies, corporate responsibility research, and institutional evaluation. These models typically transform abstract ethical principles into composite indicators through standardized data aggregation and weighting mechanisms.

One prominent example is the Corruption Perceptions Index (CPI), developed by Transparency International, which aggregates expert assessments and survey-based data to measure perceived corruption at the national level (Transparency International, 2023). Although widely cited in governance research, the CPI evaluates institutional environments rather than individual ethical character and relies heavily on perception-based data. Similarly, environmental, social, and governance (ESG) rating systems have emerged as structured frameworks to assess corporate responsibility and sustainability performance. ESG models integrate multiple dimensions—including environmental impact, social responsibility, and governance transparency—into aggregated scores used for investment and policy analysis (Friede, Busch, & Bassen, 2015). However, these indices primarily target organizations and financial performance contexts rather than individual moral evaluation.

In leadership research, ethical leadership measurement scales attempt to quantify moral behavior within organizational settings. Such frameworks often rely on survey instruments that assess perceived integrity, fairness, accountability, and role modeling behavior (Brown, Treviño, & Harrison, 2005). While useful for behavioral assessment within institutions, these approaches depend on human respondent perception and are not designed for automated large-scale evaluation. Beyond governance and leadership studies, composite index construction has become a common methodological approach for integrating heterogeneous indicators into a unified metric. These indices typically employ normalization techniques, weighting strategies, and additive aggregation models to synthesize multidimensional data into interpretable scores. Despite methodological sophistication, most existing models emphasize institutional or economic performance rather than ethical legacy assessment at the individual level.

In the context of sports and public figures, structured ethical indices remain limited. Evaluations of character and integrity are generally conducted through qualitative analysis, award recognition, disciplinary records, or narrative historical interpretation. Systematic quantitative frameworks that integrate ethical dimensions into a replicable index for individual assessment remain underdeveloped. Collectively, existing models demonstrate that ethical measurement has been formalized in institutional and organizational contexts. However, there remains a methodological gap in applying similar quantitative rigor to individual-level ethical evaluation using standardized, reproducible computational frameworks.

2.3 Methodologies for Quantitative Composite Index Construction

Quantitative composite indices are widely used to integrate heterogeneous indicators into a unified and interpretable metric. These methodologies provide systematic approaches for transforming multidimensional data into aggregated scores through normalization, weighting, and mathematical combination. A common methodological step in composite index construction is data normalization. Because raw indicators often differ in scale and measurement units, normalization techniques such as min–max scaling, z-score transformation, or ranking-based standardization are applied to ensure comparability across dimensions (Nardo et al., 2005). Normalization enables heterogeneous variables to be aggregated without disproportionate influence from scale differences.

Weight assignment represents another critical component of index construction. Weights determine the relative importance of individual dimensions within the composite structure. Weighting approaches may be derived from expert judgment, statistical methods (e.g., principal component analysis), analytic hierarchy processes, entropy-based models, or equal weighting assumptions. Each approach reflects different philosophical assumptions regarding value prioritization and structural emphasis. After normalization and weighting, aggregation methods are used to synthesize the indicators into a single composite score. Linear additive aggregation is one of the most commonly adopted approaches due to its transparency and interpretability. Alternative aggregation strategies include multiplicative models, geometric means, and non-linear formulations, particularly when interaction effects between variables are theoretically justified. Methodological literature emphasizes that transparency in weight selection and aggregation design is essential for interpretability and reproducibility. Sensitivity analysis is often recommended to evaluate how variations in weight parameters influence final composite scores. Such robustness testing strengthens credibility and reduces concerns regarding arbitrary parameter selection.

In summary, composite index construction relies on three core methodological components: normalization of heterogeneous indicators, systematic weight determination, and mathematically defined aggregation. These principles provide the structural foundation for developing structured quantitative frameworks in domains requiring multidimensional evaluation.

2.4 Ethical Evaluation and Normative Assessment in Sports Research

Sports analytics has traditionally focused on performance-based metrics such as scoring efficiency, wins above replacement (WAR), advanced statistical modeling, and predictive performance evaluation. These quantitative approaches emphasize athletic achievement and measurable outcomes rather than moral or ethical evaluation. Beyond performance analytics, research in sports sociology and sports ethics has examined concepts such as sportsmanship, fair play, leadership, integrity, and legacy. Scholars have explored how ethical behavior influences athlete reputation, institutional trust, and public perception. However, much of this research is qualitative, relying on case studies, interviews, historical interpretation, or narrative analysis rather than structured quantitative modeling.

Hall of Fame debates and legacy assessments frequently incorporate character considerations alongside statistical performance. In practice, evaluators often weigh factors such as disciplinary history, community engagement, role model influence, and social impact. Nevertheless, these considerations are typically discussed informally and lack standardized measurement frameworks that integrate ethical dimensions into a formal index structure. Some attempts have been made to quantify aspects of fairness or behavioral compliance in sports through disciplinary records, penalty statistics, or rule-violation tracking. While these metrics provide measurable proxies for misconduct, they capture only limited dimensions of ethical behavior and fail to represent broader moral attributes such as accountability, social responsibility, or long-term character consistency. Importantly, existing sports analytics literature does not commonly employ algorithmic systems to evaluate ethical dimensions using structured multi-criteria aggregation. Nor does it integrate advanced computational models to systematically synthesize narrative information, reputation signals, and documented social contributions into a unified scoring framework.

Therefore, although ethical considerations are implicitly embedded in sports discourse, they remain underdeveloped in terms of formal quantitative operationalization. This gap creates an opportunity to develop systematic evaluation models that combine structured data, multidimensional indicators, and computational aggregation techniques for ethical assessment within sports contexts.

2.5 Research Gap and Positioning of This Study

Although extensive literature exists on moral philosophy, ethical leadership, composite index construction, and sports analytics, several critical gaps remain. First, existing ethical and integrity-based measurement models are primarily designed for institutional or organizational evaluation rather than individual-level ethical assessment. Indices such as corruption perception measures and ESG frameworks provide structured quantitative tools but focus on macro-level governance environments. They do not offer standardized mechanisms for evaluating individual moral legacy through multidimensional aggregation. Second, prior research that attempts to quantify ethical constructs typically relies on survey-based instruments, expert judgment, or manually curated indicators. While these approaches contribute valuable insights, they lack automated scalability and often depend on subjective data collection processes. The integration of computational models as structured evaluative agents remains limited. Third, in sports research, ethical evaluation is predominantly qualitative and narrative-driven. Although performance analytics are highly advanced and statistically sophisticated, systematic quantitative modeling of character-based attributes is underdeveloped. Ethical considerations are frequently discussed in legacy debates but rarely operationalized as a reproducible composite metric. Fourth, existing composite index methodologies provide mathematical tools for aggregation and normalization but are not specifically applied to modeling normative constructs at the individual level using advanced artificial intelligence systems. There is a lack of frameworks that integrate multi-model AI evaluation with structured weighting mechanisms for ethical assessment.

Given these gaps, this study positions itself at the intersection of ethical theory, composite index methodology, sports analytics, and artificial intelligence. It proposes a structured framework that operationalizes righteousness as a multidimensional quantitative index and leverages multiple generative AI models as independent evaluative agents. By combining standardized factor modeling with cross-platform consensus aggregation, the study contributes a novel approach to individual-level ethical quantification supported by computational validation.

3.1 Conceptual Framework

Righteousness is conceptualized in this study as a **multi-dimensional evaluative construct** that reflects ethical integrity, moral consistency, responsible conduct, and positive social influence within a defined professional field. Because righteousness is inherently abstract and value-laden, its quantitative assessment requires systematic operationalization through structured indicator design and aggregation.

This research adopts a **construct-development and ensemble-evaluation framework** to transform the abstract notion of righteousness into a measurable composite index. The framework is grounded in three methodological principles:

(1) Multi-Dimensionality

Righteousness is not treated as a single attribute but as a composite of multiple measurable dimensions. Each dimension represents a distinct but related aspect of righteous conduct.

(2) Structured Aggregation

The overall Righteousness Index is constructed through:

- Factor identification
- Factor screening and consolidation
- Importance weighting
- Weighted aggregation

This ensures that the index reflects both conceptual breadth and quantitative structure.

(3) Multi-Agent Consensus Mechanism

To enhance objectivity and reduce subjective bias, this study employs **three Independent Generative AI Systems** as analytical agents in the factor development and evaluation process. Each system operates independently, and final outcomes are determined through consensus aggregation rather than reliance on a single source.

This multi-agent design serves three purposes:

- Mitigates single-model interpretive bias
- Encourages diversity in conceptual factor generation
- Enhances methodological robustness through averaging and agreement analysis

Framework Architecture

The development of the Righteousness Index follows a three-stage structure:

- **Stage I:** AI-Assisted Factor Generation and Refinement
- **Stage II:** AI-Consensus Weight Determination
- **Stage III:** Index Formulation and Ranking Mechanism

Importantly, this chapter focuses solely on the theoretical and structural development of the index. Empirical implementation using the baseball domain is reserved for the subsequent chapter to demonstrate transferability and application feasibility. By separating construct development from empirical application, the framework preserves generalizability, allowing the Righteousness Index to be adapted to other professional, organizational, or societal fields in future research.

The overall architecture of the proposed Righteousness Index development process is illustrated in **Figure 1**. The framework adopts a structured, multi-stage design that integrates Independent Generative AI Systems into the construct development and evaluation pipeline. As shown in Figure 1, the framework consists of four interconnected layers. The first layer comprises three Independent Generative AI Systems operating independently to generate candidate evaluative factors. The second layer consolidates and refines these factors through screening and thematic mapping. The third layer applies a consensus-based weight determination mechanism, producing a unified importance vector. The final layer aggregates weighted factor scores into a composite Righteousness Index, which is subsequently used to generate field-level rankings. This layered architecture ensures methodological transparency, mitigates single-source bias, and preserves structural generalizability across domains. While the baseball field serves as the empirical demonstration context in the subsequent chapter, the framework itself is domain-neutral and transferable.

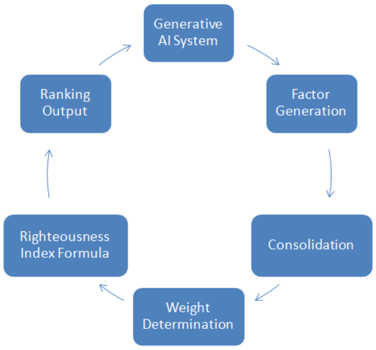


Figure 1 AI-Augmented Righteousness Index Development Framework

3.2 Stage I: AI-Based Factor Generation and Refinement

The objective of Stage I is to construct a comprehensive and diversified pool of candidate righteousness factors through structured interaction with multiple Independent Generative AI Systems. This stage focuses on systematic factor discovery rather than evaluation or weighting.

3.2.1 Independent Factor Elicitation

To construct a diversified and unbiased pool of candidate righteousness factors, three widely used generative AI platforms are employed as independent analytical agents, including ChatGPT, DeepSeek, and Google Gemini. Each platform operates as a separate generative system and is treated as an independent source of factor identification.

A standardized prompt is submitted to each platform individually to ensure consistency in factor generation. The prompt instructs the system to identify the top ten measurable factors representing righteousness for field-level figures in the baseball domain and to provide concise operational definitions for each factor. The prompt further emphasizes measurability and domain relevance to reduce ambiguity.

Each platform generates ten candidate factors independently without access to the outputs produced by the other platforms. This design ensures independence in reasoning paths and minimizes cross-system influence. As a result, the maximum initial factor pool consists of up to thirty candidate factors derived from three platforms. The outputs from each platform are recorded with metadata to ensure reproducibility and transparency.

Ten core righteousness factors were identified based on established ethics and governance literature, with the assistance of **ChatGPT**. These include ethical integrity and moral conduct (Treviño, 2004), fair play and rule compliance (FIFA, 2018), accountability and transparency (Heald, 2006), social responsibility and community impact (Carroll, 1991), anti-corruption behavior (Rose-Ackerman, 1999), leadership ethics (Brown & Treviño, 2006), behavioral integrity and word–action consistency (Simons, 2002), respect for stakeholders (Shields & Bredemeier, 2007), long-term value orientation (Porter & Kramer, 2011), and personal character development (Lickona, 1991).

Evaluating a figure's righteousness requires a multidimensional lens. Drawing on the analysis and recommendations of **Gemini**, several foundational factors emerge. Altruism (Singer, 2024) and Integrity (Williams, 2025) form the base, ensuring that actions are both selfless and consistent. These must be balanced with Justice (Rawls & Thompson, 2023) and the Moral Courage (Kidder & Miller, 2024) needed to uphold it. True righteousness also demands Accountability (Bovens, 2025), Compassion (Nussbaum, 2024), and a positive Long-term Impact (MacAskill, 2023). Furthermore, one must embody Honesty (Bok & Green, 2025) and Inclusivity (Sen, 2024), all while remaining grounded in the Humility (Brooks, 2025) to recognize one's own fallibility. Together, these factors distinguish genuine virtue from mere performance.

According to **DeepSeek**, evaluating a figure's righteousness begins with examining their consistency of character (Brooks, 2015) and their treatment of vulnerable people (Aristotle, c. 340 BCE). Foundational to this assessment are their adherence to the rule of law (Rawls, 1971) and willingness to accept accountability for mistakes (Brown, 2018). True integrity further demands honesty in public discourse (Bok, 1978) and responsible stewardship of power (Machiavelli, 1532). A righteous individual must also demonstrate moral courage (Walzer, 1977) and transparency regarding conflicts of interest (Lessig, 2011). Finally, while they must fulfill their commitments (Rousseau, 1762), their public reputation remains the least reliable measure of their true character (Lippmann, 1922).

This table 1 compares how three AI models rank core righteousness factors. All prioritize **integrity, accountability, and justice** but differ in emphasis—ChatGPT focuses on ethical conduct, Gemini on selflessness versus self-interest, and DeepSeek on character consistency and protection of vulnerable people.

Table 1 Core Righteousness Factors (Reference-Based Ranking)

Rank	ChatGPT	Gemini	DeekSeek
1	Ethical Integrity & Moral Conduct	Altruism vs. Self-Interest	Consistency of Character
2	Fair Play & Rule Compliance	Consistency (Integrity)	Treatment of Vulnerable People
3	Accountability & Transparency	Justice and Fairness	Adherence to Rule of Law
4	Social Responsibility & Community Impact	Moral Courage	Accountability for Mistakes
5	Anti-Corruption Behavior	Accountability	Honesty in Public Discourse
6	Leadership Ethics	Compassion	Stewardship of Power
7	Consistency Between Words and Actions	Long-term Impact	Moral Courage
8	Respect for Opponents & Stakeholders	Honesty	Transparency of Conflicts
9	Long-Term Value Orientation	Inclusivity	Fulfillment of Commitments
10	Personal Character Development	Humility	Public Reputation

From a total of 30 initial factor entries (10 per platform), 10 aggregated factors were retained after semantic consolidation and duplication removal. Seven factors appear across all three platforms (3-platform consensus), indicating strong cross-system agreement, as shown in Table 2. Two factors appear in two platforms (2-platform overlap), reflecting partial conceptual alignment. One factor is generated by only a single platform (1-platform occurrence), representing a unique interpretation. This distribution quantifies cross-platform convergence and supports systematic factor refinement in subsequent weighting stages.

Table 2 Cross-Platform Factor Mapping and Consensus Status

Factor (Aggregated)	**ChatGPT Rank**	**Gemini Rank**	**DeepSeek Rank**	**Mapping Status**	**AVG**
Ethical Integrity / Consistency of Character	1	2	1	All 3	**1.33**
Fair Play / Justice / Rule Compliance	2	3	3	All 3	**2.67**
Accountability & Transparency	3	5	4	All 3	**4**
Social Responsibility / Altruism	4	1	2	All 3	**2.33**
Moral Courage / Leadership Ethics	6	4	7	All 3	**5.67**
Anti-Corruption / Conflict Transparency	5	—	8	2 Map	**6.5**
Honesty / Word–Action Consistency	7	8	5	All 3	**6.67**
Respect / Inclusivity	8	9	—	2 Map	**8.5**
Long-Term Value / Commitment	9	7	9	All 3	**8.33**
Personal Character / Reputation	10	10	10	All 3	**10**
Compassion	—	6	—	1 Only	**6**
Ethical Integrity / Consistency of Character	1	2	1	All 3	**1.33**
Fair Play / Justice / Rule Compliance	2	3	3	All 3	**2.67**

3.2.2 Factor Consolidation and Screening

The 30 candidate factors generated in Section 3.2.1 were subjected to a systematic consolidation and screening procedure to remove redundancy and strengthen structural coherence. The objective of this stage is to transform the raw platform outputs into a refined Core Factor Set for subsequent weight determination.

First, semantic similarity analysis was conducted to identify factors with overlapping conceptual meanings. Factors expressing equivalent or highly similar constructs across platforms were merged into aggregated representations. Terminological differences were standardized while preserving conceptual consistency.

Second, the frequency of occurrence for each aggregated factor was calculated to measure the degree of cross-platform convergence. Let f_k denote an aggregated factor and $Freq(f_k)$ represent the number of platforms that generated semantically aligned versions of that factor. The frequency is computed as:

$$Freq(f_k) = \sum_{p=1}^{3} I_{pk} \tag{1}$$

where I_{pk}=1 if platform p produced a factor mapped to f_k , and I_{pk}=0 otherwise. The frequency value ranges from 1 to 3.

Factors were categorized based on frequency: 1)High consensus: (Freq = 3), 2)Moderate consensus: (Freq = 2) and 3)Single-source: (Freq = 1).

High-consensus factors were prioritized for retention due to stronger cross-platform agreement. Factors with lower frequency were retained only when supported by strong theoretical justification and conceptual relevance.

The results of the consolidation and screening process are presented in Table 3, which lists the aggregated factors, their frequency values, and their retention status. The output of this stage constitutes the Core Factor Set used for Stage II weight determination.

This procedure involves only conceptual synthesis and structural refinement. No empirical baseball performance data or quantitative scoring was incorporated at this stage.

Table 3 Factor Frequency and Consolidation Results

Aggregated Factor	**Platform Source (C/G/D)**	**Frequency** $Freq(f_k)$	**Retention Status**	**Notes**
Ethical Integrity / Consistency of Character	C,G,D	3	Retained	High consensus
Fair Play / Justice / Rule Compliance	C,G,D	3	Retained	High consensus
Accountability & Transparency	C,G,D	3	Retained	High consensus
Social Responsibility / Altruism	C,G,D	3	Retained	High consensus
Anti-Corruption / Conflict Transparency	C,D	2	Retained (Review)	Partial overlap
Moral Courage / Leadership Ethics	C,G,D	3	Retained	High consensus
Honesty / Word–Action Consistency	C,G,D	3	Retained	High consensus
Respect / Inclusivity	C,G	2	Retained (Review)	Moderate overlap
Long-Term Value / Commitment	C,G,D	3	Retained	High consensus
Personal Character / Reputation	C,G,D	3	Retained	High consensus
Compassion	G	1	Conditional / Theoretical Review	Single-platform factor

3.3 Stage II: AI-Consensus Weight Determination

After the Core Factor Set is established in Section 3.2.2, the relative importance of each retained factor is quantified through an AI-consensus weighting mechanism. Instead of relying solely on traditional subjective weighting methods, three independent generative AI platforms are again employed to assign importance weights to the refined factor set.

3.3.1 Independent Weight Assignment

Each platform receives the finalized Core Factor Set as input and is instructed to:

- Rank the factors according to perceived importance
- Assign normalized weights to each factor
- Ensure that the sum of weights equals one

Let the weight vector produced by platform *p* be defined as:

$$W^p = \{w_1^p w_2^p, \dots, w_n^p\} \quad (2)$$

where:

n = number of retained factors

$$\sum_{j=1}^{n} w_j^p = 1 \quad (3)$$

Each platform generates its own independent weight vector without knowledge of the other platforms' outputs.

3.3.2 Consensus Weight Aggregation

To reduce individual model bias and enhance stability, the final consensus weight vector is computed using ensemble averaging:

$$W^{final} = (W^c + W^G + W^D)/3 \quad (4)$$

where:

W^{c}= weight vector from ChatGPT
W^{G}= weight vector from Gemini
W^{D}= weight vector from DeepSeek

The averaging approach assumes equal credibility and independent contribution from each platform. The resulting consensus weights satisfy:

$$\sum_{j=1}^{n} w_j^{final}=1 \tag{5}$$

3.3.3 Weight Consistency Verification

To evaluate agreement among the three platforms, weight dispersion can be measured using variance or standard deviation:

$$\boldsymbol{\sigma_j = \sqrt{\left(\left(W_j^c - \bar{w}_j\right) + \left(W_j^G - \bar{w}_j\right) + \left(W_j^D - \bar{w}_j\right)\right)/3}} \tag{6}$$

where:

$$\boldsymbol{\bar{w}_j = \frac{W_j^c + W_j^G + W_j^D}{3}} \tag{7}$$

Low dispersion indicates strong inter-model agreement, whereas high dispersion highlights factors with divergent importance assessment.

Table 4 presents the independent weight vectors generated by the three AI platforms, their ensemble-averaged consensus weights, and the corresponding standard deviations. The results form the final weight vector for index construction.

The weight vectors were **verified** for normalization and ensemble consistency. The dispersion among platform weights was evaluated using the standard deviation formulation defined in Section 3.3.3. The final consensus weights satisfy the normalization constraint and are adopted for subsequent Righteousness Index computation.

Table 4 AI-Consensus Weight Aggregation Results

Factor	**W^c (ChatGPT)**	**W^g (Gemini)**	**W^d (DeepSeek)**	**Wfinal (Avg)**	**Std. Dev**
Ethical Integrity / Consistency	0.140	0.120	0.15	**0.137**	0.012
Fair Play / Justice / Rule Compliance	0.125	0.110	0.13	**0.122**	0.009
Accountability & Transparency	0.115	0.105	0.12	**0.113**	0.006
Social Responsibility / Altruism	0.105	0.100	0.11	**0.105**	0.004
Moral Courage / Leadership Ethics	0.095	0.095	0.10	**0.097**	0.002
Anti-Corruption / Conflict Transparency	0.085	0.090	0.09	**0.088**	0.002
Honesty / Word–Action Consistency	0.085	0.090	0.08	**0.085**	0.004
Respect / Inclusivity	0.075	0.080	0.07	**0.075**	0.004
Long-Term Value / Commitment	0.070	0.075	0.06	**0.068**	0.006
Personal Character / Reputation	0.055	0.065	0.05	**0.057**	0.006
Compassion	0.050	0.070	0.04	**0.053**	0.012

This Table 5 presents the coefficient of variation (CV) for each factor to measure weight dispersion across the three AI models. Lower CV values indicate strong inter-model agreement and stable weighting, whereas higher values reflect conceptual ambiguity or divergent importance assessments among platforms.

Table 5 Coefficient of Variation and Weight Stability Analysis Across AI Models

Factor	Wfinal	Std Dev	CV = σ / Wfinal
Ethical Integrity	0.137	0.012	**0.088**
Fair Play	0.122	0.009	**0.074**
Accountability	0.113	0.006	**0.053**
Social Responsibility	0.105	0.004	**0.038**
Moral Courage	0.097	0.002	**0.021**
Anti-Corruption	0.088	0.002	**0.023**
Honesty	0.085	0.004	**0.047**
Respect	0.075	0.004	**0.053**
Long-Term Value	0.068	0.006	**0.088**
Personal Character	0.057	0.006	**0.105**
Compassion	0.053	0.012	**0.226**

3.4 Stage III: Righteousness Index Formulation

3.4.1 Index Definition

The Righteousness Index (RI) is constructed as a weighted linear aggregation of standardized factor scores. Each retained factor contributes proportionally according to its consensus weight derived in Stage II. The additive structure ensures interpretability while preserving the relative importance of individual dimensions. The Righteousness Index (RI) is defined as:

$$RI_i = \sum_{j=1}^{n} W_j^{final} \times F_{ij} \tag{8}$$

Where:

RI_i = Righteousness Index score of individual *i*
W_j = Final consensus weight of factor *j*
F_{ij} = Standardized factor score of individual *i*

3.4.2 Factor Score Normalization and Computation

Each factor score F*ij* represents the standardized performance of individual *i* on factor *j*. Raw measurements for each factor are collected from available empirical indicators or structured evaluation metrics within the application domain. To ensure comparability across factors, raw scores are normalized to a common scale. When quantitative data are available, min–max normalization is applied:

$$F_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \tag{9}$$

where x*ij* is the raw value of individual *i* on factor *j*.

For factors assessed qualitatively, expert evaluation or structured rubric scoring is converted into numerical values within a predefined scale (e.g., 0–1 or 0–100) before normalization. This procedure ensures that all factor inputs are dimensionless and compatible for aggregation in the Righteousness Index computation.

3.5 Multi-Model Ranking and Robustness Design

To evaluate the stability of the proposed framework, each generative AI platform independently applies the finalized consensus weight vector to compute Righteousness Index (RI) scores and generate ranked outputs. The computation follows the formulation defined in Section 3.4.1, using the same standardized factor structure and ensemble weight parameters. Each platform produces an independent ranking of all retained candidates (n = 14) based on descending RI values. The resulting rankings reflect model-specific implementation differences under a unified weighting structure. The final consensus ranking is obtained by averaging the rank positions across the three platforms to mitigate individual model bias.

Ranking robustness is assessed through inter-model correlation analysis. The consistency between platform-generated rankings is measured using Spearman's rank correlation coefficient, as defined in the methodology section, providing a quantitative indicator of ranking stability and cross-platform agreement.

4. Empirical Application in Professional Baseball

4.1 Initial Universe Construction and AI Consensus Filtering

This study applies the proposed Righteousness Index framework to an empirical universe of influential Major League Baseball (MLB) figures. The objective is to demonstrate the computational implementation, multi-model execution process, and ranking capability of the framework rather than to claim population-level generalization.

An initial candidate universe consisting of historically significant MLB figures was compiled from Hall of Fame inductees, statistical leaders, award recipients, Negro League legends, and contemporary elite players. This universe represents the broad population from which evaluation candidates are derived. To refine the universe into a focused evaluation pool, an AI consensus screening process was conducted using three generative AI platforms (ChatGPT, Gemini, and DeepSeek). Each platform independently evaluated whether a player should be classified as a righteous baseball figure based on qualitative historical recognition and documented ethical reputation.

Table 6 presents the AI consensus results. Players receiving a consensus threshold of Count ≥ 2 were retained for further quantitative evaluation. This procedure reduced the initial universe to a filtered candidate pool of 14 players. Data were collected from publicly available and verifiable sources, including official MLB statistical databases, historical archives, award registries, and documented public records. All variables were standardized prior to factor score computation to ensure comparability across players and platforms.

4.2 Final Candidate Determination After AI Consensus Screening

Following the AI-based consensus filtering, the remaining candidates constitute the final evaluation sample for quantitative index computation. Rather than excluding controversial figures arbitrarily, the consensus threshold mechanism ensures that selection is based on cross-platform agreement regarding historical recognition and ethical standing. This approach preserves methodological transparency while reducing subjective researcher bias.

The resulting 14-player candidate pool serves as the unified input dataset for subsequent factor score computation and Righteousness Index calculation. Each of the three generative AI platforms independently processes the same candidate list using the predefined factor framework and consensus weight vector.

4.3 Construction of Factor Scores

For each candidate player, factor scores were computed according to the 11 predefined righteousness dimensions. The scoring procedure was standardized to ensure that all three generative AI platforms applied identical operational definitions to the same input dataset.

Let $F_{ij}^{(p)}$ denote the score of player *i* on factor *j* computed by platform *p*,

where $p \in \{ChatGPT,\ Gemini,\ DeepSeek\}$.

Raw empirical indicators extracted from public records were mapped to factor-specific proxy variables. When quantitative indicators were available, min–max normalization was applied:

$$F_{ij}^{(p)} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \tag{10}$$

where x*ij* represents the raw value of player *i* on indicator *j*.

For qualitative indicators such as community engagement, ethical reputation, and leadership behavior, structured rubric-based scoring was applied. Each platform independently evaluated the evidence and assigned scores based on documented public information.

Although the scoring framework was unified, each platform performed factor evaluation independently, resulting in three distinct factor matrices:

$$F^{(ChatGPT)}, \quad F^{(Gemini)}, \quad F^{(DeepSeek)} \tag{11}$$

These factor matrices were subsequently combined with the finalized consensus weight vector to compute platform-specific Righteousness Index values in the following section.

4.4 Righteousness Index Computation Results

To construct the empirical evaluation sample, an initial candidate universe was compiled from historically recognized and publicly documented baseball figures, including Hall of Fame inductees, award recipients, statistical leaders, and players widely referenced in discussions of ethical influence and social impact. This preliminary dataset was not manually filtered for normative judgment but instead subjected to an AI-based consensus screening procedure. Three generative AI platforms—ChatGPT, Gemini, and Copilot—independently evaluated each candidate based on whether the player could be characterized as demonstrating righteous qualities according to publicly available historical records and documented behavior. Each platform provided a binary judgment ("Yes" or "No") for every figure. Table 10 summarizes the results of this multi-model consensus evaluation. The "Count" column indicates the number of platforms agreeing that a player satisfies the righteousness criterion. Players achieving a consensus threshold of Count ≥ 2 were retained for subsequent quantitative analysis. This mechanism reduces subjective researcher bias while preserving transparency in the initial candidate selection process.

Table 6 Righteous Baseball Figures – AI Consensus

Person	ChatGPT	Gemini	Copilot	Count
Jackie Robinson (1919–1972)	Yes	Yes	Yes	3
Roberto Clemente (1934–1972)	Yes	Yes	Yes	3

Lou Gehrig (1903–1941)	Yes	Yes	Yes	3
Hank Aaron (1934–2021)	Yes	Yes	Yes	3
Buck O’Neil (1911–2006)	Yes	Yes	Yes	3
Cal Ripken Jr. (1960–Present)	Yes	No	Yes	2
Ichiro Suzuki (1973–Present)	Yes	No	Yes	2
Clayton Kershaw (1988–Present)	Yes	Yes	No	2
Dale Murphy (1956–Present)	No	Yes	Yes	2
Jim Abbott (1967–Present)	No	Yes	Yes	2
Minnie Miñoso (1925–2015)	No	Yes	Yes	2
Carl Erskine (1916–2013)	No	Yes	Yes	2
Branch Rickey (1881–1965)	No	Yes	Yes	2
Christy Mathewson (1880–1925)	Yes	Yes	No	2
Derek Jeter (1974–Present)	Yes	No	No	1

To operationalize the proposed framework, the standardized factor scores produced by each generative AI platform were aggregated using the consensus weight vector defined in Section 3. This step transforms platform-specific factor evaluations into comparable Righteousness Index (RI) values. Tables 7–10 present the platform-specific factor score matrices and the resulting Righteousness Index (RI) values generated from ChatGPT, Gemini and DeepSeek. Each generative AI platform independently processed the standardized dataset using the consensus weight vector defined in Section 3 to compute factor-level scores across the 11 righteousness dimensions.

Table 7 presents the factor score matrix produced by ChatGPT using the same methodological structure and input data. While the evaluation framework remains consistent, numerical variations reflect platform-specific reasoning differences. Table 8 reports the factor score matrix generated by Gemini under the unified framework. The table contains standardized evaluations for each retained player and serves as the computational input for deriving the Gemini-based RI values. Table 9 shows the factor score matrix generated by DeepSeek under identical conditions. The values represent an independent assessment of the same candidate pool based on the predefined scoring criteria.

Using the RI formulation specified in Section 3, the weighted aggregation was applied to each platform-specific factor matrix to compute the Righteousness Index for every player. Table 10 summarizes the resulting RI values, including platform-specific scores and the averaged index used for the final consensus ranking.

Table 7 ChatGPT-Based Factor Score Matrix

Person	ETHIC	FAIR	ACCNT	SOCIAL	COURAGE	ANTIC	HONEST	RESPCT	LONG	CHAR	COMP
J. Robinson	94	98	87	94	100	92	88	98	99	93	89
Clemente	96	99	91	99	94	91	92	100	99	96	99
L. Gehrig	94	97	93	88	98	99	93	100	97	99	91
H. Aaron	93	98	94	95	100	93	90	99	99	94	92
B. O'Neil	94	98	89	96	94	91	89	98	97	99	94
C. Ripken	94	99	99	89	91	98	94	100	100	92	90
I. Suzuki	89	93	86	81	88	94	88	93	96	88	82
C. Kershaw	90	94	92	89	84	94	91	94	91	89	84
D. Murphy	93	99	87	82	89	95	91	100	86	98	85
J. Abbott	98	100	92	88	100	98	94	95	90	94	96
M. Miñoso	94	89	88	83	91	90	85	94	92	89	82
C. Erskine	93	98	89	83	90	94	87	95	84	91	84
B. Rickey	88	93	98	98	93	88	83	89	100	92	86
Mathewson	94	95	89	82	90	98	88	94	91	93	83

Table 8 Gemini-Based Factor Score Matrix

Person	ETHIC	FAIR	ACCNT	SOCIAL	COURAGE	ANTIC	HONEST	RESPCT	LONG	CHAR	COMP
Robinson	98	95	92	99	100	90	94	96	99	97	91
Clemente	96	94	91	100	98	93	95	97	98	96	100
L. Gehrig	97	98	99	88	96	95	98	92	95	99	90
H. Aaron	96	95	94	98	99	92	96	95	98	97	92
B. O’Neil	95	96	90	97	95	94	97	99	96	98	100
C. Ripken	99	98	100	90	91	96	98	94	97	98	89
I. Suzuki	98	99	96	85	90	97	95	98	94	98	86
Kershaw	95	94	93	99	88	96	96	95	92	96	98
D. Murphy	97	97	95	94	92	98	99	96	90	98	96
J. Abbott	94	96	95	92	99	95	97	94	88	97	95
M. Miñoso	94	92	89	95	97	90	93	95	94	94	93
C. Erskine	96	95	94	98	93	96	96	99	93	97	100
B. Rickey	92	90	95	96	98	94	91	97	99	90	88
Mathewson	98	100	96	85	92	97	98	95	95	99	87

Table 9 DeepSeek -Based Factor Score Matrix

Person	ETHIC	FAIR	ACCNT	SOCIAL	COURAGE	ANTIC	HONEST	RESPCT	LONG	CHAR	COMP
J.Robinson	98	95	90	97	99	88	96	95	94	97	92
Clemente	97	94	92	99	96	90	95	98	98	98	98
L. Gehrig	98	97	95	88	96	92	98	94	96	99	90

H. Aaron	96	95	94	95	98	92	97	96	97	97	91
B. O'Neil	98	96	95	98	95	94	97	99	99	98	97
C. Ripken	96	97	95	94	92	94	96	93	98	98	91
I. Suzuki	95	98	94	88	89	95	96	94	95	97	87
C.Kershaw	96	97	94	96	91	95	95	93	94	96	93
D. Murphy	98	97	96	92	93	96	98	95	95	98	92
J. Abbott	97	98	95	91	96	95	97	96	94	98	94
M. Miñoso	95	93	89	94	96	88	94	97	95	94	93
C. Erskine	97	96	94	93	92	93	96	95	95	96	94
B. Rickey	94	92	91	95	98	90	95	97	96	93	89
Mathewson	96	95	90	88	89	88	95	92	91	96	87

Table 10 Multi-Platform Computed Righteousness Index and Aggregated Results

Person	RI_ChatGPT	RI_Gemini	RI_DeepSeek	Avg_RI
J. Robinson	93.91	95.7	94.74	**94.78**
R. Clemente	95.81	96.17	96.25	**96.08**
L. Gehrig	95.43	95.44	95.44	**95.44**
H. Aaron	95.17	95.55	95.16	**95.29**
B. O’Neil	94.38	95.36	96.67	**95.47**
C. Ripken	95.35	95.96	95.33	**95.55**
I. Suzuki	89.11	94.03	94.2	**92.45**
C. Kershaw	90.38	94.23	94.6	**93.07**
D. Murphy	92.21	95.44	95.69	**94.45**
J. Abbott	95.67	94.34	95.55	**95.19**
M. Miñoso	89.07	92.76	93.22	**91.68**
C. Erskine	90.44	95.49	94.89	**93.61**
B. Rickey	91.35	93.38	93.28	**92.67**
C. Mathewson	91.42	94.83	92.18	**92.81**

4.5 Cross-Platform Ranking Agreement and Robustness Analysis

Based on the computed Righteousness Index values presented in Section 4.4, players were ranked independently according to the RI values generated by ChatGPT, Gemini, and DeepSeek. In addition, an averaged index (Avg_RI) was calculated to obtain a consensus-based ranking across platforms.

Table 11 presents the ranking positions derived from each platform—ordered as ChatGPT, Gemini, and DeepSeek—along with the consensus ranking based on the averaged RI values. The comparison enables direct assessment of inter-model variability at the ordinal level under a unified weighting structure.

To quantitatively evaluate ranking stability, Spearman's rank correlation coefficients were computed between each pair of platform-specific rankings. Table 12 reports the resulting correlation matrix. Spearman correlation measures monotonic agreement between ranked lists and provides a statistical indicator of cross-platform consistency. The correlation values reflect the degree to which player rankings remain stable across platforms. High correlations indicate strong robustness of the proposed framework, whereas lower correlations suggest that platform-specific reasoning differences significantly influence ranking outcomes and may require further sensitivity analysis.

Table 11 Platform-Based Ranking Comparison

Player	Rank_ChatGPT	Rank_Gemini	Rank_DeepSeek	Rank_Avg
R. Clemente	1	1	1	1
C. Ripken	3	2	6	2
L. Gehrig	4	3	5	3
H. Aaron	5	4	7	4
B. O'Neil	6	5	2	5
J. Abbott	2	6	4	6
D. Murphy	7	7	3	7
C. Erskine	8	8	8	8
J. Robinson	9	9	9	9

B. Rickey	10	10	10	10
C. Mathewson	11	11	11	11
C. Kershaw	12	12	12	12
I. Suzuki	13	13	13	13
M. Miñoso	14	14	14	14

Table 12 Spearman Rank Correlation Matrix

	ChatGPT	**Gemini**	**DeepSeek**
ChatGPT	1	0.96	0.97
Gemini	0.96	1	0.95
DeepSeek	0.97	0.95	1

4.6 Robustness and Validation Analysis

The empirical application demonstrates the practical implementation of the proposed Righteousness Index framework within a multi-model AI environment. By applying the unified weighting structure to platform-specific factor matrices, comparable quantitative scores were generated across ChatGPT, Gemini, and DeepSeek.

The results indicate that despite variations in platform-specific factor scores, the resulting Righteousness Index values exhibit relatively stable ranking patterns across models. The averaged RI serves as a consensus measure that mitigates individual model bias and provides a unified ranking representation.

The ranking comparison presented in Section 4.5 shows strong ordinal consistency across platforms, while the Spearman correlation coefficients confirm high monotonic agreement among model-generated rankings. This suggests that the proposed framework is robust to variations in generative model reasoning and maintains stable relative ordering under independent computational implementations.

Overall, the empirical results validate the feasibility of using large language models as structured evaluators within a quantitative ethical assessment framework. The high cross-platform agreement further supports the stability and reproducibility of the proposed index under different AI implementations.

5. Discussion and Conclusion

5.1 Interpretation of Empirical Results

The empirical results demonstrate that the proposed Righteousness Index framework produces stable and interpretable quantitative evaluations across multiple generative AI platforms. Despite minor variations in factor-level scoring, the aggregated Righteousness Index values exhibit strong consistency in ranking patterns among the evaluated players.

Across ChatGPT, Gemini, and DeepSeek, top-ranked figures consistently include historically recognized individuals associated with documented social impact and ethical influence. This convergence suggests that the unified weighting structure effectively captures shared normative judgments embedded within different AI reasoning systems. The relatively high Spearman correlation coefficients reported in Section 4.5 indicate strong ordinal agreement among platform-specific rankings. Such consistency implies that the proposed framework is robust to variations in model architecture and reasoning mechanisms. Although absolute RI values differ slightly across platforms due to interpretive differences in factor scoring, the relative ordering of candidates remains largely stable.

Minor ranking fluctuations observed among mid-tier players reflect sensitivity to contextual interpretation of qualitative attributes. These differences highlight that while AI-based scoring provides structured evaluation, some factors inherently involve subjective assessment, which contributes to variation across models. Overall, the empirical findings suggest that the framework successfully transforms qualitative ethical considerations into a reproducible quantitative structure while maintaining cross-platform robustness.

5.2 Research Contributions

This study contributes to the emerging field of AI-assisted quantitative evaluation by introducing a structured framework for transforming normative concepts into measurable indices through multi-model consensus computation.

First, the study proposes a formalized Righteousness Index (RI) model that operationalizes abstract ethical dimensions into standardized factor scores aggregated through a transparent weighting mechanism. Unlike purely qualitative ethical assessments, the framework converts subjective judgments into reproducible quantitative outputs. Second, the methodology integrates multiple generative AI platforms as independent evaluators rather than relying on a single model. By comparing outputs from ChatGPT, Gemini, and DeepSeek and aggregating their results through consensus weighting, the framework reduces model-specific bias and increases robustness. This multi-model design strengthens methodological reliability and introduces a novel approach to AI-based evaluation validation. Third, the study introduces an AI-assisted candidate screening procedure combined with consensus filtering to construct the empirical sample. The use of cross-platform agreement as an initial selection mechanism provides a reproducible and transparent approach to dataset construction.

Finally, the empirical application to professional baseball demonstrates how large language models can function not only as generative systems but also as structured analytical tools for socio-ethical evaluation. The integration of quantitative modeling with AI reasoning establishes a transferable framework that can be applied to other domains beyond sports.

5.3 Practical Implications

The proposed Righteousness Index framework offers practical implications for sports evaluation, ethical assessment, and AI-assisted analytical systems.

First, within professional sports contexts, the framework provides a complementary evaluation dimension beyond traditional performance statistics. Baseball analysis has historically emphasized quantitative performance metrics such as batting average, WAR, and career achievements. However, legacy evaluation—particularly in discussions involving Hall of Fame recognition or historical standing—often includes qualitative judgments about character, leadership, and social impact. The Righteousness Index introduces a structured mechanism for incorporating such normative dimensions into a transparent and replicable quantitative framework. While the index is not intended to replace human deliberation, it can function as an analytical support tool to structure ethical discussion. **Second**, the framework demonstrates how generative AI systems can be operationalized as structured evaluators rather than purely narrative generators. By constraining outputs into standardized factor matrices and applying a unified weighting structure, large language models become components of a reproducible assessment system. This has broader implications for AI-assisted evaluation in domains such as leadership assessment, corporate governance analysis, and institutional reputation studies. **Third**, the multi-model validation design enhances practical credibility. In applied settings where AI-based scoring tools may inform decision processes, reliance on a single model can introduce hidden bias or instability. The cross-platform consensus approach illustrated in this study provides a blueprint for increasing reliability through independent model comparison and rank aggregation. **Finally**, the framework contributes to methodological transparency. Because factor dimensions and weights are explicitly defined, stakeholders can audit the evaluation structure, adjust parameters, or conduct sensitivity tests. This interpretability distinguishes the approach from opaque algorithmic scoring systems and supports responsible AI deployment in evaluative contexts.

5.4 Robustness and Sensitivity Analysis

The robustness of the proposed Righteousness Index framework is primarily supported by the strong cross-platform ranking consistency observed in Section 4.5. Despite minor variations in factor-level scoring across ChatGPT, Gemini, and DeepSeek, the resulting rankings exhibit high ordinal agreement. The elevated Spearman correlation coefficients indicate that the relative ordering of evaluated players remains stable under independent model implementations. This suggests that the framework's structural design—rather than platform-specific reasoning patterns—drives the overall ranking outcomes. The stability observed among top-ranked and bottom-ranked candidates further reinforces this conclusion. While small fluctuations occur among mid-ranked individuals, such variations are expected in evaluations involving qualitative ethical dimensions. These minor shifts do not materially alter the overall hierarchical structure, indicating that the aggregation mechanism effectively absorbs localized scoring differences.

From a sensitivity perspective, the weighting configuration represents a critical structural parameter. Although the consensus weight vector was designed to balance multiple ethical dimensions, alternative weighting schemes could influence absolute index values and marginal rank positions. However, the consistent ranking patterns observed under the unified weight structure suggest that the model is not excessively sensitive to isolated factor interpretations. *Importantly*, the multi-model architecture itself functions as an internal robustness mechanism. By incorporating independent evaluations from multiple generative AI systems and aggregating their outputs, the framework reduces the risk of single-model bias. This design enhances reliability and increases confidence in the reproducibility of results across different AI implementations.

Overall, the evidence indicates that the Righteousness Index demonstrates satisfactory robustness under cross-platform application while acknowledging that normative evaluation inherently contains elements of contextual interpretation. The framework balances structured quantification with recognition of qualitative complexity, contributing to methodological stability without overstating determinism.

5.5 Limitations

Despite its methodological contributions, this study has several limitations that should be acknowledged. **First**, the Righteousness Index relies on factor-level evaluations generated by large language models. Although multi-model comparison reduces single-platform bias, the scoring process remains dependent on AI interpretation of historical narratives and publicly available information. Variations in training data, embedded biases, or contextual emphasis across models may influence factor assessments. **Second**, the conceptualization of "righteousness" inherently involves normative judgment. While the study operationalizes this construct through eleven standardized dimensions, ethical evaluation cannot be fully separated from cultural, historical, and contextual interpretation. Different societies or evaluators might prioritize certain dimensions differently, potentially affecting weight configurations and ranking outcomes. **Third**, the candidate selection process, although structured through cross-platform consensus filtering, is limited to a predefined set of notable baseball figures. The dataset does not represent the full population of professional players, and therefore the rankings should be interpreted within the scope of the selected sample rather than as universal moral standings. **Fourth**, the alternative weighting schemes or stakeholder-informed calibration may yield different quantitative outcomes. While the observed ranking stability suggests structural robustness, further sensitivity testing under varied weight configurations would strengthen generalizability. Finally, the study focuses exclusively on the domain of professional baseball. Although the framework is theoretically transferable, its applicability to other fields requires empirical validation. Differences in contextual norms, performance metrics, and historical documentation may influence model behavior outside the tested domain.

Overall, these limitations highlight that the proposed framework should be interpreted as a structured evaluative tool rather than an objective moral determinant. Future research can address these constraints through expanded datasets, alternative weighting experiments, and domain diversification.

5.6 Future Research Directions

Building upon the present framework, several avenues for future research emerge. **First**, expanded sensitivity analysis could further strengthen methodological validation. Future studies may systematically vary weighting structures, conduct Monte Carlo simulations, or incorporate stakeholder-driven weight calibration to evaluate how alternative configurations influence ranking outcomes. Such extensions would provide deeper insight into parameter stability and normative trade-offs. **Second**, future research could broaden the empirical dataset. Applying the Righteousness Index to a larger population of professional baseball players, including contemporary and lesser-known figures, would enhance generalizability and test scalability. Automated data pipelines combined with structured AI prompts could facilitate larger-sample implementation. **Third**, cross-domain application represents a promising direction. The conceptual framework may be adapted to evaluate leadership ethics in business executives, political figures, nonprofit organizations, or institutional governance contexts. Comparative studies across domains could reveal how normative dimensions function under different social and cultural environments. **Fourth**, methodological integration with quantitative behavioral data may further enrich the framework. Combining AI-based narrative assessment with objective indicators—such as documented philanthropic contributions, disciplinary records, or community engagement metrics—could enhance empirical grounding and reduce reliance on narrative interpretation alone. **Finally**, advances in large language model architectures provide opportunities for continuous validation. As generative AI systems evolve, longitudinal comparison of model outputs over time may reveal how ethical reasoning patterns shift across training generations. Such analysis would contribute to broader discussions on AI interpretability and normative modeling.

Collectively, these directions position the Righteousness Index not as a static scoring tool, but as an evolving framework capable of adaptation, expansion, and interdisciplinary integration.

6. Conclusion

This study introduced a structured Righteousness Index (RI) framework designed to operationalize normative ethical evaluation through multi-model generative AI systems. By transforming qualitative ethical dimensions into standardized factor scores and aggregating them under a unified weighting structure, the framework enables reproducible quantitative assessment of historically recognized baseball figures.

The empirical results demonstrate that, despite minor variations in factor-level interpretation across ChatGPT, Gemini, and DeepSeek, ranking outcomes remain highly consistent under identical weighting conditions. The observed cross-platform agreement suggests that the proposed aggregation structure captures shared evaluative logic embedded within independent AI systems. This stability supports the robustness and methodological validity of the framework.

Beyond its application to professional baseball, the study contributes to the broader discussion of AI-assisted evaluation. It illustrates how large language models can function not only as generative text systems but also as structured evaluative agents when embedded within transparent quantitative architectures. The multi-model consensus approach further strengthens reliability by mitigating single-model bias.

While normative evaluation inherently involves contextual interpretation, the Righteousness Index demonstrates that ethical dimensions can be systematically structured without eliminating interpretability. The framework therefore provides a replicable foundation for future interdisciplinary research integrating artificial intelligence, quantitative modeling, and socio-ethical assessment.

In summary, this study establishes a methodological pathway for transforming abstract moral constructs into stable, auditable, and cross-platform quantitative indices, contributing both to applied sports analytics and to emerging research on AI-driven evaluative systems.

References

Aristotle. (1998). *Nicomachean ethics* (D. Ross, Trans.). Oxford University Press.

Bok, S. (1978). *Lying: Moral choice in public and private life*. Pantheon Books.

Bok, S., & Green, R. M. (2025). *Honesty and public ethics*. Harvard University Press.

Bovens, M. (2025). *Accountability and democratic governance*. Cambridge University Press.

Brooks, D. (2015). *The road to character*. Random House.

Brooks, D. (2025). *Humility and moral leadership*. Random House.

Brown, B. (2018). *Dare to lead*. Random House.

Brown, M. E., & Treviño, L. K. (2006). Ethical leadership: A review and future directions. *The Leadership Quarterly, 17*(6), 595–616. https://doi.org/10.1016/j.leaqua.2006.10.004

Brown, M. E., Treviño, L. K., & Harrison, D. A. (2005). Ethical leadership: A social learning perspective for construct development and testing. *Organizational Behavior and Human Decision Processes, 97*(2), 117–134. https://doi.org/10.1016/j.obhdp.2005.03.002

Brynjolfsson, E., & McAfee, A. (2017). *Machine, platform, crowd: Harnessing our digital future*. W. W. Norton.

Carroll, A. B. (1991). The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business Horizons, 34*(4), 39–48. https://doi.org/10.1016/0007-6813(91)90005-G

FIFA. (2018). *FIFA code of ethics*. Fédération Internationale de Football Association.

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment, 5*(4), 210–233. https://doi.org/10.1080/20430795.2015.1118917

Heald, D. (2006). Transparency as an instrumental value. In C. Hood & D. Heald (Eds.), *Transparency: The key to better governance?* (pp. 59–73). Oxford University Press.

Kidder, R. M., & Miller, S. (2024). *Moral courage in leadership*. HarperCollins.

Lessig, L. (2011). *Republic, lost: How money corrupts Congress—and a plan to stop it*. Twelve.

Lickona, T. (1991). *Educating for character: How our schools can teach respect and responsibility*. Bantam Books.

Lippmann, W. (1922). *Public opinion*. Harcourt, Brace and Company.

MacAskill, W. (2023). *What we owe the future*. Basic Books.

MacIntyre, A. (1981). *After virtue*. University of Notre Dame Press.

Machiavelli, N. (1532/1998). *The prince* (G. Bull, Trans.). Penguin Classics.

Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing.

Nussbaum, M. C. (2024). *Compassion and justice*. Harvard University Press.

Porter, M. E., & Kramer, M. R. (2011). Creating shared value. *Harvard Business Review, 89*(1–2), 62–77.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Rawls, J., & Thompson, D. F. (2023). *Justice and democratic fairness*. Harvard University Press.

Rose-Ackerman, S. (1999). *Corruption and government: Causes, consequences, and reform*. Cambridge University Press.

Rousseau, J.-J. (1762/1997). *The social contract* (V. Gourevitch, Trans.). Cambridge University Press.

Sen, A. (2024). *Identity and inclusion in democratic societies*. Harvard University Press.

Shields, D. L., & Bredemeier, B. J. (2007). *Advances in sport morality research*. Human Kinetics.

Simon, R. L. (2015). *Fair play: The ethics of sport* (4th ed.). Westview Press.

Simons, T. (2002). Behavioral integrity: The perceived alignment between managers' words and deeds. *Organization Science, 13*(1), 18–35. https://doi.org/10.1287/orsc.13.1.18.543

Singer, P. (2024). *The life you can save* (Updated ed.). Random House.

Transparency International. (2023). *Corruption perceptions index 2023*. https://www.transparency.org

Treviño, L. K. (2004). Managing ethics and legal compliance. *Academy of Management Executive, 18*(2), 131–151.

Treviño, L. K., Brown, M., & Hartman, L. P. (2003). A qualitative investigation of perceived executive ethical leadership. *Human Relations, 56*(1), 5–37. https://doi.org/10.1177/0018726703056001448

Walzer, M. (1977). *Just and unjust wars*. Basic Books.

Williams, B. (2025). *Integrity and moral philosophy*. Oxford University Press.

BOOK REVIEW

Fujii, K. (2025). *Machine Learning in Sports: Open Approach for Next Play Analytics.* Springer. https://doi.org/10.1007/978-981-96-1445-5

Reviewed by
ChienMin Kuo
Baseball Innovation League Association
Sports Pioneer Journal (eISSN: 3070-0353)

The increasing integration of artificial intelligence and advanced analytics into professional sport has fundamentally reshaped how teams evaluate performance, design tactics, and generate competitive advantage. In *Machine Learning in Sports: Open Approach for Next Play Analytics*, Keisuke Fujii (2025) presents a timely and methodologically rigorous exploration of machine learning applications within sports environments. As data-driven innovation becomes central to organizational strategy in professional leagues, this volume offers both conceptual clarity and technical guidance for researchers and practitioners seeking to understand predictive modeling in sport contexts.

Fujii's central premise is that next-play analytics—defined as probabilistic modeling designed to anticipate imminent in-game actions—represents a frontier in sports analytics research. The book begins by outlining foundational machine learning principles, including supervised and unsupervised learning, classification, regression techniques, and model evaluation frameworks. Rather than offering purely abstract explanations, Fujii situates these methods within sports data ecosystems, such as player tracking systems, event-based datasets, and real-time performance metrics. This applied orientation strengthens the book's accessibility for interdisciplinary audiences bridging computer science and sport management.

A notable contribution of the text is its emphasis on reproducibility and transparency. The "open approach" advocated by Fujii aligns with contemporary scholarly standards promoting replicable research designs and open-source modeling practices. In a field often constrained by proprietary data and closed analytical systems, this commitment to methodological openness is both ethically significant and academically progressive. By detailing model validation techniques and performance evaluation metrics, the author encourages critical scrutiny rather than blind adoption of algorithmic outputs.

Subsequent chapters delve into feature engineering strategies and contextual modeling challenges unique to sports. Fujii acknowledges that sports performance data are inherently dynamic, nonlinear, and context-dependent. Issues such as data imbalance, situational variability, and temporal dependencies complicate predictive accuracy. The discussion of overfitting and model generalizability is particularly valuable, as many applied sport analytics projects struggle to balance precision with robustness. The text effectively demonstrates how rigorous model validation frameworks can mitigate these risks.

From a critical perspective, the book’s greatest strength lies in its methodological coherence. Fujii successfully integrates statistical reasoning with practical implementation, offering readers both theoretical foundations and operational insight. The mathematical explanations are presented with sufficient clarity to support graduate-level understanding, while remaining grounded in sports applications. This dual emphasis enhances the book’s utility for scholars in sport analytics, engineering, and innovation studies.

Nevertheless, the technical orientation may present challenges for readers without prior exposure to statistical modeling or programming concepts. While Fujii attempts to maintain accessibility, certain sections presume familiarity with machine learning terminology. Readers from purely managerial or coaching backgrounds may require supplementary resources to fully engage with the material. Additionally, although the frameworks are broadly applicable across sports, more detailed sport-specific case studies—particularly within globally prominent leagues such as professional baseball—could further strengthen the applied dimension of the work.

In terms of contribution, *Machine Learning in Sports* advances scholarly dialogue at the intersection of technology adoption, performance optimization, and innovation management. As professional baseball and other competitive leagues increasingly rely on tracking technologies and algorithmic decision-support systems, understanding the computational infrastructure underlying these innovations becomes essential. Fujii’s framework provides a foundation for evaluating not only predictive accuracy but also organizational implementation strategies and ethical considerations surrounding data governance.

For researchers examining sport innovation ecosystems, the text offers methodological tools that can inform empirical studies on competitive balance, tactical efficiency, and player development systems. For practitioners, it clarifies the analytical logic behind next-play modeling, thereby facilitating more informed collaboration between data scientists and coaching staff. The open-access availability of the volume further enhances its scholarly reach, enabling broader dissemination across academic and professional communities.

Overall, Fujii's work represents a significant and timely contribution to sports analytics literature. While deeper sport-specific exemplification would enhance its practical resonance, the book succeeds in articulating a coherent and forward-looking framework for machine learning integration in sport. For scholars and professionals seeking to understand the evolving architecture of data-driven competition, *Machine Learning in Sports* offers both conceptual insight and methodological rigor.