ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 10 Issue 4 July - August 2022

International Manuscript ID : 23482001V10I4072022-02

(Approved and Registered with Govt. of India)

PREDICTION AND ANALYTICS OF CRICKET PLAYERS USING ARTIFICIAL INTELLIGENCE

Dr. Vikas Mehta
Associate Professor in Physical Education
Sri Guru Hari Singh College
Sri Jiwan Nagar, Sirsa, Haryana

Abstract

Cricket is one of the most prominent sports in India and millions of fans await its international tournaments. Nowadays, lots of the techniques and strategies and techniques are dependent on technology for the analysis of sports activities and these enormous technological aspects are associated with Artificial Intelligence and Machine Learning. With the usage of specialized domain of artificial intelligence, the approaches help in predictive abilities over time without being explicitly taught how to do so by humans. Predictions from machine learning algorithms can be done on the historical data. In addition to aiding in product creation, machine learning helps businesses keep tabs on shifting client preferences and organisational tendencies. Facebook, Google, and Uber are just a few of the industry leaders who use machine learning extensively. Many businesses now use machine learning as a key differentiation in the market. To know where a stock price is going or to know whether or not a consumer will buy your goods in the future seems like a magical power. Being able to accurately foresee the future gives us a significant edge. The advent of machine learning has only added to the mystique and wonder of it all. Predicting the outcome of a sporting event is done primarily with the intention of boosting a team's performance and consequently their odds of winning. Winning has several downstream effects, including increased attendance, ticket sales, retail sales, parking revenue, food sales, sponsorships, and even student enrollment and retention. In this research manuscript, the integration of machine learning is used for the prediction of the performance of cricket player. With the use of logistic regression, the prediction of batsman can be done whether that sportsperson will be able to perform well or not in the upcoming tournament.

Keywords : Prediction of Sportspersons with Machine Learning, Sports and Artificial Intelligence, Machine Learning and Sports

Introduction

Cricket is second only to football in terms of worldwide participation. As soon as the British brought it to India, it became a widely played sport. Despite not being recognised as India's official national sport, cricket is the country's most popular sport. The sport of cricket is expanding in popularity in the country, and as a result, there is a substantial betting industry developing around it. When it comes to attracting new customers, cricket betting presents an unrivalled potential for online bookmakers. India is not only a country that loves sports, but also one that loves to wager on those games. The combination of cricket and betting has led to the sport's meteoric rise in popularity [1].

In India, the sport of cricket is practically a faith. Even young children may learn the ropes of the game with relative ease. Millions of people in India watch the results of every cricket game. This shows just how popular cricket is in India. When it comes to their favourite sports team, fans will go to any mile to show their loyalty. Betting on the game is a fun and profitable way for fans to demonstrate their allegiance to their side and get some extra cash at the same time.

Machine Learning and Sports

The use of machine learning to forecast sports results has been propelled forward in large part by bookmakers and the betting industry. With a 2019 estimated worth of $85 billion, the motivation is more than clear. Incredibly, many prediction markets are already rather reliable. Although experts aren't always accurate when predicting an election's outcome, the betting markets usually end up being correct as in the US presidential election. In general, the odds in sports betting markets are quite accurate, especially for soccer. In this day and age, we can make more accurate forecasts, and this is a major factor. Machine learning algorithms are able to sift through massive amounts of data and extrapolate results (predictions) at a scale and speed that would be impossible for humans to match [2].

Many people, not only those working in sports, find it interesting to think about, forecast outcomes for, and extract useful information from sports, especially in the contexts of team management and sports betting. When evaluating ML algorithms, researchers frequently employ data segmentation using data organised in a time-series format. Aside from data segmentation, the k-cross test is also used to determine the quality of study results. Predictions in sports are typically viewed as a classification issue, with the unusual exceptions being represented numerically. Neural network models that segment data are the most common type of ML model [3].

Most machines that apply machine learning are replacing human workers. In the same way, you can't predict sports outcomes. Since the inception of professional sports, coaches, analysts, and gamblers have all attempted to forecast the outcomes of individual games [4].

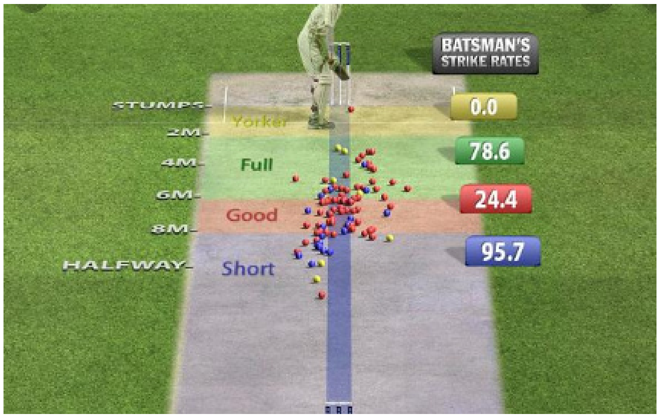


Figure 1 : Analysis of Bowling Activities using Artificial Intelligence

Usage Patterns of Artificial Intelligence in Sports

Massive volumes of data are used by machine learning algorithms to teach them and inform their predictions. Data utilised is similar to what people use to predict the outcome of matches. For the sake of illustration, let's use the sport of soccer as an example. This would contain statistics like the team's and individual's goals, assists, and possessions, as well as the outcomes of any past games between the two teams.

Figure 2 : Evaluation of Cricket Pitch with Bowling Analytics

The primary use of machine learning in sports prediction is the construction of a classification model from a training data set, where the initial data is provided to the algorithm in order for it to find patterns and generate predictions. Both supervised and unsupervised approaches can be used to train a new system. Unlike the latter, which only uses input data, the former uses both input and output data to construct its prediction models [5]. Choosing the correct data collection is the most challenging component of using ML to predict match results. Making accurate predictions with machine learning requires more than just feeding the algorithm all the data you think could be relevant. The accuracy of these forecasts varies, and it's not uncommon for them to be lower than the odds offered by bookies.

As a result, several studies use betting odds as an extra component to improve the accuracy of machine learning systems. To incorporate even more variables, artificial neural networks (ANN) can be used. ANN are designed to mimic the functioning of the human brain. This in turn leads to increasingly more accurate forecasts, but they are still often only around 50% accurate.

Improved national health is a byproduct of rising living standards, and the sports prowess of the nation's youth is a useful barometer of the industry's progress [1]. Therefore, all nations place a premium on the ability to forecast and analyse pupils' athletic performance. Teachers of PE and those in charge of PE programmes might benefit from learning to forecast their students' performances in order to have a handle on the ever-changing dynamics of their students' physical abilities [2]. Improve students' performance in physical education classes by tailoring lesson plans to their unique needs and interests. Help students develop stronger skills in the area of physical activity. Researchers have built a multilevel physical education performance analysis system and an effective prediction model [3, 4] after extensively investigating the change characteristics of physical education performance in order to forecast the performance demand of physical education colleges at all levels.

Two major hubs were identified in the process of making predictions based on the data of sports students at all levels. At first node we have the linear model. Multiple methods, including regression analysis and grey identification, are incorporated into the process. An individual sports student's data is used to produce an in-depth prediction of future results using this form. This kind of analysis may be used to relay back to sports students any updates to their sports knowledge, provided that such updates occur only seldom. However, several elements, including mental state, age, and the sports environment, will interfere with most sports students' performance while they work out. Students' sports data will increase in variability and exhibit nonlinear properties if there is sports interference of this type. Therefore, the initial stage's reliance on a single linear prediction method limits the extent to which the sports situation of students can be depicted, increases the variability of prediction results, eliminates the likelihood that the prediction system will be available, and lessens the usefulness of students' sports operation data. Introduce the nonlinear concept and conduct fresh modelling and analytic studies to protect the linear model's optimization impact and actively broaden the models' predictive capabilities. New methods of analysis such as support vector, neural algorithm, etc., have been developed and implemented. Linear analysis provides the backbone of the new modelling system, and multilevel study of the features of variance in students' athletic performance improves both the model's predictive power and its practical use [5].

New-vision prediction model construction is now the standard approach to predicting athletic success. There are three test participants in this neural network [6]. The internal correlation of student athletic data is more precisely analysed with the use of three units, improving upon earlier prediction findings. For a neural network to operate, its starting parameters must be established. The key to reliable prediction is the careful selection of appropriate starting values for the model. Currently, the empirical approach is utilised to assume the starting parameters of the algorithm, which enhances the unpredictability of the results produced by the model algorithm. Inconsistent initialization of a network model reduces its predictive power and undermines its ability to provide reliable forecasts [7, 8]. Starting with the initial parameters as a point of optimization, chaos theory is introduced and combined with a machine learning system to create a novel model for predicting the outcomes of sporting events. At first, we look at students' athletic performance and use a new chaotic analysis technique to determine the sports performance change rule [9]. Next, implement a neural algorithm to process the data quickly and accurately. Through online training with processed motion data using the particle swarm optimization approach, the necessary neural system starting parameters may be obtained with high precision. This modelling approach has been proven to have high prediction accuracy of motion information and high research value [10], as well as the ability to properly monitor the future trend of motion data.

The most widely used AI algorithm right now is machine learning. Data categorization, including related tasks like data mining and association, is possible thanks to machine learning techniques. In addition, via data mining, the ability to effectively enhance the future trend of data and make effective forecasts is realised. This type of learning is called guided learning. Quote class variables for data processing in the prediction process; data training, data mining, data modelling, and prediction are all required for the study of sports data.

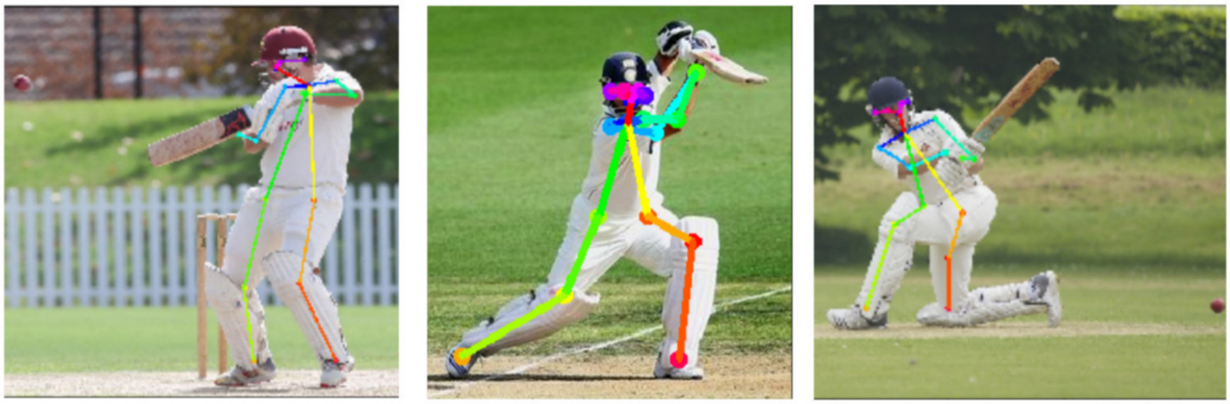
Sports contests and training procedures may be better predicted and modelled with the help of a well-designed exercise programme. Exercise effectiveness may be enhanced by machine learning prediction and creation of sports training regimens. Sports modelling may be used to make predictions about the intensity and duration of future events. The prediction may be improved by using parameter estimate when the sports competition and sports training plan are seen as time series [3]. There are inherent flaws in the conventional sports training prediction systems. With the advancement of fuzzy theory over the past few years, intelligent algorithms based on fuzzy theory and grey theory have found increased usage in competitive sports [4].

Sports games may be viewed as a multi-factor prediction model with time series because of the high intelligence of the machine learning prediction process and the various rule limitations in the games themselves. Machine learning algorithms are more organised and flexible than more conventional prediction methods. Furthermore, it does data analysis by means of machine learning, data training by use of preexisting data, and intelligent learning with the aid of time series algorithms. This paves the way for anticipating the future of sports training and contests, and maximising the impact of intelligent training.

As with the rest of our continually evolving culture, the outcomes of sporting events are never definite. The ability to accurately forecast the outcomes of sporting events is crucial to the growth of the sporting events industry. To make better informed decisions about sports training and competition in the future, it is important to not only be realistic about the challenges already present in the competitive sports arena, but also to anticipate the challenges that may develop in the future. Image recognition, cancer diagnosis, stock market forecasting, and customer retention modelling are just some of the many uses of machine learning in the scientific, medical, and financial communities. The successful use of machine learning is still in its infancy in various fields, including sports.

The algorithms (the "rules" to be followed in computations) employed in machine learning are dubbed supervised learning techniques (e.g. regression and classification) and unsupervised learning methods (e.g. clustering). Injury prediction training load factors would be an example of labelled input data, with injury incidence serving as an example of a labelled output. In contrast, unsupervised learning approaches rely solely on unlabelled input data, with no matching outputs. Since injury prediction often relies on properly labelled training data and player injuries, this work focuses on supervised methods, specifically classification (predicting classes or categories as opposed to continuous values). Any machine learning model's basic goal is to distinguish between positive events like injuries and negative events like non-injuries (a negative class). Linear and logistic regression, decision trees, random forests, k-nearest neighbours (often abbreviated KNN), support vector machines (SVM), artificial neural networks (often abbreviated ANN or NN), and "ensemble methods" (e.g. bagging; and boosting) are all examples of popular supervised machine learning algorithms. There are two main categories of machine learning algorithms: white-box algorithms (such as linear regression, logistic regression, k-nearest neighbours, and decision trees) and black-box algorithms (such as ensemble methods, random forest, artificial neural networks, and support vector machines).

White-box algorithms, also known as interpretable approaches, can be helpful because they reveal the logic behind the algorithm's decisions and reveal the full range of data that went into producing the results. This can allow practitioners and clinicians to better draw clinical and practical conclusions from the research. However, this inputs-to-outcomes mapping is not transparent in black-box algorithms. This means that for the later algorithms, further procedures used after the fact are required if the findings are to be interpreted and understood. The most important takeaway is that all these concepts refer to distinct algorithms that might be utilised, each of which may perform better or worse depending on the specifics of the task at hand.



Cut **Drive** **Sweep**

Figure 3 : Pattern Analysis of Batsman using Computer Vision

Many practical applications of machine learning, such as injury prediction, rely on unbalanced data sets. The number of negative cases (i.e., no injuries) in an imbalanced dataset is far larger than the number of positive examples (i.e. injuries). However, this might generate issues for machine learning algorithms, since they are more likely to learn from the most abundant data sets (in this example, the absence of injuries) and accurately predict the absence of injuries, but not the occurrence of injuries. Studies can use balancing strategies like oversampling (to intentionally produce more injury data points) or undersampling (to eliminate non-injury data points) to create datasets with a more even mix of non-injuries and injuries, hence improving the performance of models using such unbalanced data. In spite of the limitations of both methods, this procedure should provide machine learning models that provide no clear advantage to either injury or non-injury prediction.

Some fit measures, such accuracy and area under the curve (AUC), may be stated as a single value for evaluating classification machine learning models, while others, including precision, recall, and specificity, might have multiple values depending on the choice of the positive class. Assuming injuries are the positive class and non-injuries the negative class, accuracy is the proportion of correct predictions of injuries and non-injuries to the total number of observed injuries and non-injuries, precision is the proportion of correct predictions of injuries to the total number of correct and incorrect predictions of injuries, and recall is the proportion of correct predictions of injuries to the total number of observed injuries (often described as the true posit). (Therefore, it is often said that this statistic represents the best possible balance between accuracy and recall.) Percentages are a common way of expressing these indicators. AUC measures how well a model predicts the true positive and false positive rates, with values closer to 1 indicating better fit.

Precision, recall, specificity, and F1-score are per-class metrics that are often calculated independently for each class (such as injuries and non-injuries) and then averaged to yield a single overall score. While this makes sense in some situations, it can also be deceptive when used to datasets that aren't perfectly balanced, as is frequently the case with football injury statistics. This is because our primary focus is on how well the model works on the "minority class" (in thus example, the injury data, as there likely to be considerably fewer injury than non-injury data points), and this total score does not reflect this. Therefore, in the latter instance, recall and F1-score of just the injury class would be deemed very valuable measures, while accuracy and specificity of both the injury and non-injury data assist to prevent against making inferences that may subsequently be skewed towards the prediction of injuries. Finally, it has been pointed out that AUC, despite its widespread acceptance as a valuable assessment tool, can be deceiving when used to unbalanced data. Since research (including the ones described in the current article) utilise some but not all of these variables, and not the same metrics, comparing studies is a complex procedure.

To elaborate, a standard machine learning research would involve the following steps: data collection, data pre-processing, applications of machine learning methods (i.e. model training), and model assessment. Cleaning (e.g., missing values imputation, handling outliers, anomaly detection), transformation (e.g., data normalisation), feature selection (where only a subset of the original data are used in the model), and feature extraction (where new features are created from the original raw data, to perform better within the machine learning algorithm) are all examples of data pre-processing that can be performed after data collection.

In most cases, the performance of the machine learning algorithms is improved by doing this pre-processing stage rather of feeding them the original raw data. When it comes to measuring how well a machine learning model does its job, two primary methods emerge after data cleaning and preparation. The first method splits the dataset in half, with the larger portion serving as training data, and the smaller portion serving as validation data. This procedure is known as a train-validation split (although it is also frequently termed train-test split). By feeding a machine learning algorithm (such as a decision tree, support vector machine, or artificial neural network) with training data, one can produce a trained model. After the model has been trained, its prediction ability is evaluated using the validation data. The second method involves training a machine learning model with partial data and then testing it on further (validation) partial data. The phrase "cross-validation" describes this procedure. Some researchers, regardless of method, also reserve a subset of the dataset as "test" data; this subset is used to get an objective assessment of the models' efficacy following validation.

The aforementioned evaluation measures (accuracy, precision, recall, specificity, F1-score, and AUC) are then used to determine how well the trained model performs on the (validation or) test data [30]. Overfitting, when a model is very sensitive to the information it has been trained on yet performs poorly when applied to novel validation/test data, is what these techniques aim to mitigate. Depending on the model's success, analysts may go back to the feature selection or hyperparameter optimisation phases to make adjustments to the algorithm's settings or attempt a different machine learning technique. These repetitive and cyclical processes are typical in machine learning. During the training and validation stages, this complete iterative and cyclical process happens.

An important takeaway from the preceding explanation is that pre-processing techniques are applied to all three stages of data (training, validation, and test). However, balancing approaches are only applied to the training data. In fact, it would be counterproductive to balance the validation or test data, as doing so would distort evaluation results and obscure how well the trained model performs on real-world (and imbalanced) data. After completing the preceding stages, it is common practise to evaluate the machine learning model's prediction performance by comparing it to a baseline model. Simple machine learning algorithms or dummy classifiers based on heuristics like "predicting the most common class" can serve as baseline models (i.e. in our case non-injuries).

When it comes to selecting features, baselines often consist of the very minimum. These initial classifiers are arbitrary and are established by the researchers in each every study (i.e. there are no fixed baseline criteria that must be adhered to). Usually, researchers usually try to compare their results with those of similar prior studies; however, this is difficult to do with football injury prediction due to the novelty of the field and (as we mention below) the variations in load factors and assessment techniques utilised by different studies. The ultimate objective is to use the test data to develop a model with the highest possible assessment criteria. Knowing this procedure can help laypeople extract the most important findings from studies conducted with machine learning.

Research Methodology

Following is the dataset of cricket players taken for prediction. The dataset is associated with the behavioral analysis of cricket batsman with their outcome (Out / Retained). With the use of machine learning based approach of logistic regression, the prediction is whether that batsman will be able to handle the bowling conditions or not and will be out or not.

Following is the description of the dataset attributes used :

Player ID :

Unique Player Id Allocated for Prediction using Artificial Intelligence by the associated approach

Bowler Type (0-3)

0 : Spin

1 : Slow Seam

2 : Mixed

3 : Pace

Pitch Type (0-3)

0 : Green Pitch

1 : Dusty

2 : Dead

3 : Spin Supportive

Stadium (Local / Offshore) :

0 : Local

1 : Offshore / International

Out (0/1) :

0 : Out

1 : Retained on Ground / Not-Out

Table 1 : Sample Dataset for Training in Machine Learning Model

Player ID	Bowler Type	Pitch Type	Stadium (Local / Offshore)	Out (0 or 1)
7	3	2	0	0
5	1	3	0	0
5	0	2	0	0
1	0	2	1	1
5	2	2	0	1
2	1	0	0	0
5	1	2	0	0
5	1	1	1	1
0	3	3	0	0
1	2	0	0	0
3	0	1	0	1
3	3	0	0	1
3	1	0	0	1
4	1	0	0	1
6	3	3	1	0
2	2	3	0	0
4	3	1	0	1
0	3	0	0	1
5	1	0	1	1

3	1	0	0	1
1	2	1	0	1
0	0	2	0	1
1	2	1	1	1
1	1	0	0	1
4	3	2	0	1
5	2	1	0	1
2	2	0	0	0
4	0	1	0	0
2	2	2	0	0
2	2	0	0	0
6	3	2	0	0

Results and Outcomes

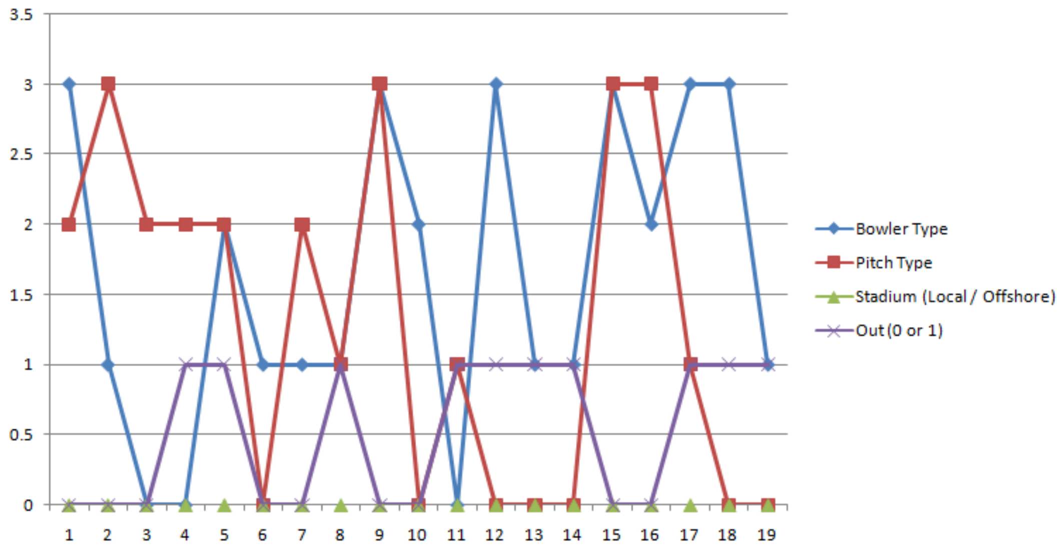


Figure 4 : Prediction and Pattern Analysis of Batsman using Logistic Regression

Table 2 : Dataset for Validation and Testing in Machine Learning Model

Player ID	Bowler Type	Pitch Type	Stadium (Local / Offshore)	Out (0 or 1)
0	3	3	0	0
6	2	1	0	0
1	1	0	0	0
4	0	2	0	0
5	1	1	0	0
1	3	0	0	1
5	3	2	0	0
4	0	3	0	0
6	1	0	0	0
6	3	2	0	0
3	3	1	0	0
1	1	3	0	0
6	0	0	0	0
4	1	3	0	0
4	0	3	0	0
6	0	1	0	0
4	1	0	0	0
2	2	2	0	0

3	3	1	0	0
0	3	3	0	0

Conclusion

This research work is focusing on the prediction of batsman performance using logistic regression that is one of the key approaches in machine learning in integration with artificial intelligence. Data and analytics were once seen as a way for professional sports teams to gain an advantage over their rivals. Data science and analytics for sports are now considered essential. In order to make choices and put new ideas into action more quickly than the competition, these businesses need to do more than just use data, so that their employees can take the most appropriate actions at the most opportune times. Numerous raw data sets are already available to sports organisations, and more are being added regularly. They may now utilise this data to improve ticket sales, player safety, and any other part of the business. DataRobot paves the way for these businesses to leverage artificial intelligence (AI), machine learning (ML), and sports to gain insights and make off-the-field decisions. Using their data, sports organisations may optimise every facet of their operations with the help of machine learning and AI solutions. All aspects of a sports team, from player acquisition and performance to marketing and attendance, can benefit from the use of predictive analytics.

References

[1] Kapadia, K., Abdel-Jaber, H., Thabtah, F., & Hadi, W. (2020). Sport analytics for cricket game results using machine learning: An experimental study. Applied Computing and Informatics, (ahead-of-print).

[2] McGrath, J. W., Neville, J., Stewart, T., & Cronin, J. (2019). Cricket fast bowling detection in a training setting using an inertial measurement unit and machine learning. Journal of sports sciences, 37(11), 1220-1226.

[3] Bhatia, V. (2020, November). A review of Machine Learning based Recommendation approaches for cricket. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 421-427). IEEE.

[4] McGrath, J., Neville, J., Stewart, T., Clinning, H., & Cronin, J. (2021). Can an inertial measurement unit (IMU) in combination with machine learning measure fast bowling speed and perceived intensity in cricket?. Journal of Sports Sciences, 39(12), 1402-1409.

[5] Asif M, McHale IG (2016) In-play forecasting of win probability in one-day international cricket: a dynamic logistic regression model. Int J Forecast 32:34–43. https://doi.org/10.1016/j.ijforecast.2015.02.005

[6] Jhanwar MG, Pudi V (2016) Predicting the outcome of ODI cricket matches: a team composition based approach. In: European conference on machine learning and principles and practice of knowledge discovery in databases (ECML-PKDD) proceedings, vol 1842, pp 111–126

[7] Sankaranarayanan VV, Sattar J, Lakshmanan LVS (2014) Auto-play: a data mining approach to ODI cricket simulation and prediction. Int Conf Data Min SDM 2:1064–1072. https://doi.org/10.1137/1.9781611973440.121

[8] Bunker RP, Thabtah F (2019) A machine learning framework for sport result prediction. J Appl Comput Inf 15:27–33. https://doi.org/10.1016/j.aci.2017.09.005

[9] Asif M, McHale IG (2019) A generalized non-linear forecasting model for limited overs international cricket. Int J Forecast 35:634–640. https://doi.org/10.1016/j.ijforecast.2018.12.003

[10] Chakraborty S, Kumar V, Ramakrishnan KR (2019) Selection of the all-time best World XI Test cricket team using the TOPSIS method. Decis Sci Lett 8:95–108. https://doi.org/10.5267/j.dsl.2018.4.001