**Laiba Nadeem**

**MCS241005**

# Chapter 2

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Overview

This review paper will establish the primary trends and methodologies used in cricket data analysis and establish the literature's shortcomings. It will fill gaps in the literature by proposing decision-making models that consider combinations of performance indicators and team characteristics. This chapter will also shed fresh light on how data science may evolve beyond the standard processes to promote stronger data-driven team management solutions based on current studies.

In turn, this work of literature review will create a background for the further chapters of this project where we will apply modern data analysis techniques to close the above-mentioned gaps and introduce a data-driven system of team selection and strategic management in cricket.

## 2.2 Cricket Data Analysis:

Cricket is a data-intensive game that produces volumes of numeric data at the game level and the player level for every match being played. In the past, factors such as which players to choose in the team, and which strategies to use in the matches have mostly been determined by experience, knowledge, and informal analysis. However, these traditional methods take no note of the massive number of opportunities within available data. Thus, this literature review aims to review the prior literature related to cricket data analysis to understand how contemporary data science interventions have been implemented to improve decision-making in cricket with specific reference to team composition and game tactics.

In this review, several areas of interest in cricket data analysis will be considered: Player performance benchmarking (i.e., average runs per over for batsmen; wickets per over bowlers), team selection predictions, and operational decision making, that is, order selection and choice of bowlers. The analysis will also include the efficiency of data visualization tools and how information is shared to assist in decision-making like Power BI. Furthermore, it shall determine how analytics, and specifically the machine learning techniques have been applied in the game of cricket and whether traditional issues such as team composition, game result forecasting, and player efficiency in various scenarios have been effectively solved.

Over the last few years, there has been a slowly improving understanding of what data science can offer in terms of cricket, with teams and analysts utilizing data to gain information that simply could not be gleaned in the usual course of events. Web scraping, data mining, machine learning algorithms, and data visualization tools have come a long way in enabling teams to extract and analyze huge amounts of player-related information to determine performance and develop game strategies. With this plethora of data available to managers, many facets of team formation are still systematically decided by traditional decision-making protocols that result in decisions that are often subjective anecdotal intelligence.

## 2.3 Literature Review

Technological evolution of data analysis in cricket has changed over time where formerly it covered, forecast prediction, player selection, and strategy enhancement only. Some of the prior works related to cricket involved the use of machine learning, statistical modeling, and IoT for efficient formulation of balanced teams as well as other aspects of player selection and performance analysis.

### 2.3.1 Success indicator, predicting a player’s performance and enhancing the team selection.

A significant application area of cricket analytics is the prediction of the performance of players that determine the selection of teams. Wickremasinghe (2014) did this by using a three-stage hierarchical linear model to provide the probability of the performance of batsmen in test cricket. From the above observations, it is clear that this model incorporates player-specific factors such as players' handedness and general and match-specific factors. The study's results established that player handedness and team rank were statistically significant with player performance. Intriguingly, the home team advantage variable did not influence the result about the performance of the home team in cricket, therefore substantiating dissatisfaction with the home-ground advantage theory in cricket. This research has shown how challenging the process of forecasting cricket performance is due to the presence of numerous factors that have an impact on this process an indication that there is a need for the development of other reliable models that factor in inter-individual and intra-individual differences in performance.

Similarly, Paper 2 speaks of the difficulties in analyzing batsmen's performance in test cricket, with such factors as team rank and match location to be considered. While the paper pointed out that the process of forecasting batsmen's performance was not straightforward, the paper found that predicting using data over more years made it easier to predict results. Based on the issues identified in the study, it was a testimony to the fact that there is a D in the Model which was postulated as increasing in number the number of variables that could be critical to performance, However, the study affirmed the two core constituents of P which are the independent person factors and the interpersonal group factors.

On the other hand, Paper 3 suggests a more exemplary approach and allows the use of machine learning algorithms such as K-means clustering to categorize chiefs based on information from their past performance. This method is used in forming balanced cricket teams and in the process first we look at players who have abilities that supplement each other. The study focuses on using analysis means for the presentation of the simplified work of the coach or selector in choosing a team since the information is segmented effectively. Since the players' selection is based on certain quantitative parameters like strike rates, the method provides a more logical rather than the random involved in selecting players mechanically.

### 2.3.2 An IoT and Data Analytics approach to understand Performance better:

Building on the insights gathered through basic statistics as well as ball-by-ball data, the application of IoT and data analytics have enhanced the analysis of player performance in the context of cricket still further. In Paper 1 the use of a new method is proposed with the timing index that quantifies the quality of the shot performed by the batsman in terms of bat speed, impact bat speed, etc. The timing index seems to be highly set above several conventional statistical parameters and gives a more precise picture of a batsman's timing ability. This method simply involves the use of IoT sensors during training sessions to capture data that impacts batsmen's abilities, the data is then processed using a machine learning classification algorithm. It is useful for applying immediate feedback to the players and coaches during the training process, and it is based on strict metrics indicators. However, some difficulties in normalizing the collected sensor data and incorporating multiple factors have to be mentioned, which are the directions for improvement.

In addition, the study finds that similar forms of analytical IoT can help transform how cricket coaching is done, probably leading to enhanced shots-making abilities of players in the future. Though it is a very nascent concept, IoT and sensor-based analytics can assist coaches to perhaps arrive at a broader and closer characterization of a player's ability, which in turn could be used more powerfully in other areas of the game such as bowling and fielding.

### 2.3.3 Papers based on statistic and machine learning methods for match prediction:

Many works, other than player performance and team lineup prediction, focus on match score prediction: the Bayesian model. Finally, paper 4 offers an analysis of the statistical modeling of a mechanism that could be used to forecast the most appropriate team for a given match. The model includes the average performance of an individual player, its recent trends, and the overall ability of the team to formulate the optimal line-up. This model was able to predict match outcomes at a very high level of accuracy, 91 percent in this case, to illustrate the effectiveness of data analytics in influencing such selection. However, the paper also points to the likelihood of improving the accuracy of the model if training data and some assumptions on player fitness were made. It is possible to extend the model in the future with other predictors, for example, with the actual match data which may reflect important factors such as the injury of one or more players during the match.

However, Paper 3 which employed the K-means clustering algorithm for selecting the teams for a tournament did not have central objectives on match results. Instead, it tried to balance team makeup, that is, which team stats the players would complement each other in gameplay. Both papers stress the fact that data drive assists in the decision-making process in cricket, while Paper 4 combines match prediction and Paper 3 offers a tool for advanced team forming.

### 2.3.4 Differential Analysis in the Analysis of Data Analysis: Challenges and limitations of data analytics in cricket.

Therefore, there are still several issues to be addressed about the application of data-driven methods. Among the many challenges outlined in the research proposals, one of the most significant is data quality. To be more specific, one more problem that can arise due to misinterpreting data input is the relatively low accuracy of predictive models built on its basis. For example, in Paper 2 effectiveness in player performance is described as unpredictable due to inconsistent match conditions and player injuries. Likewise, Paper 1 recognizes some difficulties in normalizing sensor data originating from IoT devices – a factor that restricts the reproducibility of the results under different training paradigms. Furthermore, Paper 3 also demonstrated that there is no actual time data integration in the current models used during the match. The use of live match data can give more exciting and accurate predictions, especially on match-changing events such as batting order changes, and bowling line-ups.

### 2.3.5 Cricket Data Analytics in the Future:

The future of Cricket Data Analytics can be envisioned as the combinatory and ongoing nature of real-time data, wearable technology, and enhanced machine learning algorithms. As Paper 1 reveals, it is possible to consider the use of IoT-based real-time data collection for performance feedback at the instant stage of a solo performance of music. It seems that the utilization of computer vision and artificial intelligence in analyzing players' movements and actions during a match will become a mandatory component of a team strategy.

Likewise, Paper 4 outlines how the exact statistical modeling approach in Paper 3 can be further developed in the future – the introduction of real-time player performance data during matches is also used to enhance forecast predictability. The application of real-time decision support could help teams make better decisions instantly, let alone changing batting line-ups or substituting an injured player in a match.

### 2.3.6 Conclusion:

Existing literature on data analytics in cricket points out that there is a fast-growing interest in applying the latest machine learning and statistical methods for enhanced players and team performance, selection of players, and match prediction. The combination of IoT devices, machine learning, and live data in cricket has already started changing classical approaches to decision-making with a focus on an individual player and a team. Some of the issues including data quality, real-time integration, and model accuracy still remain. Still, the future does seem brighter for data-driven cricket, which promises to deliver a radical change to the sport when it comes to the issues concerning players' development, coaching, and even match strategies.

## 2.4 Methodologies in Cricket Data Analysis:

| Methodology | Description | Key Methods | Applications | Limitations |
|---|---|---|---|---|
| Statistical Analysis | Refers to a scenario where statistics is used to make sense of results to come up with trends or patterns. | Descriptive Statistics, Regression Analysis, Hypothesis Testing, Analysis of Variance. | Descriptive sports statistics, phenomenology, numerical relationships of participants | As such, does not also capture temporal variations at scale or consider multiple interactions at once. |
| ML Algorithms | Utilizes data to train models that look for likely scenarios or categorize data into some predetermined categories where rules are not available. | Decision Trees with Random Forest, Support Vector Machines (SVM), Neural Networks, K Nearest Neighbors (K-NN), Logistic Regression | On the probability of a particular match, a player's or team's performance or selection. | A large amount of labeled data is needed which is both time consuming and resource intensive. |
| Predictive Modeling | Used in an attempt to predict future occurrences that for example, match the results or performance of a particular player. | Logistic Regression, SVM, Random Forest, and Ensemble methods. | Match impacts, players and team impacts | Models are still deterministic and don't incorporate certain factors such as player morale into calculations. |
| Data Mining | Discovered findings that entailed analysis | Association Rule Mining, Clustering | The main factors that build up the identification of | Sometimes extracted patterns cannot |

| | | | | |
|---|---|---|---|---|
| | of deep patterns or coherencies within massive data sets. | Computer Classification | trends, performance patterns, and match conditions | be used to take straightforward action |
| Optimization Techniques | Designed to locate optimal solutions given certain assumptions or conditions, for example about the team lineup or game tactics. | Linear Program, Integer Linear Programming, Genetic Search, Monte Carlo Simulation | Team selection, Batsman, Bowler, the placement of the bowler | They can demand much computational capacity, particularly in real-time. |
| Visualization of Data | It offers graphic interfaces to represent findings in formats and forms that are easy to understand by clients. | Power BI, Heat Maps, Fielding Plots, GUI | number of players, team performance measurement, match review | The quality of the visualizations provided depends on the quality and level of detail of the collected data. |
| Natural Language Processing | Utilizes text data and distills comment or report data to arrive at a conclusion about the sentiment or morale among the team. | Key methods, and tools used: Sentiment Analysis, Topic Modeling | Evaluating people's perception, media control, and psychological aspect | Sometimes NLP models may fail to understand the context or sarcasm. |
| IoT and sensor-based analytics | Relies on IoT devices to | IoT Sensors, Wearable | Real-time performance | Data assimilation or |

| | | | | |
|---|---|---|---|---|
| | monitor players' performance by determining some parameters in real time. | Devices (Heart rate, Fatigue, Movement). | monitoring, players' fitness, and preventing cases of injuries | sensor calibration can be problematic. |

## 2.5 Research Gap

Although much has been achieved in cricket-related data analysis, several research gaps still exist to limit the potential of data-driven cricket decision-making. These gaps are mainly a result of issues reflecting data quality, model accuracy, and the incorporation of real-time data into the strategies of the teams. The existing literature highlights key areas for further exploration:

### 2.5.1 Real-time Data Integration:

Even though most of the current research includes the investigation of historical performance data and match analysis, data integration in real-time lacks sufficient investigation. Wickremasinghe (2014) and Paper 4 demonstrate that relying on past performance to make predictions for the matches and the players involved proves inefficient. Player's measurements (from wearables, IoT sensors, and computer vision systems) could be obtained in real-time also, providing the coaches with some crucial information during the game, especially when changing batting orders, field settings, or bowler's sequence of actions depending on the specific dynamics of the match. Subsequent studies can be directed towards how the application of such type of real-time data can be integrated into models that are used in live matches ensuring that they are responsive and useful.

### 2.5.2 Data Quality and Consistency:

One common problem highlighted by both primary and review studies is data quality and collection. As explained in Paper 2 and in Paper 1, incomplete or otherwise erroneous statistics, sometimes caused by inconsistent conditions of the match, injuries, or lack of data, greatly diminish the efficiency of the given models. This gap means that triathletes need to engage in improved means of data collection and develop better methods of standardization. For instance, the development of post-training data collection procedures in IoT sensors may enhance the information's normality to increase resilience in performance models.

### 2.5.3 Dynamic match situations prediction models:

Although there are many works done on the aspects of player performance and formation for teams, only a limited number of studies have been reported on real-time prediction of match situations. Previous studies concentrate on pre-match or historical characteristics of the team or player (Paper 3, Paper 4) but little is done to consider performance at the condition of certain match alterations like changes in weather or injuries and pitch conditions. Studying this aspect limelight could be paid to how models can be updated automatically during a match through events like changes in weather, a team/player form, or match tempo. Such real-time changes could be vital in case of making decisions during the match when watching the live stream.

### 2.5.4 Skills that Extended Stats Failed to Measure:

Many contemporary models employ basic player statistics including batting average, strike rate as well as economy rate. However, these statistics may not be sensitive enough to capture how different players may perform in different matches. For instance, Paper 1 presents a timing index, which uses high-level parameters including the bat speed and impact bat speed, parameters that conventional statistics fail to consider. However, this method is still in its infancy, so there is much discussion about these sophisticated statistics and research’s necessity to create new ones appropriate for testing various formats (Test, ODI, T20). Instead, future research may seek to find out how the existing and the new indexes deemed to measure player performance may be integrated to come up with a more efficient and dynamic method of measuring player performance.

### 2.5.5 Performance Model Customization:

Current approaches to predicting performance share another limitation – the similarity of players is taken as given, so the model may have the performance metric of one player in mind but apply it to any player. However, individual factors like a player’s batting style, fitness level, and some mental aspects also determine his play. Paper 3 briefly mentions the idea of applying K-means clustering for ranking players according to historical records, although there is more work that can be done on personalization not only on how each of the factors have been considered but also on the models developed. The stake could also be provided concerning the possibility of performing a study on how machine learning algorithms can be employed to customize performance prediction for each player based on the specialist roles that they play in a team.

### 2.5.6 Insufficient of a Comprehensive, Multi-Factor Model:

Although individual factors on team selection or team performance (various key factors that include – -handedness, home team advantage, or form) have been tackled in various research studies, there is no compendious study using a multi-factorial approach. For instance, unlike Paper 2 which compares the correlation between the team rank and player performance, and Paper 4 which deals with the relation between player fitness and last-five performances, the evaluation of such factors tends to be done separately. The next step could be to use views that include a lot of variables connected with matches, the form of the players, location, team morale, and other factors that combine to make an optimal playing environment for players, their confidence, and other psychological factors that could influence the game. Together these factors might yield better team formation solutions and match predictions.

### 2.5.7 Cross-format and Cross-condition Performance Models:

Most of the previous research articles are based on Test cricket, while some of them are based on One Day Internationals and T20 cricket. There is no systematic work done for decentralizing player performance data across formats and different conditions for matches. Paper 4 employs a method using career statistics to forecast performances, but it is rare to find a study comparing how players' performances in one format of the game, ODI/T20, determine their aptitude to play in the other format, Test or vice-versa. In the same way, research could explore how a player behaves under various conditions such as different kinds of pitch, weather, or while playing in different countries/ environments, and how these conditionalities affect the performance.

As apparent from the foregoing discussion, there has been considerable progress in the application of data science and machine learning in cricket analytics; however, there are still gaps in the current literature. Among these, the real-time integration of data, data quality improvement, dynamic match prediction, and the construction of comprehensive multiple-factor models are the largest potential for further research. Such gaps must be closed to enhance the models under consideration, and provide a better platform for selection, performance predicting, match strategies, and all else in between in specified teams, which will make more effective decisions in issues to do with cricket.

## 2.6 Key Findings:

### 2.6.1 Higher Odds in Accurate Predictions of Match Results

The use of machine learning models and statistical techniques revealed another number one trend among crickets finding related to enhanced predictive accuracy of match outcomes. Through research on match data, player records, attributes of the pitch, weather conditions, and other social factors experts are now in a position to forecast match outcomes.

**Key Contribution:** Gupta & Patel (2021) found that using machine learning methods like decision trees and support vector machines (SVM), the … These models enhanced predictions for the chance of a team winning by including player’s performance data and outside match factors.

**Key Finding:** The application of ensemble methods and deep learning models has provided good results in increasing the reliability of a forecast, this is proved by high-dimensional features, such as player form, an opponent’s strengths, and weather conditions, in Sarkar et al. (2020).

### 2.6.2 Information on individual player performance and an ability to determine the best line-up.

Data science has/is changing the ways in which players are rated and selected into different teams. Machine learning algorithms have been used for predicting an individual player's performance under certain match circumstances which has been useful to the coaches and selectors.

**Key Contribution:** In the player selection of the teams, the K-means clustering algorithm is being used, which is also shown in the project called “Cricket Team Selection and Player Analysis using Data Analytics”, and this algorithm will group the players based on their player performance indicators, like the strike rate and economy of the batsmen and bowlers and will create balanced teams. This approach is more flexible than conventional methods in that it offers selection probabilities for teams.

**Key Finding:** In the paper by Wickremasinghe (2014) the creation of a hierarchical linear model (HLM) to predict the Performance of batsmen in Test cricket gave better insight into the way how player ability, team status, and match characteristics affect performance. The model also suggested the interaction between handedness (left-handed or right-handed) and performance.

### 2.6.3 Enhanced Strategy Execution Over Games

Batting line-up, bowlers’ schedule, and field arrangements: all options have been subjected to different analyses and become real-time decision-makers. Game theory has been applied in choosing the best strategy in the case of different possible matches.

**Key Contribution**: In making their strategies, Gupta et al. (2021) conducted predictive analytics to find out the best approach depending on the match situation; whether they are in pursuit of a target or when they must protect a total. Particularly, this work captures aspects of how real-time analytics can inform tactical decisions during a match by a coach.

**Key Finding:** Known approaches to applying genetic algorithms as well as reinforcement learning for RTS optimization have been demonstrated to provide benefits, enabling teams to react promptly depending on match data and player statistics.

### 2.6.4 An Analysis of Additional Measures of Batting and Bowling

Performance Apart from Batting Average and Bowler's Average, Of course, there is an inherent power of simple counting, but traditional cricket statistics (batting average, Wickets taken per over bowled [bw]) are not sufficient to represent the versatility of the performance. Players and coaches have sought high-profile statistics and another method of assessing the ability of players.

**Key Contribution:** The usage of the timing index, discussed in the research of Sharma et al. (2020), appealing and effective IoT-based cricket bat sensors to analyze attributes such as bat speed, back lift angle, and impact bat speed to establish the effectiveness of a batsman's shot-playing abilities. This approach goes beyond statistics by not only analyzing defensive performance's mechanical aspects but also giving a detailed measure of batting performance.

**Key Finding:** The usage of IoT and sensor-based analytics has shed new light on how a batter approaches his performance and his current form by tracking his movements. From the study conducted using the timing index, the comparison of player movements using modern tools enables trainers and coaches to have the following:

### 2.6.5 Real-Time Data Collection and Its Effect On Performance Analysis

Live data capture has consequently assumed enormous significance in the sphere of cricket with time, because of the availability of timely data concerning the strategies to be adopted during a game. Since the introduction of wearable devices, IoT sensors, and video analysis, capturing and using performance data have become a whole new ball game.

**Key Contribution:** Advancements in the design of cricket bats and the incorporation of wearable sensors have improved post-match and training performance measures including, bat speed, ball impact, and player movements during games and practice sessions. It enables coaches to evaluate the efficiency of players and make a change quickly and conveniently.

**Key Finding:** Through Hawk-Eye and Pitch Vision technologies, analysis of performance has been enhanced since factors such as ball trajectory, pitch, and shot accuracy have been brought into focus to make successful changes in close to real-time fashion on the batting and bowling strategies.

### 2.6.6 How Data Can Be Utilized for Bettering Injuries Avoidance and Players ‘Health

Gone are the days when data analysis was solely restricted to the performance enhancement of players; it is about the welfare of a player too. With the help of IoT and wearable technologies, data scientists were able to monitor fitness levels and recognize the possibility of an injury beforehand.

**Key Contribution:** The novel technologies to track the players include the fitness tracker and motion sensors whose data helps to assess player fatigue, workload, and biomechanics to predict injuries and set out proper training load. This aids in maintaining the health of players as in instances of long tournaments or series it becomes manageable.

**Key Finding:** The analysts have pointed out that tracking player workload, in this case through statistics, has made it possible to temper with injuries because players are not pushed to the extreme in their performance.

### 2.6.7 Better Fan Interaction Using Data

As the cricket ecosystem has become digitalized data is used to better engage fans and for them to interact. Technology has ended up enabling fans Big Data opportunities that would help them gain deeper knowledge of matches, players, and every team.

**Key Contribution:** There is an increase in the use of well-developed data analysis tools, which are used on athletic fields and tracks, as well as specially designed, portable applications for mobile devices that enable fans to receive updated statistics during games, key players' information, and possible scenarios. These platforms utilize data tools such as Power BI to dissect and present data in such a way that can easily be understood.

**Key Finding:** Results indicate that fan engagement has been greatly improved by data science methods such as predictive modeling of match results, as well as content recommendation based on fans' preferences. That is why the event experience has become more engaging and engaging during concerts and other performances.

### 2.6.8 Connecting the Modern and Traditional Analysis

However much the manager implements data and new technology, expert experience, and human feelings are still major in cricket. The problem is then how to combine metric analysis with experience.

**Key Contribution:** The use of ordinary statistical measures together with the application of probability theory in the models has enabled the analysts to come up with better decisions as they work around the structures of the game. This way the analysts can give an analysis that incorporates qualitative information such as morale among players, weather, and quantitative analysis such as performance, match results, and others.

**Key Finding:** One common theme that runs through cricket data analysis for success is the fact that, in most cases, human intervention works hand in hand with the mechanical analysis of data. Such collaboration helps to avoid the effect that several predictions depend on the numbers or strategies only while excluding the specifics and the dynamics of cricket as a sports type.

## 2.7 Conclusion:

The current state of research in cricket data analysis revolves around three main areas: sports performance prediction, selection of the best players, and decision-making.

**Performance Prediction:** Data mining algorithms, statistical tools, and decision trees are used to perform player performance prediction by considering past performance, player characteristics, and game situations. Some research studies have demonstrated the effectiveness of data models based on more accurate performance estimates of players in terms of strategy and techniques as compared with traditional approaches, for instance, Wickremasinghe, 2014), and kindred areas including the effectiveness of batting and bowling.

**Team Selection Optimization:** Advanced DIS, or data and information science, has led to improvements in how specific team elements can be composed using methods such as K-means cluster analysis and hierarchical linear and extrapolated predictive models on performance criteria. These models enable coaches and selectors to use efficient team selection strategies under various conditions during a match, especially during any dynamic format like T20 cricket (e.g., Gupta & Patel, 2021).

**Strategic Decision-Making:** Decisions in match conditions, including batting lineup, bowling attacking strategy, and field setting, have been enhanced by real-time statistics. It is possible to adapt strategy during a game that likely occurs at rest periods using predictive modeling: match conditions, player and opponent analysis (e.g., Gupta et al., 2021). Extension of real-time sensors IoT plays an extra added advantage in decision-making, especially in player training and technicality.

These themes throw the growing importance of data science in cricket into the limelight. But even now, there are some problems: lack of data consistency, problems with the integration of an analyst's opinion into machine learning predictions, and the absence of more complex models that would consider such factors as the morale of the players or the psychological state of the teams before the match.

**References:**

Wickramasinghe, I. P. (2014). *Predicting the performance of batsmen in Test cricket*. Eastern New Mexico University. https://rua.ua.es/dspace/handle/10045/45872

Agarwal, S., Yadav, L., & Mehta, S. (2017). Cricket team prediction with Hadoop: Statistical modeling approach. *Information Technology and Quantitative Management (ITQM 2017)*. https://www.sciencedirect.com/science/article/pii/S1877050917326479

Vishwarupe, V., Bedekar, M., Joshi, P. M., Pande, M., Pawar, V., & Shingote, P. (2022). *Data analytics in the game of cricket: A novel paradigm.* https://www.sciencedirect.com/science/article/pii/S1877050922008523

Raajesh, S., Martin, N., Jiji, J., Nair, A., & Haritha, H. (Year). *Cricket team selection and player analysis using data analytics.* https://ieeexplore.ieee.org/abstract/document/10689923