

CRICKET DATA ANALYSIS

¹MOHAMMED JUNAID, ² ROOPA R

Student, Department of MCA, BIET, Davangere

Assistant Professor Department of MCA, BIET,
Davangere

ABSTRACT:

As the second most popular sport in the world, cricket has gained popularity as a team sport on a global scale (Pathak and Wadhwa, 2016). There is a huge need for cricket data analytics due to the abundance of available cricket data and the advancement of machine learning (ML) technology. Over the past 20 years, there has been a significant rise in the use of ML in the cricket arena. This paper performs a thorough analysis of the research that has been published over the previous 20 years (2001–2021) on the use of machine learning in cricket.

1. INTRODUCTION:

According to Kampakis and Thomas (2015), cricket, a bat-and-ball sport that is played in 106 countries that are members of the International Cricket Council (ICC), is now a multi-billion dollar industry. There are three primary types: Twenty-Twenty Cricket, Test Cricket, and One-Day International Cricket (ODI). Due to the dynamic nature of both the ODI and T20 formats, cricket has quickly become a popular team sport (Bandulasiri, Brown, & Wickramasinghe, 2016). With this growing popularity, franchise T20 leagues have emerged in a number of international regions.

An essential component of any successful cricket team is now data analysis. According to Morgulev, Azar, & Lidor (2018), the findings of cricket analytics provide a better understanding of the players as well as the game, which is extremely beneficial to those involved in the sport, such as current players, technical staff, and managers, as well as to the education of future players. Sarlis and

Tjortjis, 2020). Administrators frequently look for novel ways to boost the performance of cricketers and give them a competitive edge because of the game's rapid development. The competitive sports environment is characterized by intense and stressful situations.

The obligation of the entertainers in the game is to deal with the strain and mental reactions to realize their true capacity (Devonport, 2015). The goal of player-performance management is to maximize player performance while reducing the likelihood of injury in various sports (Naglah et al., 2018; Xu and Tang, 201). Cricket data analysis plays a crucial role in this procedure.

Due to numerous limitations, sports data analytics cannot rely solely on conventional statistical methods. Conventional statistical methods typically rely on presumptions regarding the data, and cricket data may not satisfy these prerequisites. Besides, occasions connected with cricket are not free occasions and are impacted by various human

elements; As a result, choosing the right data analysis methods is critical (Horvat & Job, 2019; 2003, Karlis and Ntzoufras). Due to the large number of related variables, most traditional metrics fail to address game hypotheses today.

In spite of the limitations exhibited by conventional statistical data analysis methods, the development of machine learning (ML) has had a significant impact on sports data analytics.

Machine Learning (ML) is a subfield of artificial intelligence that uses a collection of computer algorithms to enable systems to automatically learn from experience and advance. The primary goal of ML techniques is to automate the knowledge engineering process to a higher level by taking the place of labour-intensive human activities. The wealth of computational influence and information has given gigantic prominence to ML strategies utilized in Sports information examination.

Over the past two decades, the ML process and its applications have developed to a very high level of maturity. Additionally, the growth of machine learning techniques and the abundance of sport-related data have elevated sports data analytics to new heights. ML techniques, in contrast to conventional computer

systems, permit these systems to learn from data without being explicitly programmed or imposed with rules (Grues, 2015). With the influence of ML, the introduction of electronic devices for the collection of sports data has demonstrated a significant boost in sports data analytics. Due to their influence on the devices used to collect data (Jamil, Iqbal, Ahmad, & Kim, 2020), ML techniques' power has increased (Morris, Mundt, Goldacre, & Jacqueline, 2020; Weir, Alderson, Smailes, Elliott, and Donnelly, 2019), and the handling the data accumulated through the gadgets to improve the comprehension of a definitive clients (Ahmadiet al., 2014; Rommers and other, 2020). The greater part of these electronic gadgets come as wearable gadgets. According to Acikmese, Ustundag, & Golubovic (2017), the development of AI expert systems for use in sports has been made possible by combining these wearables and ML algorithms. These ML-based applications quickly learned how to use sports data analytics to build models and make predictions about future outcomes based on existing sports data. This allows users to make more informed sports decisions. Applications of machine learning quickly learned how to use sports data analytics to build models and make predictions about

future outcomes using data already available to sports. These predictions are accurate, allowing better decisions to be made in sports. The creation of these ML systems necessitates domain-specific expertise. To enhance ML calculations, it utilizes a cycle called include designing. This is done with the intention of optimizing the ML algorithms by extracting features from the raw data with the assistance of domain expertise. Even though there are a lot of sports data, applying machine learning to big data is interesting because it requires expert knowledge of the subject, the learning algorithms used, and software engineering (Koseler & Stephan, 2017).

Orderly surveys assist with gathering experimental proof about the research's consistent development involving ML in the cricket space. Additionally, the author is aware of no such systematic review of cricket and ML-related two decades of research findings. As a result, the purpose of this study is to fill the aforementioned gap in the literature, and the results will be beneficial to players, coaches, and sports administrators. Lastly, this effort will assist researchers in identifying research gaps and providing a concise overview of cricket's current research areas.

2. LITERATURE REVIEW

B. Padmaja, Y. Mohana Roopa, and P. Sri Harsha Vardhan Goud, "Player Performance Analysis in Sports: examines the investigation job of machine learning in the improvement of exhibitions of players and the group in various games and how wearable innovation assists the players with realizing their presentation levels and further enhancements. A system that processes raw data into statistical data for each sport, team, and player is described in the paper [1]

Amal Kaluarachchi, Aparna S. Varde, "CricAI: A Classification Based Tool to Predict the Outcome in ODI Cricket" Id and CricAI Toll discussed it in. It can assist in adjusting certain variables to increase your chances of winning the actual game. Predicting a team's chances of winning a One-Day International cricket match was the topic of the paper [2]

Tavana, M., Azizi, F., Azizi, F., Behzadian, M.: a system for fuzzy inference that can be used to select players and form teams in sports with multiple players. A two-phase framework for team formation and player selection was proposed [3]

Dr. Muhaimenur Rahman and others Using machine learning, they analyzed Bangladesh ODI cricket data and discovered the predicted outcome, highlighting the significance of some features.[4] Passi and others applied some AI calculations. In order to assess a player's performance, they established some parameters and utilized some equations to generate ratings [5]

Md. based on a player's performance Minhazul Abedin and others looks at Random forest, K- nearestneighbors (KNN), Support Vector Machine (SVM), and Decision Tree supervised classification models. [6]

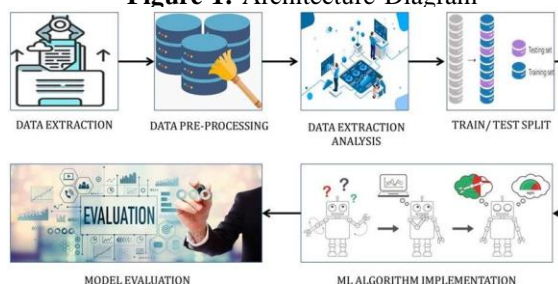
Madan Gopal and others investigate different machine classifiers like SVM Random Forests, Calculated Relapse, Choice Trees and KNN. [7] Shah also defined new performance metrics for players. Both the new measure for batsmen and the new measure for bowlers take into account the quality of each batsman they are playing against. [8] Mukharjee. utilized Social Network Analysis to evaluate the team's bowlers and batsmen [9].

This combined bowling rate was used by Bhattacharjee and Pahinkar to examine how bowlers performed in the Indian Premier League(IPL). [10]

3. METHODOLOGY

The application of machine learning techniques to cricket in contemporary literature was the focus of this manuscript's systematic investigation. Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA; standardized guidelines) The guidelines for systematic reviews (Moher, Liberati, Tetzlaff, Altman, & Prisma Group*, 2009) were followed. The whole survey convention includes three stages: identification, selection, and screening the review protocol's three phases are outlined.

Figure 1: Architecture Diagram



In the proposed system 3 types of machine learning models will be used. A classification model will be trained to predict the winner of the match. The proposed system will use Lasso regression for predicting the score of a match.K means clustering algorithm will be used to cluster players.

3.1. Data collection

First, we distinguished all distributed work utilizing four significant data sets, specifically, Google Researcher, Science Direct, Scopus, Web of Science. In February 2022, this search using broader research domains was carried out. The keywords "Applications of Machine Learning in Cricket" were used in the online search. All distributed exploration work fulfilling the above looking through data was to begin with recognized.

3.2. Pre-Processing

Also, we chose the most pertinent distributions in view of the following standards.

- Cricket-related studies, including but not limited to player performance, game outcome, team performance, score prediction, pitch-related studies, cricket commentary studies, and cricket video studies, were taken into consideration.
- The robot or electronic versions of cricket are thrown out because only the human game is taken into consideration.
- Studies using no less than one ML method were chosen.
- From 2001 to 2021, studies published in a journal with peer review were taken into consideration.
- Publications from conferences and journals were included.

3.3. ML. Model development

This section looks at the ML techniques that are used a lot in cricket .As per the looked into studies, the accompanying ML procedures have been utilized as the well known ML methods to investigate cricket information.

3.4. User Interface Design

The performance classification of the player: Batting and bowling statistics are the most commonly used to measure players' performance, despite the fact that the three main departments of the game are fielding, batting,

and bowling (Wickramasinghe, 2014b).

3.5 Data Set Used

In the ML workflow, text preprocessing is an important step. It basically involves cleaning the data so that only the relevant information can be extracted. Checking for duplicate or irrelevant data is the first step in cleaning data. Over various time periods, we gathered data by employing various hashtags. It can frequently prompt excess tweets by something similar or various clients having different normal hashtags. We killed trickery in information assortment stage itself. The subsequent steps for textdata preprocessing are as follows:

Table-1 Data Set

Descriptive statistics of the used data set.

Format	Mean	SD	Min	Q1	Q2	Q3	Max
First Class	861.8	774.3	44	281.0	909.5	1490.2	1584
ODI	27239.6	93191.0	128	385.5	1301.5	5603.8	350899
T20	537502.0	1400273.0	140	994.0	16720.0	123514.0	4000000
Test	1666.0	1856.2	354	1010.0	1666.0	2323.0	2979

4. RESULTS

The purpose of this section is to present and talk about the results of this systematic review. First, a discussion is held to determine the cricket research areas in which ML can be applied. The reviewed publications are then described with some descriptive statistics.

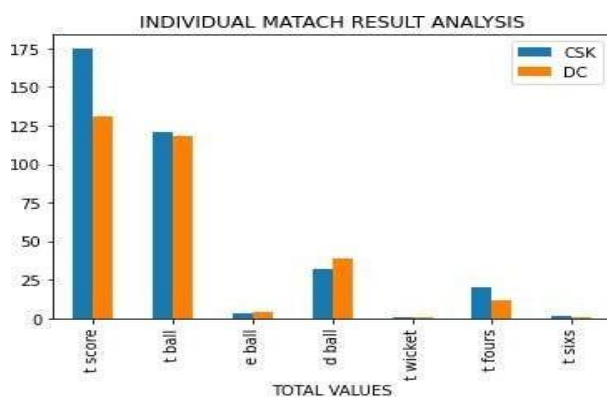


Chart -1 Individual Match Result Analysis

Second, all of the selected articles are reviewed in the research question section to determine which studies have

addressed the research questions regarding which feature extraction, ML, and accuracy estimating techniques were utilized. The data sets and the number of attributes used in each study are examined in the following subsection. Then, discoveries of the pre-owned information decrease and element extraction strategies are introduced. Following that, a discussion of the most frequently employed ML methods is provided. The final section aims to quantify the ML techniques' accuracy.

4.1 Cricket study areas:

Cricket and machine learning (ML) studies have grown exponentially since 2001, according to findings. Compared to the other formats, the ODI game has received significantly more attention in these publications. This graph shows the number of ML-based studies on each cricket format from 2001 to 2021.

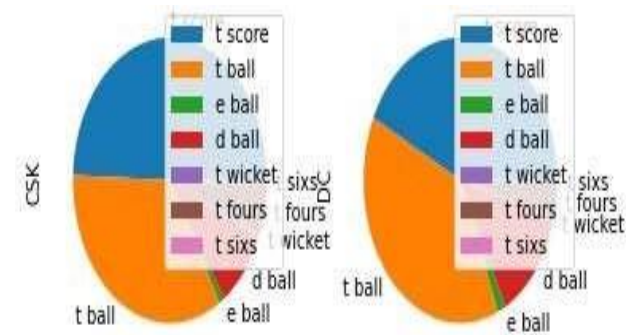


Chart-2 Cricket study areas:

Further research reveals that about 35% of all cricket-related studies have focused on predicting the outcome of games.

Around 16% of the examinations have been directed investigations in player performance grouping. The researched areas are described in greater detail below. The reviewed articles' contributions to each section are summarized in the sections that follow.

4.2 Prediction of game outcome

The expanded ubiquity and the commercialization of the game, result forecast of a cricket game has happened to the greatest amount of significance. To predict the game's outcome, researchers employ a wide range of ML

methods and a variety of performance indicators that represent various game elements. attributes of fielders, bowlers, and batsmen (Hasanika, Dilhara, Liyanage, Bandaranayake, & Deegalla, 2021; Karthik and other, 2021; Modani, Kilaru, Kaur, Sinha, & Khetan, 2020) are the performance measures that are utilized the most frequently in the prediction. What's more, different properties like home game advantage (Kaluarachchi and Aparna, 2010; The prediction was based on the outcome of the toss (Pathak & Wadhwa, 2016; Kumar, Santhadevi, & Barnabas, 2019), the pitch's behavior (Tekade, Markad, Amage, & Natekar, 2020), and the outcome of the toss. More work in result forecast should be visible in Basit et al. (Shakil, Abdullah, Momen, and Mohammed (2020), Hatharasinghe and Poravi (2019), and Vistro, Rasheed, and David (2020)

Batting and bowling statistics are the most commonly used to measure players' performance, despite the fact that fielding, batting, and bowling are considered the three central departments of the game (Wickramasinghe, 2014b). Individual players' abilities have a decent impact on the game's result. Therefore, the game relies heavily on individual skill assessments in numerous ways; This can be accomplished by combining player-level parameters and machine learning techniques (Wickramasinghe, 2020a). Classifying players into various skill levels is a common practice (Aburas, Mehtab, & Mehtab, 2018; 2020; Manage, Kafle, and Wijekularathna Wickramasinghe, 2020b), assessing or positioning them (Ahmad et al., 2021; Anik, Yeaser, Hossain, & Chakrabarty, 2018; Premkumar, Chakrabarty, & Chowdhury, 2020) identifying or predicting the level of player performance Mody, Malathi, and Jayaseeli, 2021; Rupai, Mukta, and Islam, 2020), and determining which batsman or bowler will be the best (Rani et al., 2020) or the up-and-coming talent (Ahmad et al., 2017). A strong team can be formed with the assistance of identifying the best performers because the performance of individual players directly affects the team's rank (Wickramasinghe, 2014a). In Ishi and Patil (2020), Mahbub, Miah, Islam, Sorna, Hossain, and Biswas, and Mahbub, Miah, and Biswas, ML methods are used to select the best team.

- Classification of the bowling action and bat style: As a batsman's game, cricket necessitates the ability to choose and execute shots. In order to identify and evaluate the batsmen's shot-making abilities, ML methods and body-worn internal measurement units (IMUs) were utilized (Dias, Mitchell, & Harland, 2020). In addition, the bowler's characteristics, such as bowling action (Salman, Qaisar, & Qamar, 2017), bowling volume, ball releasing speed, and the identification of intensity zones, can be studied using ML and wearable devices (McGrath, Neville, Stewart, Clinning, & Cronin, 2021; 2021; McGrath, Neville, Stewart, Clinning, Thomas, and Cronin; 2019 by McGrath, Neville, Stewart, and Cronin; Silva and Ranaweera, 2019). In addition, exact assessment of the blower's responsibility can be checked and anticipated with the utilization of ML, and it will, thus, limit the conceivable future player wounds (Jowitt, Durussel, Brandon, and Lord, 2020).

4.3 The features used and the cricket data:

The quality and the accuracy of the study directly depend on using a representative sample. In addition, bias in the inference can be minimized using a larger sample. Furthermore, selecting a good representation of features also directly impacts the quality of the outcome of the ML-based data analysis. Therefore, selecting a good data set is foremost essential. This section summarizes the sizes of data sets used in the selected studies. According to the reviewed studies, the average number of instances of data used in research in the T20 format is 537,502, which is the highest compared to the other formats of cricket. In ODI-related studies, approximately an average of 27,240 instances of records have been used. For the Test and First-Class cricket, these are approximately 1666 and 862, respectively.

Because it has a direct impact on the accuracy of the ML technique, feature selection is an essential component of its use. Cricket data is now easily accessible as a result of the game's popularity and technological advancements. In addition, the cricket game's feature set grew in tandem. The findings indicate that, in comparison to other game formats, studies using First-Class games have used an average of 144 features. Second, the average number of features used in the Test cricket study was 115, while the number of features used in the ODI and T-20 cricket studies was 40 and 12, respectively. represents the dissemination of the

quantity of elements utilized in every one of the configurations of the game.

4.5 Utilized methods for feature extraction and data reduction:

Highlight choice is a basic move toward ML. According to Guyon & Elisseeff (2003), the goal of the feature selection method is to select a subset of the input data's variables in order to effectively describe the data, minimize effects from noise or irrelevant variables, and still produce accurate predictions. The following methods were able to identify the primary feature selection method based on the reviewed studies.

- Chi-Square
- Relationship Based Element Determination
- Illustrations Based Element Determination
- Head Part Examination
- Recursive Element Disposal
- Relapse

Among the above list, Relationship based highlight determination techniques, Recursive End, and the Chi-Square procedures were the most often utilized

4.4 Frequently used ML techniques:

This section looks at the ML techniques that are used a lot in cricket.

As per the surveyed investigations, the accompanying ML strategies have been utilized as the wellknown ML procedures to dissect cricket information.

K-Means, Random Forest, Decision Trees, kth Nearest Neighbor (kNN), Artificial Neural Networks (ANN), Support Vector Machine (SVM), and XGBoost are all examples of regression.

Predicting a game's duration Player performance modeling is the most frequently used of the ML techniques listed above in the reviewed papers. SVM is used 45 percent of the time. RF, which is used by 42% of people, came in second place, followed by NB, which came in third, with 36% of people using it. SVM was used in 45 of the reviewed studies, and 26% and 24% of them were published in 2019 and 2021,

respectively. RF is the second- most frequently used ML method. RF has been used in 42% of published research, and 29% have been used in 2020 and 2021. NB was the third most widely used ML technique in published works between 2010 and 2021. NB has been used in 36% of all published works. Of them, 26% of the work was distributed in 2019, and 22% were utilized in 2020. More information, including some of the above-mentioned research areas, ML techniques According to the findings, the number of studies utilizing ML techniques with cricket data has increased in recent years. Confusion matrix-based ML accuracy technique 57.7 F-score 15.5 RMSE

6.2 ROC 9.3 Other methods 3.1 Cohen's kappa statistic 2.1 MAE 2.1 MCC 2.1 R2 -statistic 2.1 two decades Fig. 5 shows the dispersion of distributions utilizing the above-expressed ML procedures from 2001 to 2021.

4.6 Quantifying the ML technique's accuracy:

It is essential to evaluate the accuracy of the ML model because all ML models are driven by data. This may be the most significant component of ML models. According to the reviewed studies, confusion matrix-based accuracy measures are the most common (around 58%) method for evaluating the ML model's accuracy. F-Score, the Receiver Operating Characteristic (ROC) curve (approximately 9%), and the Root Mean Square Error (RMSE) (approximately 6%) are the remaining popular methods. The confusion matrix- based technique is the primary method for accuracy testing. Some confusion matrix-based methods that researchers have utilized include accuracy, balance accuracy, custom accuracy, precision, recall, sensitivity, and specificity.

4.7 Result And Snapshots



Image 1 : Main Page

Image 2: User Option Selection Page

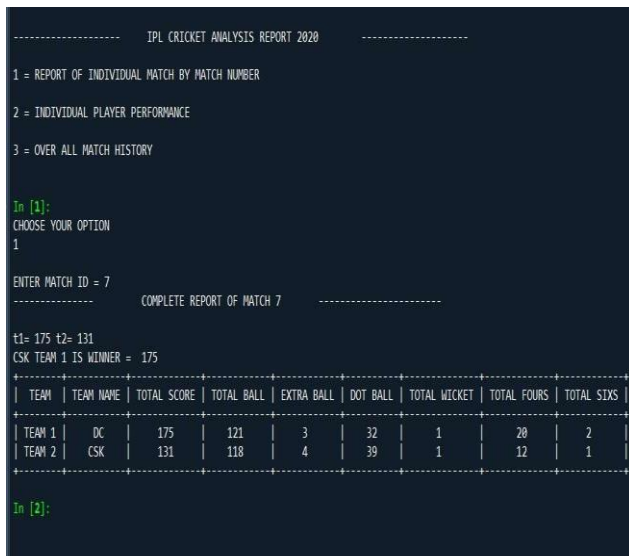


Image 3: Individual Player Selection

5. Discussion and conclusion

In view of the directed efficient survey, it is obvious that the volume of examination in cricket utilizing ML innovation has been expanding beginning around 2001. This inclination is a positive sign for the game, as the discoveries of these examination results will ultimately help the game's development. The investigation of various aspects of cricket that were previously impossible to carry out using the methods that were in place is now possible thanks to improved ML techniques. Additionally, the incorporation of ML methods and electronic wearable devices into cricket research has numerous advantages. By lowering the risk of injury, research findings enable cricketers to perform more quickly.

The results of this study show that the majority of studies on cricket players used the same set of batting and bowling characteristics.

The game evolved by adding new game-related features as technology came into use. Consequently in future exploration, the time has come to look for

these original elements while utilizing the as often as possible utilized traits of the game. In addition, the nature of the current game has changed as a result of the introduction of novel cricket rules like fielding restrictions, the number and type of balls used, and player substitution.

Accordingly, assuming we assemble cricket information that ranges a very long while, there is an issue with the similarity of the information, which can bring about conflicting results.

There are three sports in cricket: batting, bowling, and fielding, each of which is equally important to the outcome of the game. The reviewed articles indicate that not all three of the above departments have received the same number of studies. Even in studies on bowlers, fast bowlers have been the primary focus. The use of medium pacers and spinners in research is uncommon. When looking at studies about how well players perform, one thing that could affect their performance is the kind of cricket ball used in the game.

Cricket balls come in a variety of brands, including Dukes, Kookaburra, and SG. Unfortunately, there hasn't been a lot of research done in this area to investigate how the type of cricket ball affects player performance. Moreover, there is a critical number of studies connected with batting, however there are not really any concentrates on connected with defenders. However, despite the fact that fielding is crucial to the outcome of the game, there have been few studies conducted on the topic.

From a ML perspective, some reviewed studies haven't shown important details about the ML process. Some of them include the data's nature (size, number of attributes, correlation structure, and sparsity), the feature selection technique, the accuracy quantifying technique, and the level of accuracy the ML technique achieves.

The reviewed articles showed how various ML techniques were used in various cricket research areas. Because the success of the ML technique is dependent on a number of factors, it is difficult to recommend just one for a particular cricket application. According to these articles, KNN, regression, and RF have been used more frequently over the past two decades than any other ML technique.

While SVM continues to be popular in cricket applications, other ML techniques like neural networks

and Naive Bayes do not show increased popularity. One of the critical perceptions was the absence of the utilization of some famous ML procedures, like profound learning

support learning and normal language handling. In order to address cricket's more complex problems, it is anticipated that future research will make use of cutting-edge ML methods like automated machine learning, multi-modal learning, multi-objective models, and tiny ML.

This systematic review is constrained in a number of ways, as we point out. To begin with, the concentrate on thought to be just four data sets. Furthermore, this study was unable to take into account the T-10 variant of cricket due to limited resources. This review found a number of ML-based cricket research projects and other areas that need more attention from researchers. More research is needed in the future because cricket is a game that is constantly evolving. ML will make a significant contribution in this direction, which will ultimately aid in the game's further development.

References:

- [1] Abbas, K., and Haider, S. (2019). Comparison of the Duckworth-Lewis-Stern method and the machine learning approach. In 2019, an international conference on information technology frontiers (pp. 197–1975). IEEE.
- [2] Aburas, A. A., Mehtab, M., and Mehtab, Y. (2018). Cricket world cup expectations utilizing KNN smart bigdata approach. In the 2018 International Conference on Computing and Big Data's Proceedings (pp. 18–22).
- [3] Bandulasiri, T. Brown, and I. Wickramasinghe are the authors. description of the outcome of the one-day cricket format. 26(4), 21–32, Operation Research and Decisions.
- [4] Basit, M. B. Alvi, F. H. Jaskani, M. Alvi, K. H. Memon, and R. A. Shah Using machine learning, ICC T20 cricket world cup 2020 winner prediction In 2020 IEEE 23rd global multitopic gathering (pp.1–6). IEEE.
- [5] G. Deval, F. Hamid, and M. Goel are the authors. When should a test cricket match's third innings be declared? 303(1) of the Annals of Operations Research, pp. 81–99.
- [6] T. Devonport (2015) Understanding pressure and adapting among cutthroat competitors in sport. Psychology of Sport and Exercise, 127.
- [7] P. Dias, S. R. Mitchell, and A. R. Harland An

innovative experimental method for predicting shot execution in cricket batting using movement data. Proceedings of the Multidisciplinary Digital Publishing Institute, 49(1), 41.

[8] P. Sri Harsha Vardhan Goud, Y. Mohana Roopa, B. Padmaja,” Player Performance Analysis in Sports: with Fusion of Machine Learning and Wearable Technology” Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC

