Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)
IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

Data mining based Classification of Players in Game of Cricket

Balasundaram A
Assistant Professor (Sr. Grade),
School of Computer Science and Engineering,
Vellore Institute of Technology (VIT), Chennai, India
balasundaram2682@gmail.com

Ashokkumar S
Assistant Professor,
Department of CSE, Saveetha School of Engineering,
SIMATS,Chennai, India
sabariashok2016@gmail.com

Jayashree D
Assistant Professor,
Department of Information Technology,
S.A. Engineering College, Chennai, India
jayashreesagayam13@gmail.com

Magesh Kumar S
Associate Professor,
Department of CSE, Saveetha School of Engineering,
SIMATS,Chennai, India
mmce6450@gmail.com

***Abstract*— Cricket is an outdoor sport that is extremely popular in Asian countries. Generally, the game is played in three different formats. Each format throws its own set of challenges and teams that are chosen in accordance to the format demands. The outcome of a cricket match is highly unpredictable if the gap between the skill levels of opponent teams is very narrow. It should also be noted that, Cricket is a team game, it is not one player, who determines the course of outcome of the match. Hence selecting the right team for a match by assessing the available player performance and playing conditions has a significant bearing on the outcome of the match. This work is further focused towards developing a robust prediction model that can classify players and accordingly help in predicting the Cricket team especially in the ODI format.**

Keywords*— *Player classification; Cricket team prediction; decision tree; support vector machine; random forest

I. INTRODUCTION

There is a common saying that "Cricket is a game of glorious uncertainties". It is this highly unpredictable nature of the game that creates huge interest among the followers of this game. The nature of the game is such that the best team playing on that particular day emerge as winners. There are instances, where teams posting low totals end up on the winning side by dismissing the opposition for a lower score. At the same time, one cannot discount the possibility of a huge score being successfully chased by the opponent team as well. The nature of the game is such that the course of the match can be altered any ball by the fall of a wicket or a series of good scoring shots. These statements clearly underline the highly unpredictable nature of the cricket.

Several works have been carried out in the past to predict the various facets of the game which are discussed in detail in the subsequent sections. The work involves usage of Machine Learning (ML) algorithms such as Support Vector Machines (SVM), Random Forest and Decision Trees. Machine Learning algorithms find significance across multiple domains and allied technologies of Artificial Intelligence (AI) such as Computer Vision (CV) [25-31] and predictions and making decisions in cloud environment [20-24]. They are highly preferred choices for performing efficient classification across domains. This work involves developing a predictive model for Cricket using these classification and prediction algorithms and is illustrated.

II. RECENT WORKS

Several works have been carried out to understand the dynamics of ODI cricket format and to make a judicious prediction. This section highlights some of the recent works carried out towards cricketing predictions.

Subramanian Rama Iyer et al. [1] have used neural networks to analyze the player's performance and in turn their impact towards winning the game. The players were classified as performers, moderate and failures and based on that the team performance was predicted. Soomro et al. [2] have analyzed the player availability for matches and have leveraged mobile application technology for predicting player's injury and subsequently the availability for team selection. Muhammad Asif et al. [3] have proposed a highly dynamic 'in play' model that predicts the outcome of ODI games as and when the game is in progress. The model is built using logistic regression and takes into consideration various factors of the game as input to make accurate predictions.

Neeraj Pathak et al. [4] have performed an exhaustive study of how various machine learning and data mining techniques such as Naïve Bayes, Random Forest and Support Vector Machines can be used across cricket to make effective predictions. Hugh Norton et al. [5] have used a technique

called Monte Carlo simulation to estimate and determine the conditional probability for winning a ODI cricket match. Tom Allen et al. [6] have used finite element analysis to understand the impact of ball hitting the bat and accordingly predict the trajectory and speed with which the ball flies off the bat.

Sohail Akhtar et al. [7] have used a probabilistic model and covariate analysis to understand and predict the outcome of Cricket especially in test match format. Shubhra Singh et al. [8] have proposed a predictive tool and data visualization model that was devised using an open source data non-relational database namely HBase. Vishnu Sarpeshkar et al. [9] have developed a predictive mechanism to study the ball swing and the movement of batsmen while playing deliveries. Using these predictions, the skill levels of players were analyzed.

M. M. Rahman et al. [10] have used data mining techniques to analyze the performances of Bangladesh cricket team. Decision tree algorithm was used to make the predictions. M. J. Hossain et al. [11] devised a predictive model based on statistical data and genetic algorithm that helps to predict the ODI squad of Bangladesh cricket team. A. N. Wickramasinghe et al. [12] devised a predictive model that analyses the data prevailing in social media such as twitter feed to predict the outcome of a match. They also proposed a means to predict the player of a match even before the commencement of the game.

D. Thenmozhi et al. [13] have used data mining approaches such as random forest, K-nearest neighbor, Support Vector Machines and Gaussian Naïve Bayes classifier to predict the outcome of Indian Premier League matches. K. Ananthapadmanabha et al. [14] have devised a predictive model that analyzes the player and team performances and predicts who could be the potential candidates that may be targeted by bookies for match fixing. M. M. Hatharasinghe et al. [15] performed an exhaustive study of understanding the different types of predictive models and approaches used to predict the outcome of a cricket game and identified the challenges and limitations of these existing models.

N. Rodrigues et al. [16] developed a prediction model using multiple random forest regression to analyze the player performance and predict the outcome whether the player can be included in the playing squad. A. I. Anik et al. [17] used machine learning based approaches and statistical data analysis to predict the performance of players in the upcoming matches by considering the data pertaining to Bangladesh ODI squad. D. Saraswat et al. [18] used a data mining technique named Weighted Association Rule Mining to analyze the performance of Indian ODI cricket team.

J. Kumar et al. [19] used MLP networks along with decision tree algorithm to predict the outcome of any ODI cricket match. M. K. Nallakaruppan et al. [32] analyzed the impact of weather and conditions on the outcome of cricket matches. T. Singh et al. [33] developed a data mining based prediction model that used linear regression and Naive Bayes classifier to perform this operation.

K. Abbas et al. [34] have used machine learning algorithms to study the impact of DLS method on the outcome of ODI matches curtailed due to weather conditions. P. Kansal et al. [35] used several data mining classification and clustering algorithms to study the performance and impact of players on the outcome of IPL matches. A. A. Aburas et al. [36] used data analytics and business intelligence techniques to predict the outcome of ODI matches played as part of ICC world cup 2019. V. Phanse et al. [37] performed an extensive study of the different DLS method and its application in Cricket and studied the impact of the same on the outcome of the game. Also the limitations and shortcomings of this method was discussed in detail.

From all the above discussion, it can be observed that it is highly difficult to make predictions in the game of Cricket. The objective of the work is to use various data mining techniques develop a model that can make best predictions about the different aspects of this game of Cricket in ODI format.

III. Proposed Approach

A. Components of the proposed system

The initial step involved in developing the system is to pre-process the raw data. The raw data pertaining to players of countries are fetched from espncricinfo portal. Also, the statistical information pertaining to the players are fetched from espncricinfo statsguru repository. From this unprocessed data, several features that are instrumental for predicting in cricket are to be extracted as subsets using data pre-processing techniques. These subsets are called as feature subsets. Once the feature sets are finalized, these features are considered for a select set of training data comprising of players and matches. Typical data mining classification and prediction algorithms are applied over these extracted feature sets and a model is trained. Figure 1 shows the typical block diagram of the system.

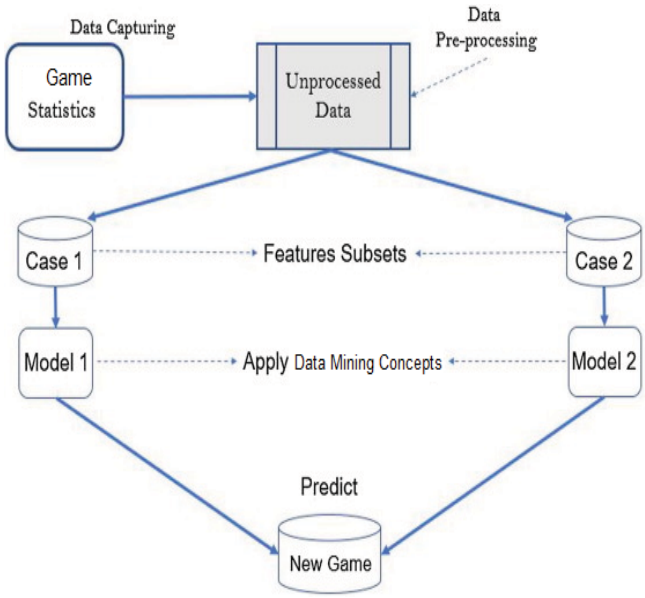

Fig. 1. Block Diagram of the proposed system

Once the model is trained successfully, it is then tested against new cases and finally applied across real time environment to check for the predictions. In this work, the data mining algorithms used are k-means clustering, decision trees,

random forest and support vector machines. The software used to perform data mining operations over the selected dataset is WEKA and the details pertaining to WEKA is provided in the subsequent section.

B. *Machine Learning Algorithms*

Machine Learning (ML) involves imparting Artificial Intelligence (AI) to a system so that the system self learns and tunes itself according to the situational demands. Several data mining algorithms have been developed to classify and predict outcomes. Fig. 2 shows the taxonomy of various classification and clustering algorithms that perform the same.

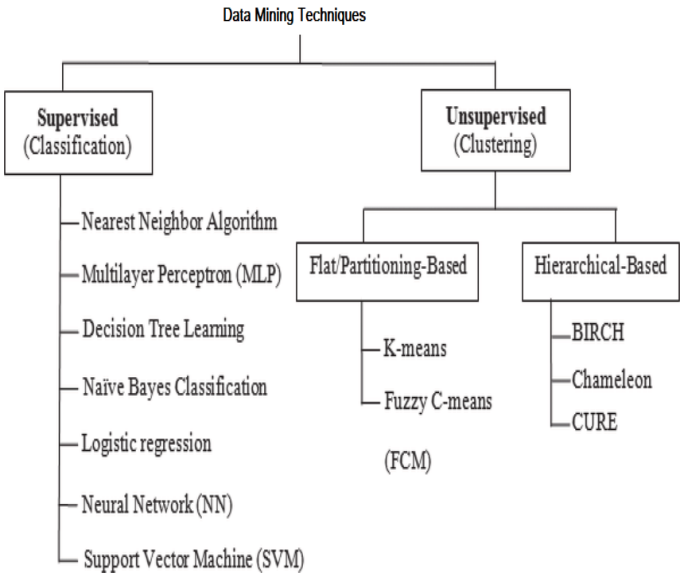

Fig. 2. Different Classification and Clustering Algorithms

C. *Knowledge flow of the proposed system*

Typical knowledge flow as depicted in fig.3 involves configuring the data source as input and proceeding with extracting the features and applying the data mining algorithms on top of the same. The data fetched from espncricinfo portal is populated into an excel file and the CSV loader is used to feed the data into the knowledge system. The dataset is then passed on to the class assigner where the typical classes are provided to the data. The class labelled data is then passed to the cross fold validation to create training and testing data. The test set and training set is then fed to the classifier algorithms and the results are observed. The classifier block is changed according to the type of classification algorithm that is to be used. In this case, the proposed system uses, decision trees, Support Vector Machines and random forest classification algorithm.

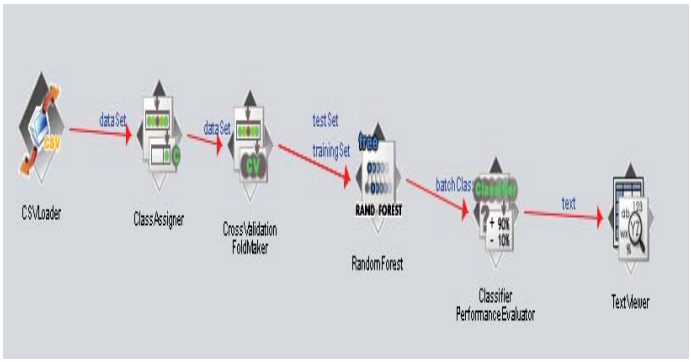

Fig. 3. Knowledge Flow of the Proposed System

D. *Experimental setup and requirements*

Typical hardware components required for developing and testing the predicting model are as follows:

- HP pavilion laptop
- Intel i5 processor @ 1.60 GHz
- Minimum of 8 GB RAM
- System Type: x-64

Typical software components required for developing and testing the predicting model are as follows:

- WEKA 3.9.4 and above compatible with windows
- 64 Bit Windows Operating System
- Microsoft Excel
- Microsoft Word
- Datasource: espncricinfo statsguru

WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. WEKA is open source in nature and can be easily downloaded. In this work, WEKA 3.9.4 is used to develop the predictive model.

IV. Experimental Results and Discussion

The experimental setup involved extracting play data from espncricinfo statsguru portal. Player data pertaining to cricketing teams of India, Bangladesh, Afghanistan, Pakistan, Australia, England and Sri Lanka were collected. The data coverage was fixed to the past 20 years. Several features were extracted for each player who represented their country in the ODI format for the past 3 years.

From, the pool of players 200 that were extracted, a select set of 40 players were used as training data and the remaining 160 players formed the testing data. The first step involved, classifying the players into three categories namely: batsman, bowler and all-rounder. The following confusion matrix was obtained.

TABLE I. CONFUSION MATRIX FOR PLAYER CLASSIFICATION

Category	Batsman	Bowler	All-Rounder	Total
Batsman	**56**	1	3	**60**
Bowler	1	**61**	4	**66**
All-Rounder	2	2	**30**	**34**
Total	**59**	**64**	**37**	**160**

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. In the above table, there are three classes namely batsman, bowler and all-rounder. Some important factors determine the efficiency of a confusion matrix are true positives, true negatives, false positives and false negatives.

True Positives (TP) denotes a set of records correctly classified according to the correct class label. True Negatives (TN) denotes a set of negative records classified correctly under negative class bucket. The diagonal highlighted in bold in the above figure represents the true values while the remaining are false positives and negatives respectively. Based on these values the metrics such as precision, recall and accuracy can be derived. Precision (P) is denoted as follows:

$$\text{Precision} = TP / (TP+FP) \quad (1)$$

$$\text{Recall} = TP / (TP+FN) \quad (2)$$

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (3)$$

The precision values for batsman, bowler and all-rounder are 0.94, 0.95 and 0.81 respectively. The recall values for batsman, bowler and all-rounder are 0.93, 0.92 and 0.81 respectively. The prediction accuracy of the classifier is 91.87% when decision tree classification is used, 93.46% when SVM is used and 95.78% when random forest used. Fig. 4 shows the prediction accuracy comparison for different methods. It can be observed that the accuracy levels are high when random forest is used.

From fig.4, it can be seen that the accuracy of Random Forest is higher when compared to Decision Tree and SVM.

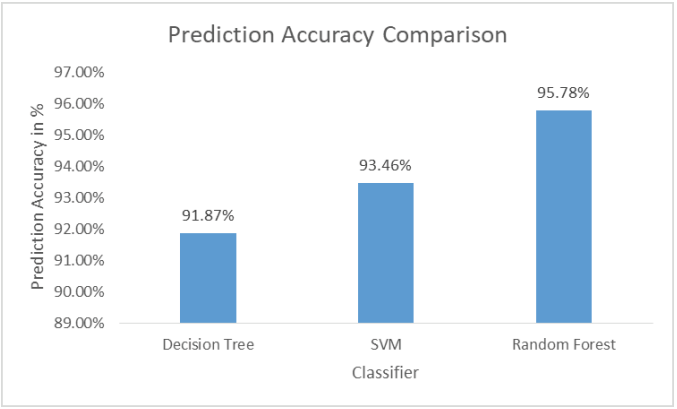

Fig. 4. Prediction accuracy comparison for different ML algorithms

V. CONCLUSION AND FUTURE WORK

This research work has proposed a classification model to classify the players based on the different features possessed by the players and a prediction as to which player can be included as part of the team squad in ODI matches. The classification was performed using three classification algorithms namely decision tree, SVM and Random forest. From, the results obtained from the experiments, the precision and recall values are noted to be above 90%, which is a good indicator of classification. The prediction accuracy of 95.78% is achieved for random forest, which means that the developed model was instrumental in predicting the right selection of players into the team. Future work will be towards fine tuning the above model to make different predictions such as win-loss, player performance in a particular game, predicting series results.

REFERENCES

[1] Subramanian Rama Iyer, Ramesh Sharda, Prediction of athletes performance using neural networks: An application in cricket team selection, Expert Systems with Applications, Volume 36, Issue 3, Part 1, 2009, Pages 5510-5522, ISSN 0957-4174.

[2] N. Soomro, R. Sanders, M. Soomro, Cricket injury prediction and surveillance by mobile application technology on smartphones, Journal of Science and Medicine in Sport, Volume 19, Supplement, 2015, Page e6, ISSN 1440-2440.

[3] Muhammad Asif, Ian G. McHale, In-play forecasting of win probability in One-Day International cricket: A dynamic logistic regression model, International Journal of Forecasting, Volume 32, Issue 1, 2016, Pages 34-43, ISSN 0169-2070.

[4] Neeraj Pathak, Hardik Wadhwa, Applications of Modern Classification Techniques to Predict the Outcome of ODI Cricket, Procedia Computer Science, Volume 87, 2016, Pages 55-60, ISSN 1877-0509.

[5] Hugh Norton, Steve Gray, Robert Faff, Yes, one-day international cricket 'in-play' trading strategies can be profitable!, Journal of Banking & Finance, Volume 61, Supplement 2, 2015,Pages S164-S176, ISSN 0378-4266.

[6] Tom Allen, Olivier Fauteux-Brault, David James, David Curtis, Finite Element Model of a Cricket Ball Impacting a Bat, Procedia Engineering, Volume 72, 2014, Pages 521-526, ISSN 1877-7058.

[7] Sohail Akhtar, Philip Scarf, Forecasting test cricket match outcomes in play, International Journal of Forecasting, Volume 28, Issue 3, 2012, Pages 632-643, ISSN 0169-2070.

[8] Shubhra Singh, Parmeet Kaur, IPL Visualization and Prediction Using HBase, Procedia Computer Science, Volume 122, 2017, Pages 910-915, ISSN 1877-0509.

[9] Vishnu Sarpeshkar, David L. Mann, Wayne Spratford, Bruce Abernethy, The influence of ball-swing on the timing and coordination of a natural interceptive task, Human Movement Science, Volume 54,2017, Pages 82-100, ISSN 0167-9457.

[10] M. M. Rahman, M. O. Faruque Shamim and S. Ismail, "An Analysis of Bangladesh One Day International Cricket Data: A Machine Learning Approach," 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, 2018, pp. 190-194, doi: 10.1109/ICISET.2018.8745588.

[11] M. J. Hossain, M. A. Kashem, M. S. Islam and M. E-Jannat, "Bangladesh Cricket Squad Prediction Using Statistical Data and Genetic Algorithm," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 2018, pp. 178-181, doi: 10.1109/CEEICT.2018.8628076.

[12] A. N. Wickramasinghe and R. D. Yapa, "Cricket Match Outcome Prediction Using Tweets and Prediction of the Man of the Match using Social Network Analysis: Case Study Using IPL Data," 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2018, pp. 1-1, doi: 10.1109/ICTER.2018.8615563.

[13] D. Thenmozhi, P. Mirunalini, S. M. Jaisakthi, S. Vasudevan, V. Veeramani Kannan and S. Sagubar Sadiq, "MoneyBall - Data Mining on Cricket Dataset," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICCIDS.2019.8862065.

[14] K. Ananthapadmanabha and K. Udayakumar, "Match fixing network analysis to verify nearness among internal participants of a cricket match," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017, pp. 1043-1048, doi: 10.1109/RTEICT.2017.8256758.

[15] M. M. Hatharasinghe and G. Poravi, "Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-4, doi: 10.1109/I2CT45611.2019.9033698.

[16] N. Rodrigues, N. Sequeira, S. Rodrigues and V. Shrivastava, "Cricket Squad Analysis Using Multiple Random Forest Regression," 2019 1st International Conference on Advances in Information Technology (ICAIT), Chikmagalur, India, 2019, pp. 104-108, doi: 10.1109/ICAIT47043.2019.8987367.

[17] A. I. Anik, S. Yeaser, A. G. M. I. Hossain and A. Chakrabarty, "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 2018, pp. 500-505, doi: 10.1109/CEEICT.2018.8628118.

[18] D. Saraswat, V. Dev and P. Singh, "Analyzing the performance of the Indian Cricket Team using Weighted Association Rule Mining," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, Uttar Pradesh, India, 2018, pp. 161-164, doi: 10.1109/GUCON.2018.8675115.

[19] J. Kumar, R. Kumar and P. Kumar, "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 343-347, doi: 10.1109/ICSCCC.2018.8703301.G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[20] Manickam M., Balasundaram A. and Ashokkumar S. (2020). Structure Optimized Multi Layer Trespass Perception System in Cloud. International Journal on Emerging Technologies, 11(3): 77–81.

[21] Magesh Kumar, S.; Ashok Kumar, S., Balasundaram, A. Providing Enhanced Resource Management Framework for Cloud Storage. Int. J. Eng. Adv. Technol. (IJERT) 2019, 9, 3903–3908.

[22] Subbiah S., Palaniappan S., Ashokkumar S., BalaSundaram A. (2020) A Novel Approach to View and Modify Data in Cloud Environment Using Attribute-Based Encryption. In: Ranganathan G., Chen J., Rocha Á. (eds) Inventive Communication and Computational Technologies. Lecture Notes in Networks and Systems, vol 89. Springer, Singapore. https://doi.org/10.1007/978-981-15-0146-3_20.

[23] S. Magesh Kumar, Balasundaram A, Sathish Kumar P J. (2020). An Improved Optimization Algorithm for Word Search on Disk. International Journal of Control and Automation, 13(4), 952 - 957.

[24] B. Ananthakrishnan, "An efficient approach for load balancing in cloud environment," International Journal of Scientific & Engineering Research, vol. 6, no. 4, pp. 36-40, 2015.

[25] Balasundaram, A., Chellappan, C. An intelligent video analytics model for abnormal event detection in online surveillance video. J Real-Time Image Proc 17, 915–930 (2020).

[26] A. Balasundaram and C. Chellappan, "Vision Based Motion Tracking in Real Time Videos," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, 2017, pp. 1-4, doi: 10.1109/ICCIC.2017.8524504.

[27] A. Balasundaram, C. Chellappan, "Vision Based Gesture Recognition: A Comprehensive Study", The IIOAB Journal, Vol.8, Issue.4, pp.20-28, 2017.

[28] Balasundaram, A.; Ashok Kumar, S.; Magesh Kumar, S. Optical Flow Based Object Movement Tracking. Int. J. Eng. Adv. Technol. (IJERT) 2019, 9, 3913–3916.

[29] Balasundaram, A.; Chellappan, C. Computer Vision based System to Detect Abandoned Objects. Int. J. Eng. Adv. Technol. (IJERT) 2019, 9, 4000-4010.

[30] Balasundaram, A. Computer Vision based Detection of Partially Occluded Faces. Int. J. Eng. Adv. Technol. (IJERT) 2020, 9, 2188-2200.

[31] Balasundaram, A, Ashokkumar, S. Study of Facial Expression Recognition using Machine Learning Techniques. JCR. 2020; 7(8): 2429-2437.

[32] M. K. Nallakaruppan, S. Nazz, K. Madhuvanthi, S. Karthikeyan and M. Medarametla, "Predicting the weather for Uninterrupted Cricket matches and Outdoor Sports events," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 451-458, doi: 10.1109/CONFLUENCE.2019.8776929.

[33] T. Singh, V. Singla and P. Bhatia, "Score and winning prediction in cricket through data mining," 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), Faridabad, 2015, pp. 60-66, doi: 10.1109/ICSCTI.2015.7489605.

[34] K. Abbas and S. Haider, "Duckworth-Lewis-Stern Method Comparison with Machine Learning Approach," 2019 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2019, pp. 197-1975, doi: 10.1109/FIT47737.2019.00045.

[35] P. Kansal, P. Kumar, H. Arya and A. Methaila, "Player valuation in Indian premier league auction using data mining technique," 2014 International Conference on Contemporary Computing and Informatics (IC3I), Mysore, 2014, pp. 197-203, doi: 10.1109/IC3I.2014.7019707.

[36] A. A. Aburas, M. Mehtab and Y. Mehtab, "ICC World Cup Prediction Based Data Analytics and Business Intelligent (BI) Techniques," 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Zhengzhou, China, 2018, pp. 273-2736, doi: 10.1109/CyberC.2018.00056.

[37] V. Phanse and S. Deorah, "Evaluation and Extension to the Duckworth Lewis Method: A Dual Application of Data Mining Techniques," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, 2011, pp. 763-770, doi: 10.1109/ICDMW.2011.79.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.