Player Performance Prediction in Football Game

Richard Pariath[1], Shailin Shah[2], Aditya Surve[3], Jayashri Mittal[4]

UG Student, Department of Computer Engineering, St. Francis Institute of Technology, Mumbai, India[1].
UG Student, Department of Computer Engineering, St. Francis Institute of Technology, Mumbai, India[2].
UG Student, Department of Computer Engineering, St. Francis Institute of Technology, Mumbai, India[3].
Assistant Professor, Department of Computer Engineering, St. Francis Institute of Technology, Mumbai, India[4]

Abstract— **In the game of football (soccer), the evaluation of players for transfer, scouting, squad formation and strategic planning is important. However, due to the vast pool of grassroots level player, short career span, differing performance throughout the individual's career, differing play conditions, positions and varying club budgets, it becomes difficult to identify the individual player's performance value altogether. Our Player Performance Prediction system aims at solving this complex problem analytically and involves learning from various attributes and skills of a football player. It considers the skill set values of the football player and predicts the performance value, which depicts the scope of improvement and the capability of the player. The objective of this system is to help the coaches and team management at the grassroots as well as higher levels to identify the future prospects in the game of football without being biased to subjective conditions like club budget, competitiveness in the league, and importance of the player in the team or region. Our system is based on a data-driven approach and we train our models to generate an appropriate holistic relationship between the players' attributes values, market value and performance value to be predicted. These values are dependent on the position that the football player plays in and the skills they possess.**

Keywords—*football, player performance, scouting, market value, data mining, soccer, team sport, player evaluation, sports analytics.*

I.INTRODUCTION

Football, also known as soccer to the western part of the world, is a team based sport played between two teams, each consisting of eleven players with a spherical ball. This sport is played in over 200 countries and in almost all weather conditions such as snow, rain, summer, etc. Football is governed by FIFA (Fédération Internationale de Football Association) as the highest body and further divides into various other bodies depending on the region and nationality. The competitiveness of the game varies from region to region based on the participation of the people, media coverage, and club budget. This, in turn, brings varying differences in the level of players and also fluctuates the market value and the skill level based on region, hype generated by the media, competitiveness of the league in which they play and their experience. The bigger the role the player plays in his team, the more likely they may be valued in the market like being the finest penalty taker, or spot free kick specialist, or other roles such as being a playmaker, chance creator, having excellent speed, etc. In India, despite the decrease in the youth participation in sports, particularly in the past few decades, the industry is putting in various means and efforts to improve the sports environments in the form of grassroots level programs, facilities, tournaments, coaching, public awareness, scholarships, etc. The problem however lies in the fact that it's difficult to search, analyze and coach the players in every part of the country; especially in rural India which consists of 70 percent of the 1.25 billion people approximately. To overcome this difficulty the clubs recruit scouts of vast experience and regional understanding to identify players.

The AIFF is trying to improve the situation by collaborating with various clubs and companies that make it possible to teach the coaches who may be inexperienced by bringing in connecting sessions with the experienced ones, hosting various tournaments at school, city, district, state level, establishing football academies and community initiatives.[2]

The proposed model is aimed specifically at the grass-root level players of India, further scaling to other soccer leagues. The system is trained as per the in-game values of the 2017 version of EA Sports FIFA [1]. The reason for choosing values based on a game is that it seemed to be the only source for a reliable, near accurate and open form of data available for football players spanning across several leagues. Moreover, the very nature of the game being a team based sport makes it difficult to analyze the players due to their dependencies on the skill-set of other team members, varying positions, formations, club budget, competitiveness in the league and injuries across their career span.

Our model is designed to estimate the performance value of the player based on the attributes and skill sets that the player possesses. Coaches can then take advantage of this performance value and train the player, reshuffle the team, recruit, and loan or sell the player. Another value added to this process is the market value of the player obtained through the performance value of the player. However, there will be an approximate deviation in that value by a certain amount due to irregularities in the demand for a particular position, club budget, contract period, injuries and current on-field performance.

The main contribution of our work is:

- To create a model that finds the performance value of the player from various attributes which is based on each player's position. The reason is that the factors contributing majorly to the performance value of the player are based on different positions in the game.

This means that categorizing the players into Goalkeepers, Midfielders, Defenders, and Attackers

- We also conclude that the overall performance of the player seems to be a better attribute to look into for recruitment since the market value may be influenced by release clauses of the players with the club, age, competitiveness of the league or club budget

The system has been designed for the following points

- ○ Help team management and coaches to make team and strategy related decisions.
- ○ To contribute to the knowledge, and study the statistics obtained from game data.
- ○ Our results have important implications for football managers and scouts, as data analytics facilitates precise, objective, and reliable estimates of market value that can be updated at any time.

The proposed work can be summarized in five steps. The overall steps are shown in Fig. 1. The first step is to extract data, we extract data from sofifa.com [1] website 2017 version. In the second step, we visualize the data. The third step is to reduce the dimensions of the data, after which we model the system and generate the predictions.

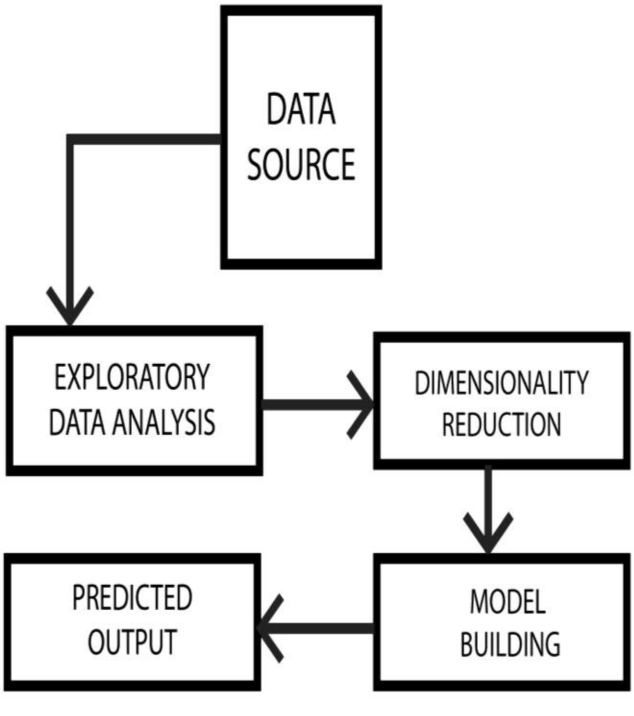

Fig. 1: The overall workflow of the system from extraction till prediction

II. RELATED WORK

Prediction of the market value of the player using data mining approach of linear regression is more suitable than using a crowd-based approach to predict the numerical target value. The model uses Data-driven approach such as using linear regression as it is much better than the crowd-based approach where the prediction is based on the decision given by the people.[3] Comparison of the transfermarkt.com market value (TMVE) and its own prediction, had smaller median errors, RMSE (Root Mean Square Error) and larger Pearson correlation.[3] The output obtained from the different models needs to be compared to check which model gives the best accuracy, [4] contains a comparison of many models but the dataset is too small. It gives references to the various models employed for prediction of the market value. It also lists the methods used to check the goodness of fit of the linear regression.

Similar work done in [5] outlines tools that enhance data analytics for U.V.A. Football (University of Virginia Football). One of the tools that they have outlined has been built with the goal of removing the guesswork out of play and providing the coaches with specific probabilities of potential outcomes of each play they could run. They have listed tools that provide the coaches with analytics that enhance their scouting and understanding of the opponents [5]. To help coaches to select players is also done by [9] using a visual approach providing them with better decisions on which recruits to pursue [9].

The player performance feature selection is discussed in [3], [4]. The player performance features are ever changing, hence monthly as well as yearly data is available on various websites [1].The challenges and the problems faced to get exact data related to all players' and to access confidential information like in-game performance is discussed in [6]. It solves this issue by deriving individual and team movements in soccer using an event-based dataset and analyzing the movements of players [6].

The game of soccer is an association game where the players of a team play together; thus winning depends on many factors. [7] Shows a novel model to find the interesting patterns in a football match that can guide the football coach and players to establish effective system against the opponent tactics. The achieved pattern from the decision model can predict the cooperation custom and pass pattern effectively.

The player evaluation in role-playing games, also a summarization of model accuracy, shows the ability to handle new data for better accuracy and fault tolerance. It also contains the analysis of how to reduce the authorial burden. Ideally, a player modeling technique should minimize the amount of effort that the game's author needs to put into creating the player models [8] [11].

III. PROPOSED WORK

A. Data extraction

The dataset has been scrapped from sofifa.com [1] which extracts the data from the EA Sports FIFA game franchise [1]. This data was scrapped on November 2017 from the FIFA 17 dataset which has around 36 player attributes. We divided the scraped data into various categories based on the player position. This is done using open source scraping tools [12].

We present a methodology for data-driven player market value estimation other than the crowd based approach used for market value prediction [3].Our Dataset consist of approx. 21,280 players divided into 4 sections as - Attackers, Midfielders, Defenders, and Goalkeepers. Each section has a number of attributes ranging from physical dimensions such as height, preferred foot etc. to skill set attributes such as dribbling, sprint, crossing, finishing, etc. The complexity of our problem increases as some players are present only in a particular category viz. Midfielders, Goalkeepers, and Defenders whereas some players can switch positions based on their skillset and formation applied by the team. Hence a single player present in one category may be present in another category too.

Initial data contains a lot of noise such as missing values, NULL values etc. which makes it necessary to normalize the data using preprocessing techniques. Later for modeling the system, dimensionality reduction was needed to identify the major attributes of the football players which affect the performance level more compared to certain attributes which barely affect their performance and indirectly the market value. Here we select the most relevant performance-related features like Dribbling, Shot Power, Attacking, Finishing, Head Accuracy, Acceleration, Crossing, Skill, Curve, and Ball Control. All these values are numerical with some values below 100 and some values like Power and Skill with values above 100. There is a need to normalize the Market values in the Training data set to a standard number system viz. 'X' millions of Euros, which are initially present in thousands as well as millions in the raw dataset.

B. *Dimensionality Reduction*

In the data visualization of the dataset, we generated a heat-map of all the supposedly independent attributes within the dataset.Fig.2 shows the heat map result and the correlation amongst various attributes. Heat map represents the internal dependency between the predictor variables of the dataset. This helps to find out whether there is any multicollinearity within these attributes. Ideally, these attributes (used as features to predict the outcome) should be independent of each other.

We found out that certain attributes like Growth of the player over time and the Power feature were internally correlated to a large extent. Such attributes; if considered in building the model, may drop the model's accuracy to some extent. Hence we need to drop these attributes from the dataset to be used for training.

The other method that we have used in this system is Principle component analysis (PCA) [10], A PCA is an orthogonal transformation of a set of (possibly correlated) variables into a set of uncorrelated variables that account for as much of the variance in the original variables as possible. In our model, we set the desired number of dimensions to different numbers to check and compare the model accuracy and select the best one. This will be used to further see if there are any more improvements in the model accuracy.

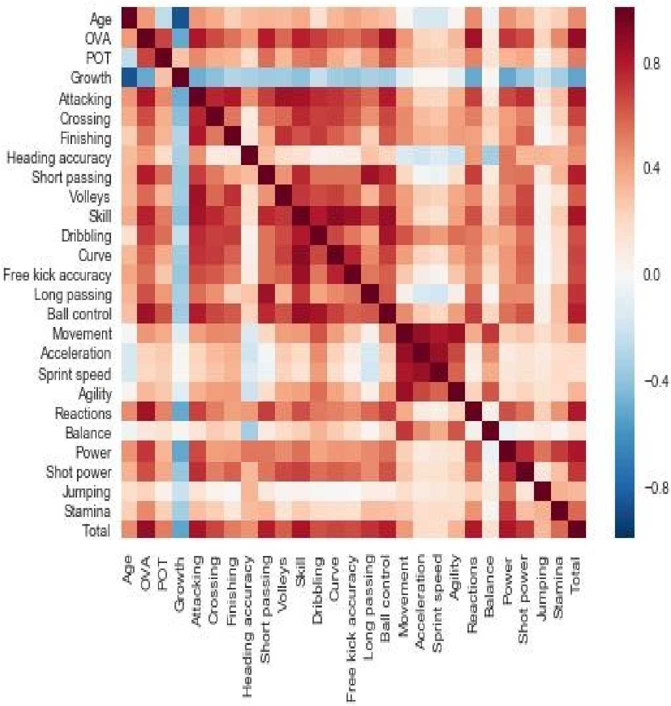

Fig. 2: The heat map of the dataset showing the relationship amongst the attributes

C. *Creating a Model*

We use a supervised learning algorithm and built a model for predicting the overall performance of the player and also to predict the market value of the player for scouting. As we have discussed earlier that market value is not the best indicator of the performance because market value is having a bias of a given overall performance parameter.

Each user is described by a feature vector which we have extracted from the sofifa.com [1] website 2017 version and also considering only the required attributes which are filtered. We create separate models for the different player positions and thus we consider different feature attributes for the different position by considering the significance of those attributes. For example, a player belonging to the attack category would have his performance value influenced heavily by finishing, shot power, etc. while a player in midfield category will have acceleration and crossing. The predictor variable is the market value (for scouting purpose) and the overall performance (for the coaches). Each dataset (based on positions) has 6000-9000 of players. We divide this dataset into training and testing dataset. The training dataset is then used to train the model. We have implemented the linear regression supervised learning algorithm and we use it for predicting the market value and the overall performance.

Linear regression is used for building the model of market value because the pattern of the market value is exponential in nature and it is giving a very low accuracy in case of the linear regression which follows the trend shown by the data. We convert the market value to the log of the market value which gives a better prediction. For the market value estimate, we show the model accuracy and discussions on bases of linear regression with a log of market value.

The model created for the overall player performance value as the output variable gives a good accuracy showing that it is a good indicator of the various skills of the player and hence can be used by coaches for player analysis and to calculate overall performance and the potential of the player.

1) Overall Performance Model

The Overall Performance of the player has a mean of 66.7; Fig. 3 shows the PDF (Probability Distribution Function) and the CDF (Cumulative Distribution Function) of the performance attribute. It can be seen that the values are properly defined in an interval. The performance model gives an accuracy of 84.34% and standard deviation of 0.84.

Although R-square value is a good performance indicator of the model we check the residual plot to see whether the model is biased or not. The residuals should not be either systematically high or low. So, the residuals should be centered on zero throughout the range of fitted values. In other words, our model is correct on average for all fitted values. Further, in the linear regression context, random errors are assumed to produce residuals that are normally distributed. Fig. 4 shows the residual plot of the performance of the player as it is randomly scattered we can say that the model is correct and gives a proper prediction on average on all values.

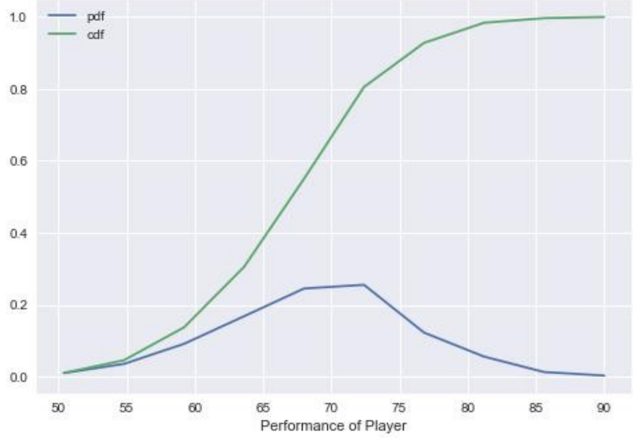

Fig. 3: The PDF and the CDF of the performance attribute

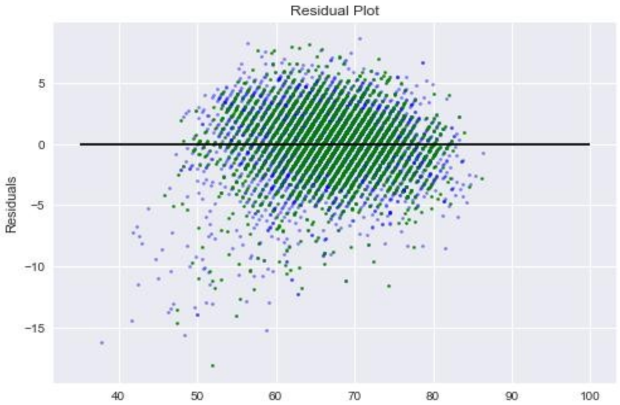

Fig. 4: The residual plot of performance

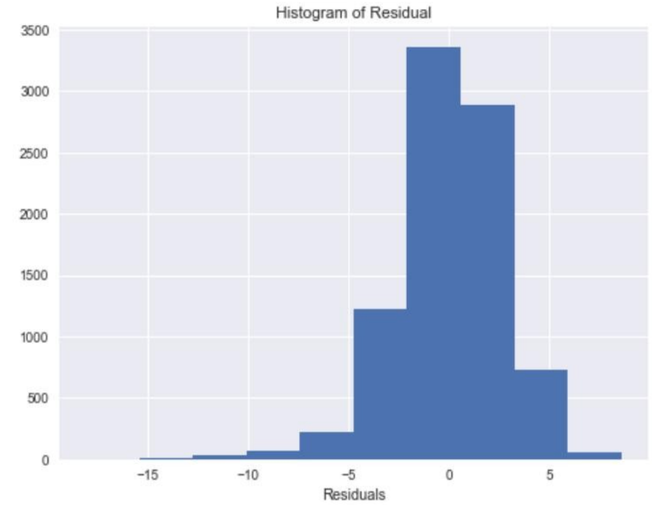

Fig. 5: The histogram of residuals

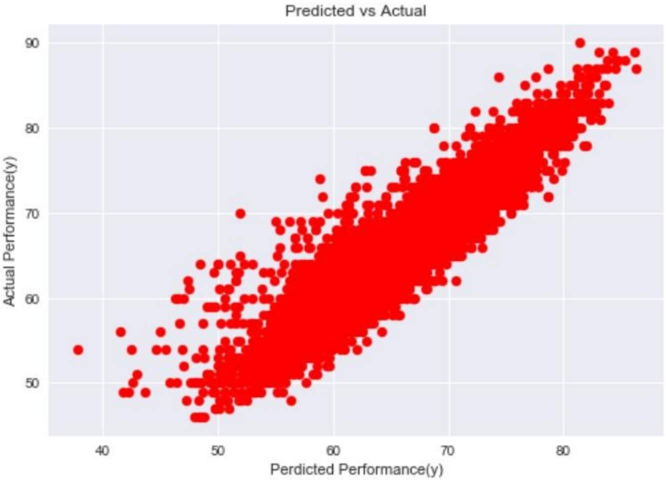

Fig. 6: The plot of the predicted performance to the actual performance

Fig. 5 show the histogram of the residuals showing the number of values in the given range. From the histogram we can see that the maximum residuals are in the range of zero, showing zero error in prediction. Fig. 6 shows the plot of predicted performance to the actual performance. The graphs show that the prediction was generally accurate except for the lower tail. This was not surprising because players in their very early years had no reputation and hence the model is unable to detect the exact performance at such an early stage career.

The model performance related to the R-square, RMSE and the median error is listed in the table. The model prediction is a bit low only because of some outlier in the data.

TABLE I

R SQUARE	*Mean absolute error*	*Median absolute error*	*RMSE*
0.84	2.01	1.61	2.67

2) *Market Value Prediction*

The model built for the market value prediction is a linear regression model. Fig. 7 shows the relation between the overall performance 'OVA' and the market value of the player. It can also be seen that after a certain threshold for a given value of the overall performance the player market value varies a lot. Thus there are many other factors that play role like the type of contract, the player's relationship with the team and the coach, how famous is the player etc. Thus we require a model that can fit into this pattern without getting over-fitted or under-fitted. However, the log-normal distribution of the market value gives an indication of the spread of the market value. Fig.8 shows the plot of the log-normal distribution of the market value. Fig. 9 shows the scatterplot of the overall performance and the log of the market value and the linear regression line which fits the data points.

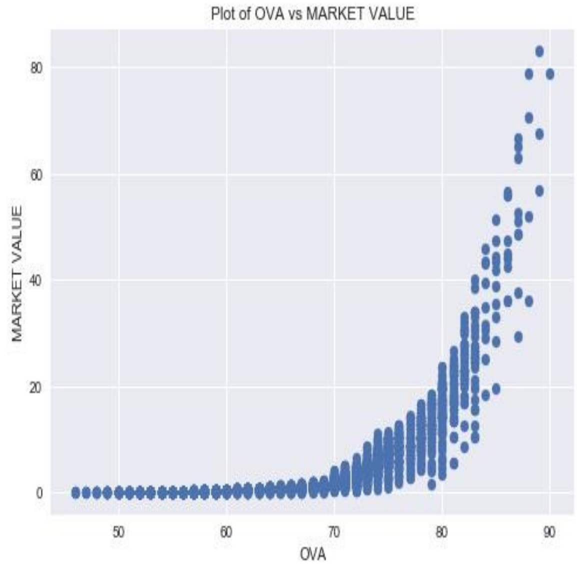

Fig. 7: The plot of the overall performance to the market value of the player

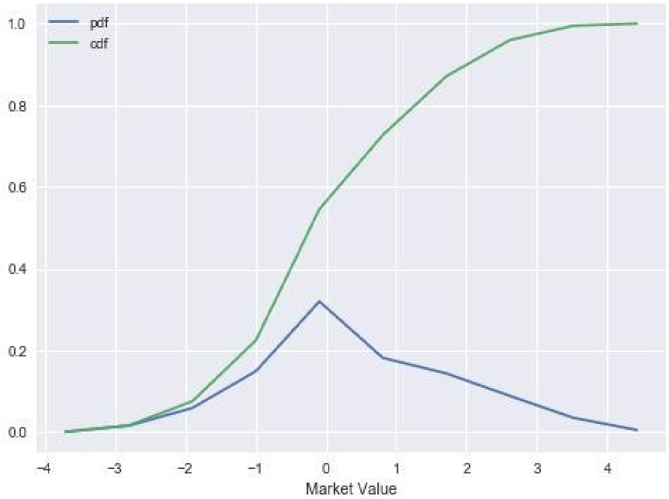

Fig. 8: The plot of PDF (Probability Distribution Function) and the CDF (Cumulative Distribution Function) of the log of the market value.

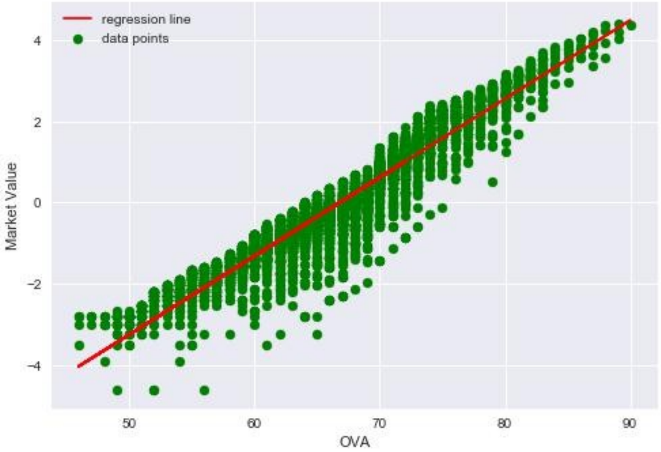

Fig. 9: The scatterplot of the overall performance and the log of the market value with the linear regression line.

TABLE II

R SQUARE	*Mean absolute error*	*Median absolute error*	*RMSE*
0.91	0.304	0.244	0.410

The model's accuracy is around 91%; thus a very good accuracy after applying the log of the market value and then we obtain the market value of the player. The RMSE score is 0.410. Table II summarizes the results obtain.

The results obtained here is the linear regression model developed between the overall performance of the player and the market value. The second model uses the prediction done in the first model and based on predicted value we predict the market value hierarchically. The overall value is the better indicator of performance as the market value for young players can be obtained only after some years of experience in the game of football.

IV. CONCLUSION

In this paper, we present a simple but comprehensive linear regression model which develops an ideal relationship between the performance features of the football players and the overall performance value. This model helps in scouting and coaching of the football players using data modeling approach.

We have a proven model accuracy of 84.34 % which is followed by the second linear regression model in the hierarchy with an accuracy score of 91%. This second model predicts the future market value of the players on the basis of the overall performance value predicted by the first model.

The model accuracy analysis shown is for the Midfielder dataset, similar accuracy has been obtained for all different positions - Attackers, Midfielders, Defenders, and Goalkeepers for which we built separate models on the same lines as of the midfielder dataset. The model prediction accuracy shows how a simple model can solve a huge problem related to scouting and coaching.

The major impact of this system would be an advantage in identifying the grass-root level talented players who fail to receive exposure as compared to the other renowned football players.

Currently, the system uses the sofifa.com [1] dataset for training purposes. This can be improved in the future by gathering the latest data by spider-cameras over the football fields. A further in-depth analysis will help us to find out the effects of physical attributes of the players which also may be affecting the performance metrics of the players over the span in which they play the game.

Thus our model can be used by the coaches to improve the performance of grassroots level players and help to allow team managers to manage their team in a just and better way.

References

[1] Dataset from https://sofifia.com

[2] https://www.the-aiff.com

[3] R. Stanojevic and L. Gyarmati, "Towards Data-Driven Football Player Assessment," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, 2016, pp. 167-172.

[4] H. Elkins et al., "Implementing data analytics for U.Va. Football," 2017 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 2017, pp. 202-207

[5] He Y. Predicting Market Value of Soccer Players Using Linear Modeling Techniques. Technical Report. University of California, Berkeley. 2015

[6] L. Gyarmati and M. Hefeeda, "Competition-Wide Evaluation of Individual and Team Movements in Soccer," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, 2016, pp. 144-151.

[7] B. Chai and X. Xu, "A Novel Decision Support Model to Discover the Interesting Pattern in Football Match," 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, 2009, pp. 67-70.

[8] Brent Harrison, Stephen G. Ware, Matthew W.Fendt, David L. Roberts, “A Survey and Analysis of Techniques for Player Behavior Prediction in Massively Multiplayer Online Role-Playing Games”

[9] Walter, A. Citera, K. Knowles, M. Lowen, C. Oldenburg, H. Shahin, W. Scherer, and C. Tuttle, “Implementation of a recruit visualization tool for UVA football,” *2017 Systems and Information Engineering Design Symposium (SIEDS)*, 2017.

[10] M. F. I. Ibrahim and A. A. Al-Jumaily, "PCA indexing based feature learning and feature selection," *2016 8th Cairo International Biomedical Engineering Conference (CIBEC)*, Cairo, 2016, pp. 68-71.

[11] B. Harrison and D. L. Roberts, ‘‘Using sequential observations to model and predict player behavior,’’ in Proc. 6th Int. Conf. Found. Digital Games, 2011, pp. 91–98.

[12] C. Lewis and N. Wardrip-Fruin, ‘‘Mining game statistics from web services: A World of Warcraft armory case study,’’ in Proc. 5th Int. Conf. Found. Digital Games, 2010, pp. 100–107.