Kobe Braynt Shot Prediction using Machine Learning

Taimur Shahzad[1], Zahoor Ur Rehman[1], Eliza Batool[1], Zain Ullah[2]
[1] Department of computer science, COMSATS University Islamabad Attock campus, Pakistan
[2] Faculty of Information Engineering, Sapienza University of Rome Italy
[1](fa20-rcs-011@cuiatk.edu.pk)

Abstract--*Kobe Bryant was one of the best players of Basketball. Data regarding his 20 years of playing games are available on the Kaggle. We transform the categorical features by PCA and normalize the data by the min-max normalization technique. Machine learning techniques (Logistic Regression, Random Forest, Linear Discriminant Analysis, Naïve Bayes, Gradient Boosting, Adaboost, and Neural Network) are applied on pre-processed data to classify whether he made shot or not. The prediction accuracy of LR, RF, LDA, NB, GB, ABC and ANN is 67.84%,64.22%,67.82%,0.61%,67.8%,68% and 67% respectively on hold an out method. The experimental results show that Adaboost has the highest prediction accuracy as compared to other methods with 5 cross-validations. Finally, we have got satisfactory results as compared to our benchmark (Kaggle).*

Keywords: Shot Prediction, Machine Learning, Neural Network, Basketball

I. INTRODUCTION

Basketball is an internationally renowned game; and National Basketball Association, also known as NBA; is the most popular league of the game. Kobe Bryant was one of the best players in NBA. He denoted his retirement by scoring 60 focuses in his last game for Los Angeles Laker on April 12, 2016. Starting at the age of 17 at NBA, Kobe acquired the game's most noteworthy honours all through his long profession.

The aim of the research is to predict every field goal attempted during his 20-year career, whether he made the shot or not. Dataset was taken from Kaggle, where already some of the participants had worked.

Data was analysed by different machine learning analysts; use of logistic regression and neural network to predict the shot prediction type [1], running the feedforward NN for the analysis of nonlinear sports data [2], identification of the basketball strategy using player tracking data [4], use of LSTM for Multi-Modal Trajectory prediction of NBA player [11], use of neural network for selection of Most Valued Player of NBA [12], predicting the NBA game results using ANN and decision tree [14], use of mixture density network and bidirectional LSTM for basketball trajectory prediction [15], and classification of the NBA offensive plays using neural network [15].

In our work, we first apply PCA for column transformation, followed by minmax scalar normalization and then pass this scaled data to different machine learning algorithms for classification whether he made the shot or not.

We applied Logistic Regression, Random Forest, Linear Discriminant Analysis, Naïve Bayes, Gradient Boosting, Neural Network and Adaboost on processed data. The performance of our method is evaluated by subsequent evaluation metrics; accuracy, precision, recall and F-measure.

Organization of the remaining paper is as follows: in section 2 we overview the related work about basketball player prediction and shot selection. Section 3 describe the methodology in steps for our shot selection prediction. Section 4 and 5 narrates Results and conclusion respectively.

II. LITERATURE REVIEW

In [1] different features are used to compare contrast the performance of NN, logistic regression gradient boosting models. Accuracies were nearly around 65% ±0.1%. Yet it is concluded that changing the features may also decrease the accuracy. Hence the importance of some features is significant i.e., shot type. Also, good quality data affects the overall performance. Furthermore, deep NN can play a vital role in good shot selection and the accuracies can speak up. Moreover, large data can help and predict better results as it spans over the whole season or even multiple seasons. Here NBA (higher quality league) dataset is being used to run the algorithm. First B Serbia's Men Basketball league dataset of 890 games comprising the 5 seasons from 2005- 2010 is being used by [2] to run the feedforward NN for the analysis of nonlinear sports data. Several analysis are being on the dataset where shooting from different field positions and basic basketball parameters for winning are the most contributed. 11 divisions of the field positions are being analyzed including 6 two-point and 5 three-point positions and the results show that two-point shots are the most important analysis element among all. Moreover, both offensive and defensive approaches are being analyzed, concluded on the fact that it is crucial to winning a game under the hoop. Despite the BSV real-time program, more data and some new software could be a contribution to the future in this regard.

SVM is considered a powerful classification algorithm, but as it has no rule generation capability, thus having a limitation. A hybrid FSVM is introduced by [3] that not only includes the quality of classification but also generates rules for decision making using the fuzzy approach. Hence is used to better predict the win and loss outcomes as compared to SVM alone. Furthermore, the work can be extended to predict the win and loss scores, not only for basketball but for some other sports too. Shooting Prediction for NBA: a comparative study of Random Forest XGBoost is conducted using the 203591 shots of the 2014-2015 regular season. XGBoost has the highest accuracy as compared to the Random Forest (RF) but RF is a good classifier too [4]. Theoretical/Mathematical Prediction of shoots under various circumstances is presented and compared with actual/ observed NBA datasets [5].

Two phased DEA (Data Envelopment Analysis)-MLA approach using multiple input multiple outputs is presented in [6] using guard position of 26 players' data. Their efficiency frontier is calculated using NN, linear regression, and SVM, the former one has an error rate of < 1% while the latter except for SVM has an error rate < 2%. The ranking is done through Andersen Petersen's model. Further, the work can be extended using the Distance Based Analysis (DBA) and its comparison with DEA, for the new player prediction too. In [7] the data-driven defensive strategy learning including the analytical and classification model is narrated. The former has the one-on-one relationship of players, while the latter is against the pick-and-roll play. The technique presented is the spatiotemporal pattern recognition; and classification is done for two defensive and one offensive technique. Classifiers being used are SVM, DT, and KNN; amongst all SVM is the best having accuracy of 68.9%. Further, the work can be extended using the alternative methods of labeling instead of the analytical method being widely used.

The group learning approach for basketball players' prediction is introduced in [8] using the Mixture of Finite Mixtures model. The process being used is Log Gaussian Cox Process (LGCP) and the algorithm is Markov Chain Monte Carlo (MCMC). The study estimated the groups' number and their configuration along with the qualitative summary of shots played by various players. Data being used is NBA 2017-2018. Further, the work can be extended using a unified approach despite the two-stage grouping, and secondly, the auxiliary information can also be incorporated. Imitation learning method implementation using RNN for soccer is presented in [9] using the top-tier soccer league data (2019). Contributions made so far include the method application on movement traces of soccer, accuracy check, and influence of single to relative behavior.

Further, the work can be extended for the multi-agent Intelligent ML framework using Naïve Bayes, ANN, and DT for predicting the results of the game played and are then compared with the previous data. Defensive rebounds (DRB) are the most significant feature of all. Also, Three-Point Percentage (TPP), Free Throws (FT), and Total Rebound (TRB) ensure 2-4% accuracy improvement [10]. Analysis was conducted of NBA player and shot prediction using Random Forest and Xgboost [19].

Data being analysed by various analysts using various machine learning techniques, such as; identification of the basketball strategy using player tracking data [4], use of LSTM for Multi-Modal Trajectory prediction of NBA player [11], use of neural network for selection of Most Valued Player of NBA [12], predicting the NBA game results using ANN and decision tree [14],], use of mixture density network and bidirectional LSTM for basketball trajectory prediction [15], and classification of the NBA offensive plays using neural network [15]. But to the best of our knowledge, there isn't any study presented on Kobe Bryant shot prediction dataset; by the transformation of the categorical features and comparing the results of ensemble, traditional and neural network methods on this dataset.

III. PROPOSED METHODOLOGY

In our work we downloaded the Kobe shots data set from Kaggle. We first performed pre-processing on this data and then passed the processed data to machine learning techniques. The machine learning techniques classified whether he made the shot or not. We evaluated the performance of these method using different evaluation metrics. Figure1 shows the workflow of our proposed work. The implementation was performed using python language and Jupyter notebook as a development tool.

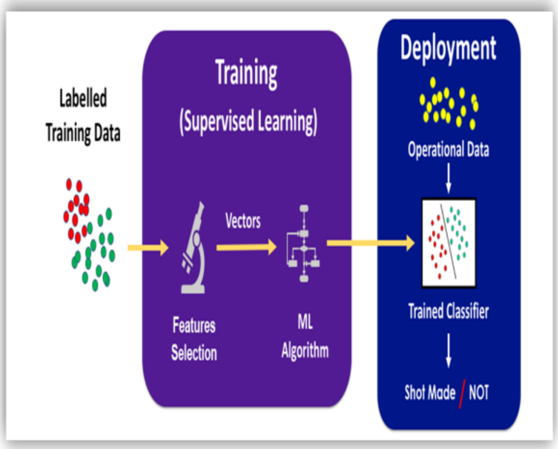

Fig. 1. Workflow diagram

3.1 Dataset

Dataset contains more than 30000 data values and 25 columns in which SHOT_MADE_FLAG is the class label; Dataset have missing values and categorical columns to handle, therefore, after handling the missing values and categorical data we finally got 90 columns and 26000 data values. (For categorical we have used One Hot Encoder and for missing values we simply discard them. Detailed sample of the dataset and class distribution bar chart are shown in Figure 2 and 3 respectively.

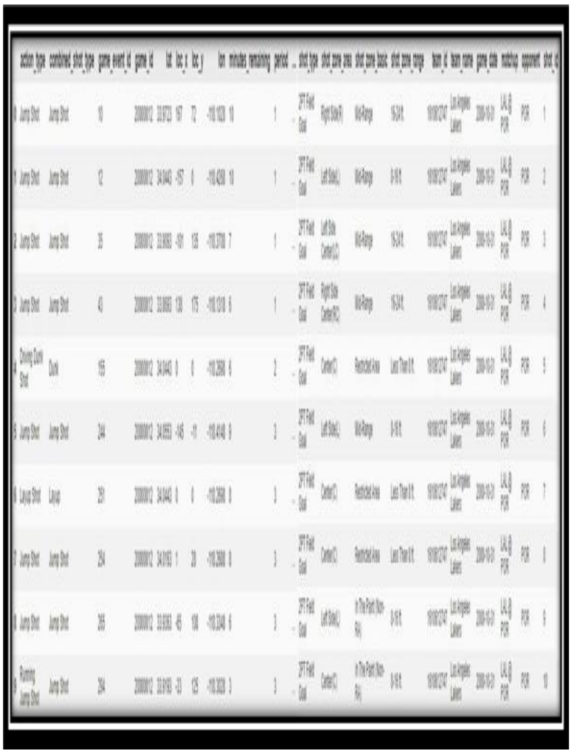

Fig. 2. Kobe shot selection dataset

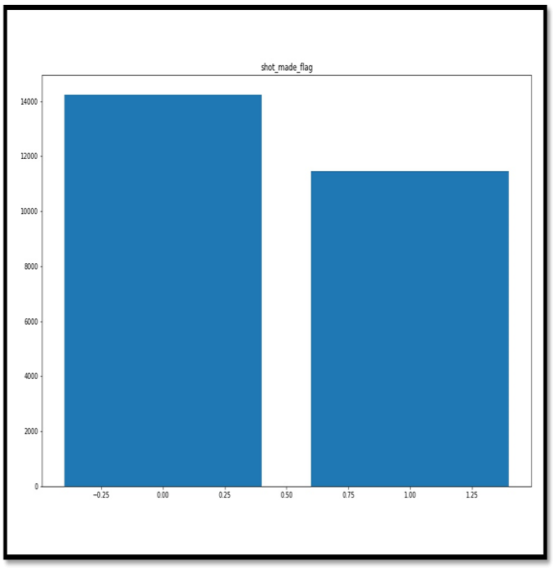

Fig. 3. Class Label Distribution

3.2 Pre-processing

We removed the missing value's row from this dataset and converted the categorical columns into discrete by applying one hot encoding method. In our data set we have some categorical features. We applied one hot encoding on it and performed transformation and generated new features using SK learn column transformer and feature. Scaling was performed using minmax scalar. Figure 4 shows the train and test split distribution.

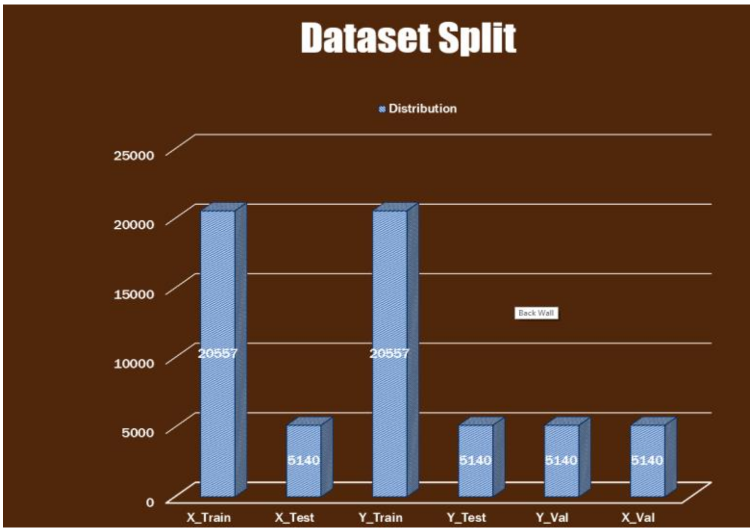

Fig. 4. Train Test Split distribution

3.3 Machine Learning Method

3.3.1 Logistic Regression:

It is a supervised learning model used for classification. It works very well when we have small dataset and have a binary class variable. Logistic regression has a coefficient for each attribute and intercept or bias. These are the parameter of logistic regression. We calculated the weighted sum by multiplying the coefficient with attribute value and added in bias. Then weighted sum was passed to the activation function (for-example; sigmoid) and the output is produced. In our work, we gave scaled data to logistic regression, and it predicted the respected class by the above procedure. Code snippet for shot selection prediction using logistic regression is as follows:

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.metrics import confusion_matrix
LR = LogisticRegression(C=0.01, solver='lbfgs').fit(X_train,y_train)
yhat = LR.predict(X_test)
print("Accuracy: ", metrics.accuracy_score(y_test, yhat))
print (classification_report(y_test, yhat))
print(LR.coef_)
```

3.3.2 Random Forest

It is an ensemble learning classifier and is used for classification as well as regression models. It constructs multiple trees of decision trees and output the class that is the mode of classes. The random forest has certain hyper-parameters that require tuning for predicting accurate results. We trained Random Forest to predict shot on the parameters given in Table 1.

Table 1 Random Forest Parameter

Hyper parameter name	Values
Bootstrap	True
Criterion	Gini
min_samples_split	2
n_estimators	50
random_state	None
max_features	Auto
min_samples_leaf	1

Code snippet for shot selection prediction using Random Forest is as follows:

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.ensemble import RandomForestClassifier
clf=RandomForestClassifier(n_estimators=50)
clf.fit(X_train,y_train)
yhat4=clf.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, yhat4))
print("Confusion Matrix:")
print(confusion_matrix(y_test, yhat4))
print("Classification Report")
print(classification_report(y_test, yhat4))
```

3.3.3 Linear Discriminant Analysis (LDA):

LDA is a technique for dimensionality reduction, and it can also be used for classification problems. In our work, we used LDA to predict shot prediction that estimates the probability of a new instance belonging to each class. Output class is the one with the highest probability. We used solver parameter as svd, Shrinkage as auto.

3.3.4 Naïve bayes:

Bayes theorem is the basis of this classification algorithm. It assumes that particular feature's presence in a class, has nothing to do with any other feature's presence. It is particularly good and performs well on a large dataset.

3.3.5 Gradient Boosting (GB):

GB is a machine learning classifier that creates a strong predictive model by combining many weak learning models. The decision tree is mostly used as a weak learner. It has three main elements: Loss function, Weak Learner, and additive model.

3.3.6 Adaptive boosting (Adaboost):

Adaboost is an ensemble boosting classifier that increases the accuracy of classifiers by combining multiple classifiers. It is a meta-algorithm and may be used at the side of many different gaining knowledge of algorithms to enhance its performance. It is miles bendy within the sense that the construction of subsequent phases is about apart to allow for one's situations divided into preceding phases. Touchy to noisy and outside facts. Adaboost is an algorithm for building "stable" divisions as a linear thing of the "simple" "susceptible" category, very last divisions primarily based on the average vote of susceptible dividers, flexible electricity as opposed to repetition, uses training to reset weight each education sample makes use of weight to determine opportunities

3.3.7 Artificial Neural Network

Machine learning algorithm Neural Network remains one of the best and most used in the past decade and we considered it for this problem. We have proposed that we will use an artificial neural network with few deep hidden layers. Figure 4 clearly explain the structure of our MLP model. After pre-processing of dataset we have 90 features and 25000 rows and our model has 90 neurons as input. So, the model is now having sequence of layers explained below:

Input 90 neurons → 64 → 64 → 32 → 8 → 1

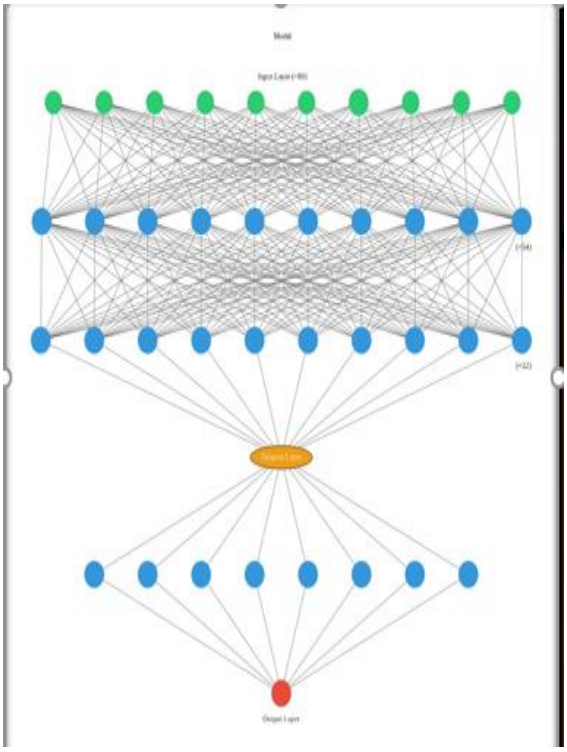

Fig. 4. The architecture of ANN for Shot Prediction

Input and hidden layers have activation function RELU, and the Final output layer has activation of the SIGMOID function.

IV. RESULTS AND DISCUSSION

We perform the comparison between different machine learning techniques to predict whether a shot is made or not. Logistic regression, Random Forest, Linear Discriminant Analysis, Naïve Bayes, Gradient Boosting, Artificial Neural network, and Adaboost are applied using hold an out and 5-fold cross-validation. Table2 shows the results of the hold an out method for shot prediction. Logistic regression has the highest F measure as compared to other methods using the hold an out method. Table2 shows the results of the 5-fold cross-validation method for shot prediction. Adaboost has the highest prediction accuracy as compared to other methods using 5-fold cross-validation. The overall performance of AdaBoost is highest with hold an out and 5 cross-validations.

4.1 Prediction results of machine learning techniques using hold an out method

We apply machine learning techniques for shots prediction using the hold an out and cross-validation method. Table2 shows that Logistic regression has the highest F score as compared to other methods.

Table 2. Results comparison of machine learning method withhold and out for shot prediction

Method	**Accuracy**	**Precision**	**Recall**	**F1**
Logistic regression	**67.84%**	**69%**	**68%**	**67%**
Random Forest	64.22%	64%	64%	64%
LDA	67.82%	69%	68%	66%
Naïve bayes	0.61%	0.73%	0.57%	0.51%
GB	0.678%	0.69%	0.66%	0.66%
Adaboost	0.68%	0.69%	0.66%	0.66%

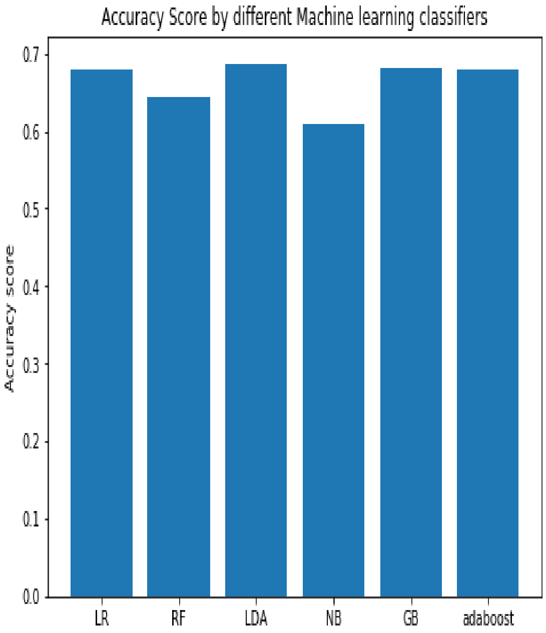

Fig. 5. Accuracy graph of Machine Learning method

Figure 6 shows the accuracy and Figure 7 shows the F1 score of LR, RF and LDA, GB, ABC, and NB on Kaggle Kobe Braynt shots dataset and compare the performance of these classifiers on this dataset. The experimental results showed that LDA performs very well in term of F measure as compared to LR, RF, GB, ABC, and NB. The performance of NB is low as compared to LR, RF and LDA, GB, ABC, and NB. Figure 8 shows the loss graph of the neural network for shot prediction.

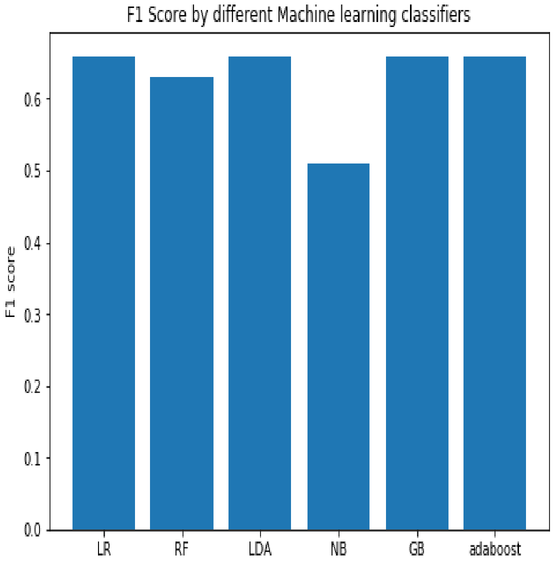

Fig. 6. F1 score graph of Machine Learning method

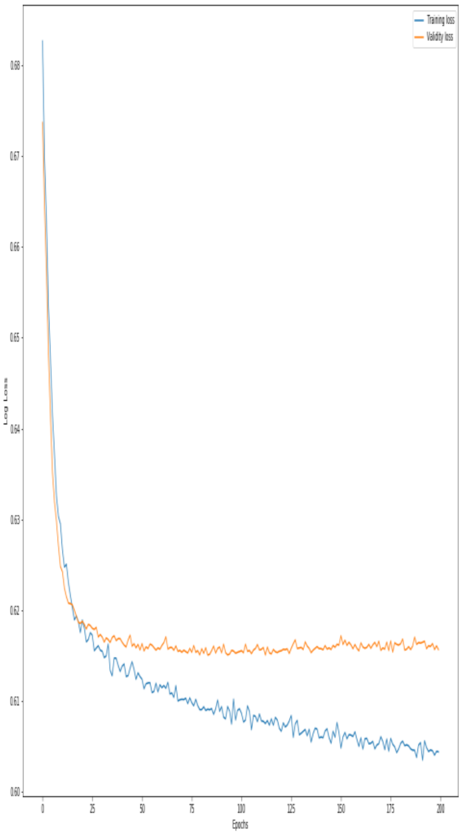

Fig. 6. Training and validation graph of neural network for shot prediction

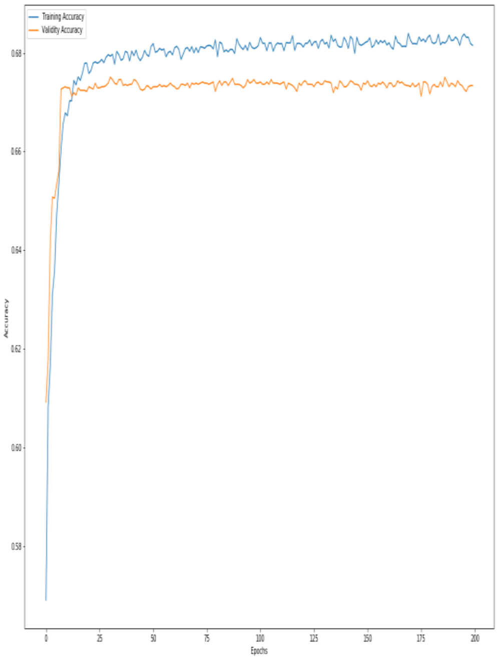

Fig. 7.training and validation accuracy of neural network for shots prediction

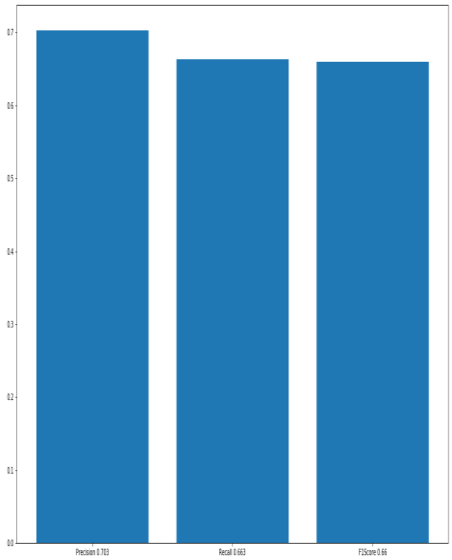

Fig. 8. Precision and recall of neural network for shots prediction

4.2 Prediction results of machine learning techniques using 5 fold cross-validation

Table 3. Results comparison of machine learning method with 5-fold cross validation for shot prediction

Method	Accuracy
LDA	0.68%
RF	0.657%
Adaboost	0.681%
GB	0.681%
Naïve bayes	0.62%
LR	0.59%

Table3 shows the results of LR, RF and LDA, GB, ABC, and NB using 5-fold cross-validation. Adaboost achieved the highest accuracy as compared to other methods.

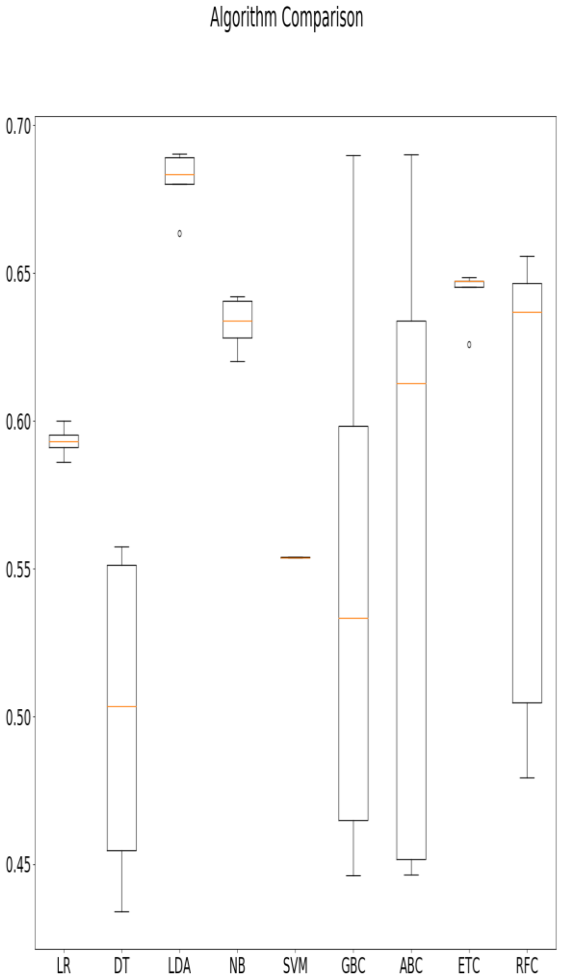

Fig. 9. Accuracy performance comparison using boxplot

In our experiment, we concluded that by using cross-validation the result of all methods increases as compared to the hold an out method.

The results of our method improved because when we used cross-validation in our case. We have taken the 5-fold of data and 5 different distributions of data. We trained and evaluated a complete model on each fold. Finally, the averaged result of all folds is presented.

V. CONCLUSION

Predicting games and players performance using Machine Learning is still used and highly recommended. In our work we applied Logistic Regression, Random Forest, Linear Discriminant Analysis, Naïve bayes, Gradient Boosting, Adaboost and Neural Network on Kaggle dataset. We applied these models using hold an out and 5-fold cross validation to predict whether Kobe made the shot or not. The experimental results shows that Adaboost performs best as compared to other methods with both hold an out and 5-fold cross validation. So in conclusion, we understood from the results that the data is truly random, because in most examples with exact same features Kobe had Made and Not Made shots, so getting better results from a random data is quite difficult but still if this dataset trained by a good Deep learning model maybe it will give better results. In the future, we can select relevant features from given features by applying the mutual information feature selection technique. The Voting Ensemble method can also be applied to selected features to improve the results.

REFERENCES

1. Murakami-Moses, Max. "Analysis of Machine Learning Models Predicting Basketball Shot Success."
2. Ivanković, Z., et al. "Analysis of basketball games using neural networks." *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE, 2010.
3. Jain, Sushma, and Harmandeep Kaur. "Machine learning approaches to predict basketball game outcome." *2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)(Fall)*. IEEE, 2017.
4. Oughali, Maram Shikh, Mariah Bahloul, and Sahar A. El Rahman. "Analysis of NBA players and shot prediction using random forest and XGBoost models." *2019 International Conference on Computer and Information Sciences (ICCIS)*. IEEE, 2019.
5. Skinner, Brian. "The problem of shot selection in basketball." *PloS one* 7.1 (2012): e30776.
6. Radovanović, Sandro, Milan Radojičić, and Gordana Savić. "Two-phased DEA-MLA approach for predicting efficiency of NBA players." *Yugoslav Journal of*

Operations Research 24.3 (2014): 347-358.

7. Tian, Changjia, et al. "Use of machine learning to automate the identification of basketball strategies using whole team player tracking data." *Applied Sciences* 10.1 (2020): 24.
8. Hu, Guanyu, Hou-Cheng Yang, and Yishu Xue. "Bayesian group learning for shot selection of professional basketball players." *Stat* (2020): e324.
9. Lindström, Per, et al. "Predicting player trajectories in shot situations in soccer." *International Workshop on Machine Learning and Data Mining for Sports Analytics*. Springer, Cham, 2020.
10. Thabtah, Fadi, Li Zhang, and Neda Abdelhamid. "NBA game result prediction using feature analysis and machine learning." *Annals of Data Science* 6.1 (2019): 103-116.
11. Tian, Changjia, et al. "Use of machine learning to automate the identification of basketball strategies using whole team player tracking data." *Applied Sciences* 10.1 (2020): 24.
12. Hauri, Sandro, et al. "Multi-Modal Trajectory Prediction of NBA Players." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021.
13. Chen, Yuefei, Junyan Dai, and Changjiang Zhang. "A neural network model of the NBA most valued player selection prediction." *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence*. 2019.
14. Thabtah, Fadi, Li Zhang, and Neda Abdelhamid. "NBA game result prediction using feature analysis and machine learning." *Annals of Data Science* 6.1 (2019): 103-116.
15. Zhao, Yu, et al. "Applying deep bidirectional LSTM and mixture density network for basketball trajectory prediction." *Optik* 158 (2018): 266-272.
16. Wang, Kuan-Chieh, and Richard Zemel. "Classifying NBA offensive plays using neural networks." *Proceedings of MIT Sloan Sports Analytics Conference*. Vol. 4. 2016.
17. Oughali, Maram Shikh, Mariah Bahloul, and Sahar A. El Rahman. "Analysis of NBA players and shot prediction using random forest and XGBoost models." *2019 International Conference on Computer and Information Sciences (ICCIS)*. IEEE, 2019.
18. https://www.kaggle.com/matt4byu/kobe-bryant-shot-selection-analysis with-xgboostAdditional-Methodology
19. https://www.kaggle.com/c/kobe-bryant-shot-selection
20. https://www.kaggle.com/chitramdasgupta/kobe-bry