Can a Video Game and Artificial Intelligence Assist for Selecting National Soccer Squads ?

Dr. Eren Unlu

***Abstract*—We have used the FIFA19 video game open dataset of soccer player attributes and the actual list of squads of national teams that participated in World Cup 2018, which almost coincides in time with the game's release date. With the intended rationale behind that numerous expert game developers should have spent considerable amount of time to assess each individual player's attributes; we can develop and test data science and machine learning tools to select national soccer teams in an attempt to assist coaches. The work provides detailed explanatory data analysis and state-of-the-art machine learning and interpretability measures.**

***Index Terms*—data science, machine learning, artificial intelligence**

I. Introduction

Electronic Arts (EA Games) video game producing giant based in California, USA has make a dataset available regarding its most popular video game in 2018-2019, *FIFA19*, on *Kaggle*, a website gathering challenges, datasets and collaboration for data science [1] [2]. The dataset contains each registered player's wide array of information such as his preferred foot, position on the soccer field, nationality, his club, weekly wage etc. and a list of numerous integer attributes such as acceleration, shooting ability etc.. These particular attributes are directly connected to virtual ability of the player in game play.

It would not be an invalid assumption to consider developers of EA Sports and particularly its world wide renowned 27 years long video game series *FIFA* have carefully crafted these attributes. As game play is considered highly realistic and has been getting better and better through three decades; the player attributes can be regarded as expert evaluated metrics which reflect the real world accurately. Thus, we have come up with the novel idea of using these video game originated features to investigate the possibility of developing an artificial intelligence backed oracle to select national soccer squads. As game's release date is September 2018 and the FIFA World Cup 2018 was played between June 14, 2018 and July 15, 2018; the squads of nations in this competition can be used for supervision. For instance, [3] clusters the players into four distinct groups as goalkeepers, defenders, midfielders and attackers based on PCA reduced two dimensional attributes.

A new vectorial representation of soccer field positions is suggested in order to reflect the highly dynamic interchangeability in contemporary soccer and offer a versatile option for coaches who would use the system. Selected and engineered features are used in two dimensionality reduction and supervised classification methods. Our initial results on test squad confirms the validity of an approach for the integration of machine learning models and data science techniques using video games' expert attested metrics for human decision making process in sports.

II. Feature Engineering

We have selected 7 nations for training and evaluating our models : Argentina, Germany, Italy, France, Belgium, England, Spain and one nation for testing : Colombia. We have acquired the list of full squads of each nation in World Cup 2018 and carefully associated them with their profiles in FIFA'19 dataset.

A. Multiple Position Encoding

FIFA'19 dataset has 29 distinct categories for positions reflecting the most up to date tendencies in modern soccer. Each player is assigned to exactly one of these. Fig. 1 (a) shows the histogram of entire game dateset players' positions, whilst Fig. 1 (b) is only for the World Cup 2018 selected players of 8 nations considered in this paper. There exists a high degree of imbalance among positions, where some of them are significantly under represented.

In order to provide a more versatile framework for human decision makers while using the proposed approach, we have defined a binary vector, where each of 29 original positions can be represented as a unique code. Each index of the vector corresponds to *GK* (goalkeeper), *CntDef* (Center/Defender), *WngDef* (Wing/Defender), *CntMid* (Center/Midfielder), *WngMid* (Wing/Midfielder), *CntAtk* (Center/Attacker), *WngAtk* (Wing/Attacker) in order. Fig.1 illustrates the multipositional paradigm where each category in FIFA'19 belongs to one or multiple among the new seven positions. For instance, RCM (Right Center Midfielder) is defined both as a Wing Midfielder and Center Midfielder, thus encoded as 0001100.

The benefits of this type of encoding is three fold. Firstly, as number of categories are reduced the class imbalance problem is solved up to a reasonable degree. Next, the system provides a versatile option for users as they can query the algorithm with intended combinations. Lastly, the one hot encoded fashion representation allows the machine learning algorithms to associate player attributes with the meaningful and interpretable generalized positions.

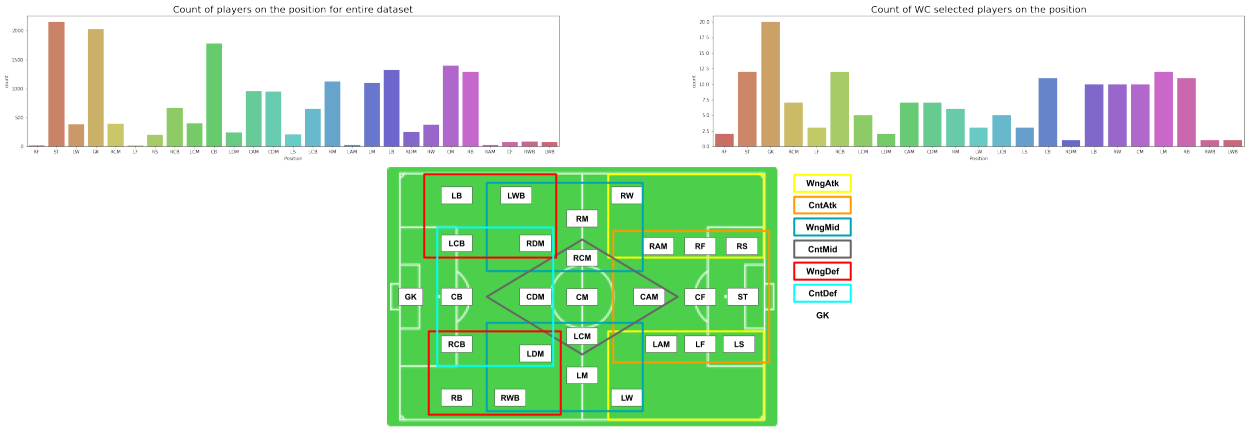

Fig. 1. (a) Number of players for each position category (29 available positions) in FIFA'19 dataset. (b) Number of players for each position among who has been selected for World Cup 2018 squads of the 8 considered nation. (c) Our suggested versatile multiple position encoding for the FIFA'19 categories.

B. Selected Attributes and Encoding

We have chosen 36 attributes from the dataset : *'Crossing', 'Finishing', 'HeadingAccuracy', 'ShortPassing', 'Volleys', 'Dribbling', 'Curve', 'FKAccuracy', 'LongPassing', 'BallControl', 'Acceleration', 'SprintSpeed', 'Agility', 'Reactions', 'Balance', 'ShotPower', 'Jumping', 'Stamina', 'Strength', 'LongShots', 'Aggression', 'Interceptions', 'Positioning', 'Vision', 'Penalties', 'Composure', 'Marking', 'StandingTackle', 'SlidingTackle', 'GKDiving', 'GKHandling', 'GKKicking', 'GKPositioning', 'GKReflexes'*.

These attributes are defined by FIFA'19 developers as integers from 0 to 100, proportional to player's real life skills. In addition, there exists an attribute called *Overall*, where it represents the general skill level of the player as the name suggests. However, we preferred not to include this parameter in the models as it indicates much more general assessment rather than a fine grade evaluation of a specific skill, which can introduce a high degree bias.

Selecting the players for a soccer squad is a ranking problem in its nature. We need to fill a limited number of places with the best possible players coming from a much larger set. This specific task is especially non-trivial as there generally exists numerous skilled players for the same position, where even small differences between players can hugely impact their chance to be selected for the national team. In according with these, we have decided to encode each of these 36 attributes with a player's ranking for that specific attribute among his nation's players.

At the end, each player is represented with a vector of 43 features; 7 for positions and 36 for attributes; normalized between 0 and 1.

III. Principal Component Analysis

Based on the aforementioned feature encoding, we have performed a 2 component Principal Component Analysis on the all players of the 7 considered nations in the training dataset. First 2 components explain the 72.1% of variance on the covariance matrix. The scores of players on two components are shown in Fig. 2. As expected, goalkeepers are clustered in a very distinct subspace (top left side). The players who were selected for their nation's world cup squads are highlighted with orange color.

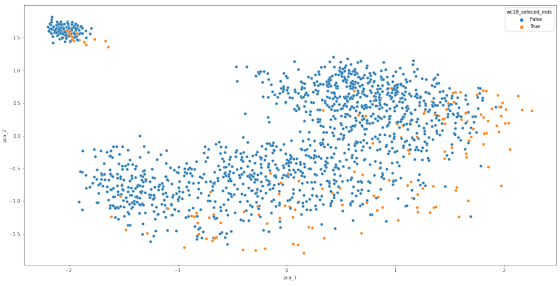

Fig. 2. PCA scores of all players in the 7 training nations on first 2 components (Total explained variance : 72.1%). The players selected for World Cup 2018 are highlighted. The goalkeepers intrinsically constitute a very distinct cluster (top left side of the graph) as expected. Players qualified for the world cup are clearly tended to have higher first and lower second component scores.

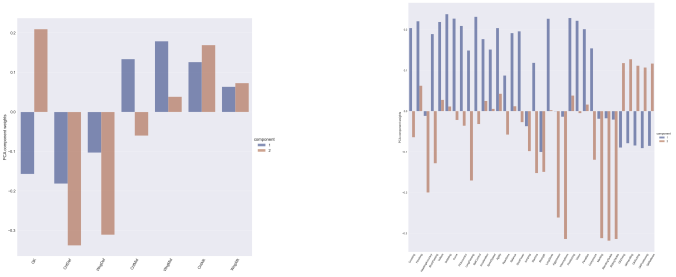

Fig. 3. (a) PCA weights of first 2 components of 7 postional features. (b) PCA weights of first 2 components of 36 player attributes.

Position related shifts on Fig. 2 can be verified by inspecting the directions and magnitudes of principal components on Fig. 3(a). For instance, the goalkeeper positional attribute has large

TABLE I
LOGISTIC REGRESSION CONFUSION MATRIX ON VALIDATION SET. (F1 SCORE : 0.76)

	WC'18 not selected	**WC'18 selected**
WC'18 selection not predicted	220	1
WC'18 selection predicted	75	21

negative weights on both components, shifting players top left of the plane. The similar effect is observed for goalkeeping related skills (Fig 3(b)). The players selected for national squads have relatively lower scores on both components compared to their competitors. On the other hand, parallel conclusions can be drawn for rest of the players.

IV. MACHINE LEARNING MODELS FOR SELECTING NATIONAL SQUADS

A. Overcoming High Degree of Class Imbalance

As the number of players who are qualified to represent their nations are much lower compared to rest, it is highly significant to use a class balancing algorithm before. Due to the very low number of world cup qualified players and the nature of the dataset, we have chosen the upsampling Synthetic Minority Over-sampling Technique (SMOTE) [4]. SMOTE generates pseudo minority class members by placing the new artificial datapoint on the feature plane based on one of the randomnly chosen k-nearest neighbor of a minority member, where the synthetic features are determined between these two instances.

B. Logistic Regression

Logistic Regression [5] inherently and naturally suits the task for this type of a qualification, where we would expect to have an exponential jump of selection probability even for a small increase in quality (middle of the S-curve). Similarly, after a threshold (sufficient skill level to qualify for national squad) the increase in features would not improve the selection probability a lot (tail of the S-curve).

We have used 20% of the 7 training countries' players for validation. Table 1 shows the results on the validation dataset, with an f1 score of 0.76.

Interpretability of machine learning algorithms both on data point (local interpretability) and dataset level (global interpretability) are paramount of interest; especially for a complicated, multi-faceted and vague task. In this context, the benefits of a detailed interpretation are multiple. First, the potential human users can back validate the decisions of acceptance or rejection of a particular player. Second, they can gain insights from overall interpretation results to further fine tune their decisions or construct a new kind of decision process even they do not use the algorithms directly.

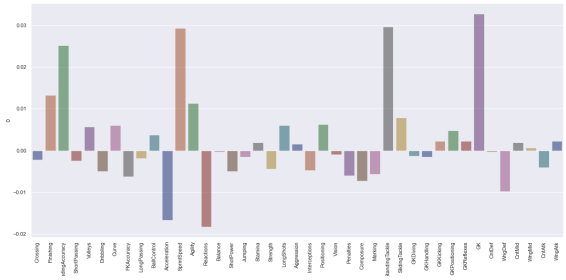

Fig. 4. Permutation importance of features for logistic regression.

Being a model agnostic algorithm, permutation importance is one of the most popular methods to interpret machine learning models globally [6]. The central idea is trivial, where the values of data points on a particular feature column are randomly shuffled numerous times and its effect on the overall classification or regression is measured. As the results deviates more for a certain feature, its level of importance is regarded as linearly higher.

We have evalulated all interpretation results on the validation set. Fig. 4 shows the permutation importance weights of attributes and positional encodings of the players. As it can be seen from the graph, for the positive contribution increasing the probability of selection of a player for national squad, as expected GK, the indicator that a player is a goalkeeper or not, appears as the most important factor; as goalkeepers are highly distinct by nature from rest of the players and a smaller minority. Among attributes increasing the chance of a player to be qualified for the national squad are $FinishingAccuracy, HandlingTackle$ and $SprintSpeed$. For the negative effect $Reactions$ and $Acceleration$ distincts themselves.

For local interpretability, meaning the explanation of a particular instance's classification result based on its input features, we preferred $LocalInterpretableModelAgnosticExplanations(LIME)$ algorithm [7]. LIME algorithm directly starts from instance based local explanation by training simpler machine learning models with much smaller traning datasets in the periphery of particular data points. By permutations of the dataset each local decision can be explained by the weighted combination of features.

Next, we have tested the algorithm's performance on the test nation, Colombia. For this purpose two intuitive approachs are followed : The probabilities of selection of the world cup squad of Colombia are measured and an alternative squad is proposed by choosing the highest probability of selection among Colombia excluding the World Cup squad for each player's same position. Fig. 6 and Fig. 7 shows the results, respectively.

TABLE II

TWO SIMILAR FOOTBALLERS ON THE VALIDATION SET WITH SAME AGE, SAME NATION AND SAME FIFA'19 OVERALL SCORE WHERE ONE IS SELECTED FOR NATIONAL SQUAD WHEREAS THE OTHER NOT. WE HAVE USED THIS PAIR OF PLAYERS TO EVALUATE THE LOCAL INTERPRETATIONS.

Name	Age	Pos.	Nation	WC'18	Overall	Team	
P. Jones	26	Cent. Mid.	Eng.	Qual.	79	Man. Utd.	
J. Wilshere	26	Cent. Def.	Eng.	Not Qual.	79	West Ham Utd.	

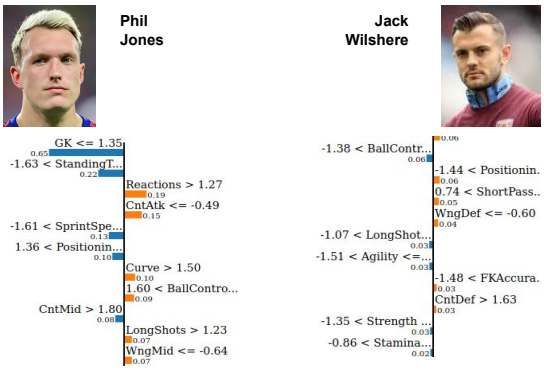

Fig. 5. Most important features explained by LIME on the decision of logistic regression algorithm for P. Jones and J. Wilshere.

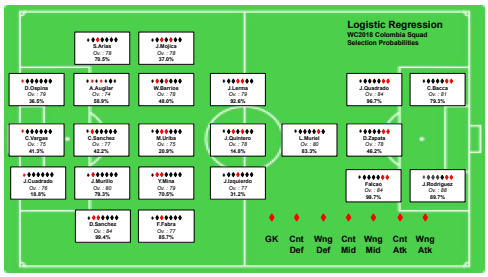

Fig. 6. Logistic regression selection probabilities of Colombia World Cup 2018 squad.

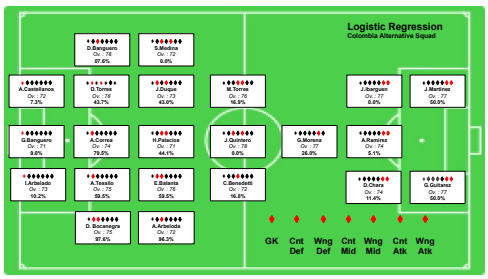

Fig. 7. Alternative squad proposed by the logistic regression algorithm for Colombia

TABLE III

K-NN CONFUSION MATRIX ON VALIDATION SET. (F1 SCORE : 0.54)

	WC'18 not selected	WC'18 selected
WC'18 selection not predicted	151	1
WC'18 selection predicted	144	21

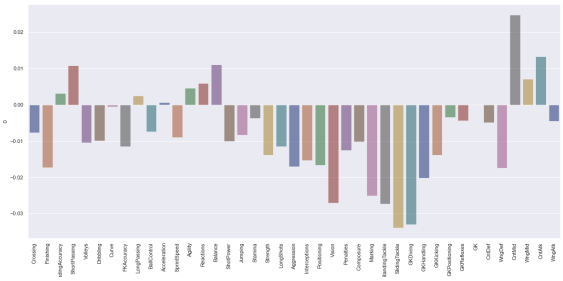

Fig. 8. Permutation importance of features for k-NN.

C. k-Nearest Neighbor

Another suitable algorithm considering the nature of the challenge would be the k-Nearest Neighbor (k-NN) [8]. Especially, for a task where the features (positions and attributes) of the data points (players) would be very close at the fringe of classification border. In other words, the player qualification criteria shall be highly fine in order to select the most fit ones. Thus, k-NN classification is needed to be tested for this particular challenge.

Via a cross-validation process, we have determined the most efficient neighborhood for the algorithm as 7. Table 3 shows the confusion matrix of the k-NN algorithm on the validation set. It is interesting the observe that the algorithm has performed much worse compared to logistic regression, with an average f1-score of 0.54. We have concluded that the particular reason for this is due to the fact that the mixture of positions and attributes as a single feature vector perturbed the efficient neighbor querying.

Fig. 8 shows the permuation feature importances for k-NN algorithm, where we observe a great deviation from the importances of logistic regression (Fig. 4), which further explains the significantly reduced accuracy. In order to further investigate, the LIME interpretations of the same pair of players are illustrated in Fig. 9. Finally, Fig. 10 and Fig. 11 shows the selection probability of real squad and the alternative squad proposed by the k-NN algorithm, respectively.

D. Linear Discriminant Analysis

One interesting approach for this specific task would be the utilization of a supervised discrimination algorithm, due to the fact that we would like to maximize the separation on feature plane as much as possible; as the qualification criteria

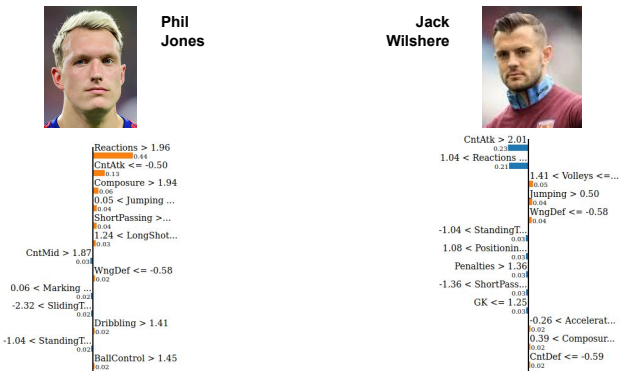

Fig. 9. Most important features explained by LIME on the decision of k-NN algorithm for P. Jones and J. Wilshere.

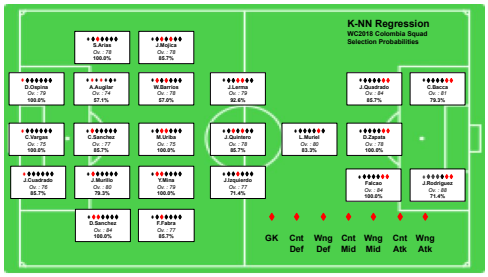

Fig. 10. k-NN selection probabilities of Colombia World Cup 2018 squad.

shall be very fine for similar talented players. We have used Linear Discriminant Analysis (LDA), also known as Fisher Discriminant Analysis [9]. However, as there exists only 2 classes the dimension of LDA plane is only one. This very fact may reduce the accuracy. Table 4 shows the confusion matrix on the validation set, where LDA achieved an average f1-score of 0.71.

We have also shown the weights of each feature on the single LDA discriminant component in Fig. 12. Interestingly, the permutation feature weights and the weighs of features on discriminant axis are not highly correlated, which may indicate an underlying inconsistency to explain to relatively poorer accuracy. As for the previous classification algorithms,

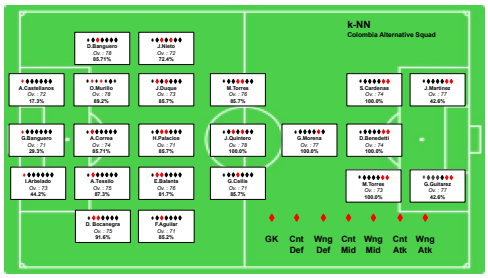

Fig. 11. Alternative squad proposed by the k-NN algorithm for Colombia

TABLE IV
LDA CONFUSION MATRIX ON VALIDATION SET. (F1 SCORE : 0.71)

	WC'18 not selected	**WC'18 selected**
WC'18 selection not predicted	204	1
WC'18 selection predicted	91	21

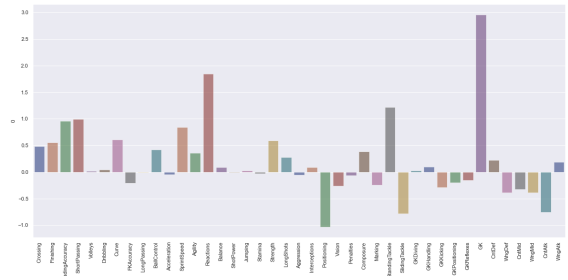

Fig. 12. Weights of features on single LDA discriminant component.

local interpretations on the same pair of English soccer players in validation dataset and computation of the selection probabilities of World Cup qualified players and the alternative squad selection for test nation, Colombia, are performed.

E. Neural Networks

Finally, we have tested the performance of a neural network architecture in order to introduce high degree of non-linearity and the exploit the capabilities of deep learning for explaining complex patterns. A 3-hidden layer architecture is used where each layer has 64,128 and 32 neurons in order. Neural networks had given the most accurate results, with an average f1-score of 0.89 for this particular architecture.

V. CONCLUSION

In this paper, we have used the open dataset of players from the video game FIFA'19 where certain soccer talent correlated player attributes are marked to investigate the possibility of a machine learning and data science backed system to select

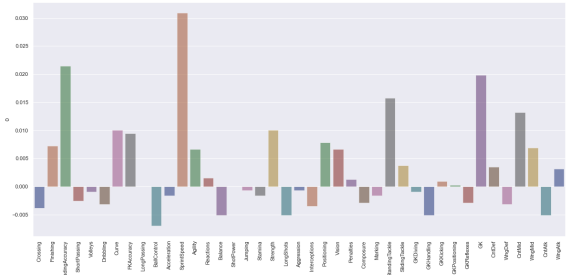

Fig. 13. Permutation importance of features for LDA.

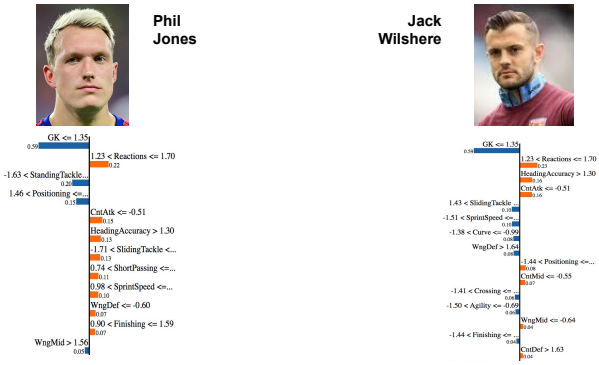

Fig. 14. Most important features explained by LIME on the decision of V algorithm for P. Jones and J. Wilshere.

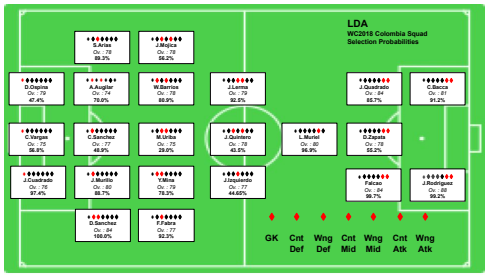

Fig. 15. LDA selection probabilities of Colombia World Cup 2018 squad.

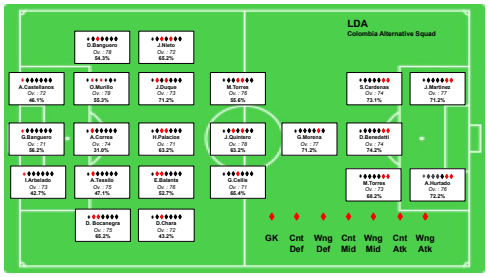

Fig. 16. Alternative squad proposed by the LDA algorithm for Colombia

TABLE V
NEURAL NETWORK CONFUSION MATRIX ON VALIDATION SET. (F1 SCORE : 0.89)

	WC'18 not selected	**WC'18 selected**
WC'18 selection not predicted	264	5
WC'18 selection predicted	31	17

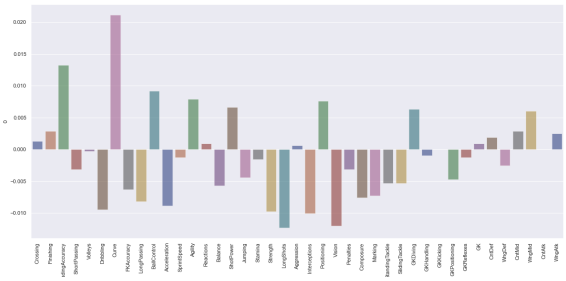

Fig. 17. Permutation importance of features for the neural network architecture.

Fig. 18. Most important features explained by LIME on the decision of neural network algorithm for P. Jones and J. Wilshere.

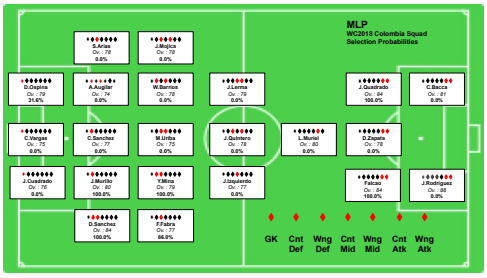

Fig. 19. Neural network architecture selection probabilities of Colombia World Cup 2018 squad.

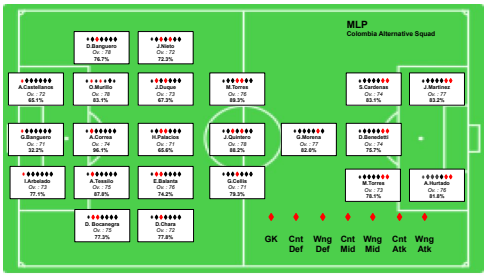

Fig. 20. Alternative squad proposed by the neural network algorithm for Colombia

players for national squads. For this purpose, the squads of 8 nations in World Cup 2018 (which coincides with the release date of the game) are used as a supervision tool. Several dimensionality reduction, visualisation, classification and machine learning interpretation techniques are used. Early results and observations indicate the promise for the approach. The central rationale of this work base itself on the fact that these the player attributes in these types of popular and high budget video games shall be as a result of a detailed and long investigative process of numerous professionals. Hence, integration of these valuable information with observed facts in a supervised manner (such as World Cup 2018 squads in this case) can introduce innovative and efficient artificial intelligence applications in sports. To the best of our knowledge, this paper is the first such an attempt in the literature.

REFERENCES

[1] E. Sports, "Fifa 2019," 2019.

[2] T. W. Neller, "Ai education matters: lessons from a kaggle click-through rate prediction competition," *AI Matters*, vol. 4, no. 2, pp. 5–7, 2018.

[3] C. Soto-Valero, "A gaussian mixture clustering model for characterizing football players using the ea sports' fifa video game system.[modelo basado en agrupamiento de mixturas gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos fifa de ea sports]." *RICYDE. Revista Internacional de Ciencias del Deporte. doi: 10.5232/ricyde*, vol. 13, no. 49, pp. 244–259, 2017.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[5] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[6] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[8] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient knn classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016.

[9] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," in *Institute for Signal and information Processing*, vol. 18, no. 1998, 1998, pp. 1–8.