Temporal Match Analysis and Recommending Substitutions in Live Soccer Games

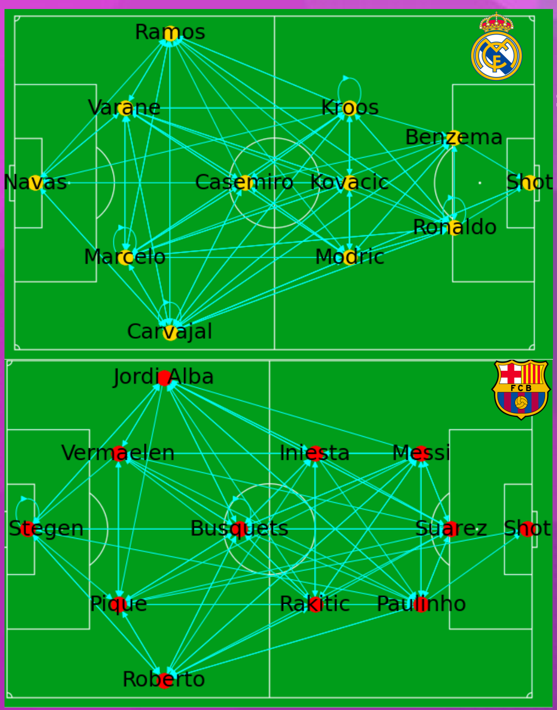
Yuval Berman, Dr. S. Mistry, Prof. A. Krishna, PhD

Introduction

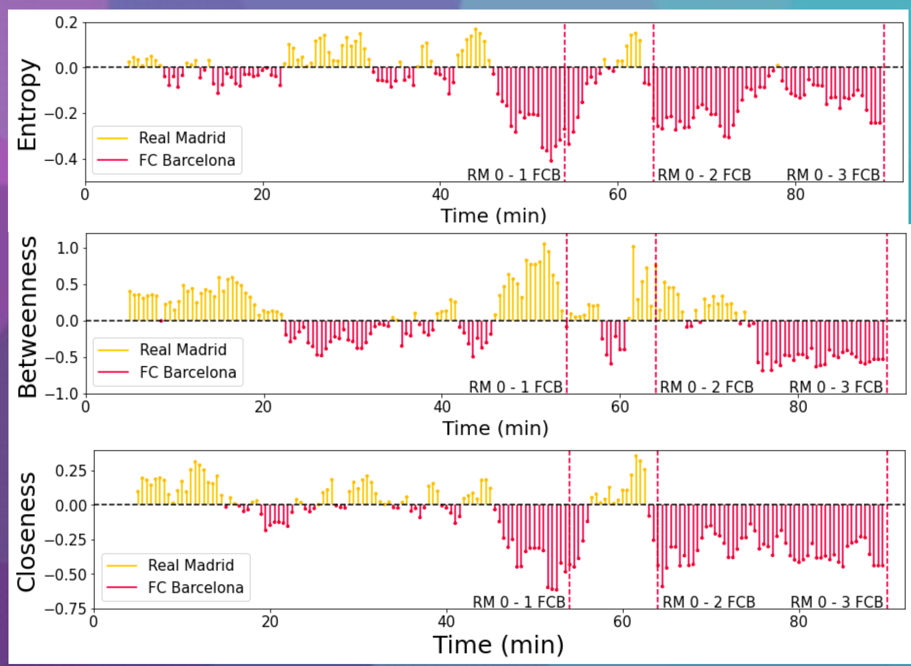
Uncertainty is an intrinsic feature in a live soccer game. Teams continuously adapt to the situation in the game, applying new tactical formations, substituting players, and so on. Due to this fluid nature, dynamic decision making is normally driven by human experts; the coaches.
As the live game generates large amounts of data in a very short period, effective analysis of dynamics in the game would provide valuable new insight and enable coaches to deploy data-backed, real time decisions to impact the game's outcome.

We investigate using network metrics and event logs data to correlate temporal network metrics to match prediction and use those results in order to create novel automated substitution models based on performance of players and wider team tactics.

Soccer Passing Networks and Match Outcome Correlation

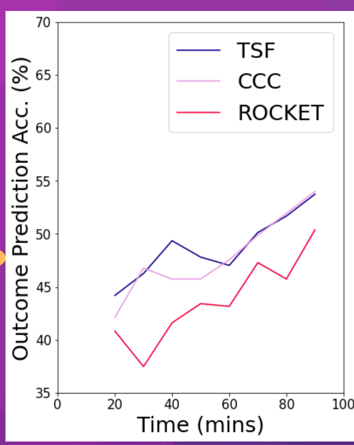


Created passing networks from soccer matches



FCB = FC Barcelona, RM = Real Madrid

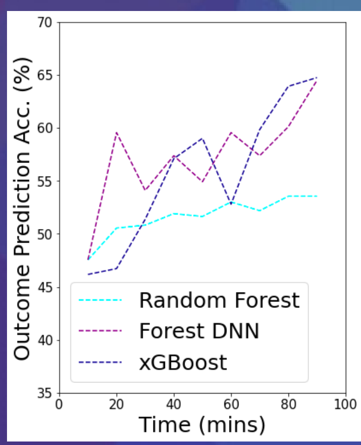
Calculated metrics of the network on temporal, 10 minute sliding windows- metrics such as betweenness and closeness centralities, and entropy (which is a measure of the unpredictability of the passing network)



TSF = Time Series Forest
CCC= Column Concatenator Classifier
ROCKET = RandOm Convolutional KErnel Transform

Used state-of-the-art multivariate timeseries classification algorithms to attempt to predict the outcome of a game from network metrics, giving a maximum accuracy of 54%.

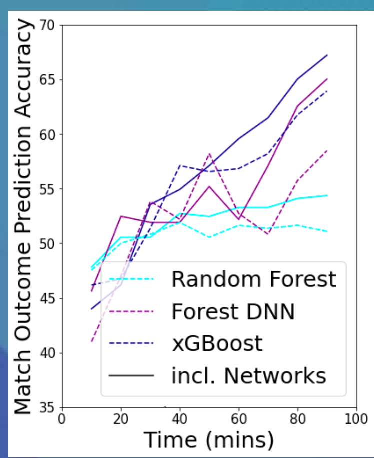
+



DNN = Deep Neural Network

So we decided to combine it with aggregate player features, developed by Makins et al. (2021), and use the state-of-the-art algorithms they used.

≈



We achieved an overall average improvement in all 3 algorithms:

1.4% in Random Forest
2.7% in Forest DNN
1.2% in xGBoost

Developing Substitution Models

Model 1- player performance and xGBoost

Method	Case Study- FCB vs RM
Find **poorest** performing player for your team based on **feature importance player** rating from **xGBoost**	**T. Vermaelen** had a player rating of **35** at the 60 min mark
Evaluate **current chances** of winning, drawing, losing the game based on **xGBoost predictions**.	**7% loss**, **24% draw**, **69% win** for **FCB** according to **xGBoost**
For **every** given **bench** player, evaluate the **chances of winning** the game if **that player had been playing** for the last 60 minutes instead .	There are **6 possible players** on the bench to choose from- try them all
Recommend substitution based on player which increases chance of winning (or decreases losing) by the **largest margin** according to xGBoost.	**A. Vidal** takes predictions to **3% L**, **15% D**, **82% W** => **Sub him on** for **Vermaelen**

Results

14% Of substitutions matched the naive model of best bench player for worst onfield player

8.7% Average win percentage increase in recommended substitution

71% Of players recommended to be subbed out actually were

78% Of players recommended to be subbed in actually were

6.1% Of recommended subs occurred exactly as suggested

Model 2 - Team Tactics Clustering

Method	Case Study- FCB vs RM
Normalise all features, and then **multiply** them by xGBoost's **feature importance** to create a **fingerprint** representation of a game	Done **before** match analysis
Use **kMeans** to **classify** all matches as one of 7 clusters in **N-dimensional** space	**Both** Barcelona and Real Madrid were in **cluster 2**
Find which cluster is **ideal** to be in **based on opponent's cluster** (i.e. which cluster leads to **most wins** against a particular cluster).	Being in **cluster 0** is most **ideal** against **cluster 2**
Iterating through **all on field players**, suggest the **substitution** which **decreases** the **distance** to the **desired cluster's centroid** by the **largest** amount.	To get as **close as possible** to **cluster 0**, substitute on **D. Suarez** for **S. Roberto**.

Results

0.1% Of substitutions matched the naive model of best bench player for worst onfield player

10% Average reduction in distance to desired cluster centroid

53% Of players recommended to be subbed out actually were

71% Of players recommended to be subbed on actually were

2.2% Of recommended subs occurred exactly as suggested

Conclusion

In this research we have shown an improvement in an existing analytics model by adding network metrics. We use the model to create novel substitution models that are based on state of the art Machine Learning algorithms such as xGBoost and kMeans. These models bear resemblance to actual substitutions made but provide the opportunity for coaches to make more data-backed decisions. This research can be further extended by making the models even more context aware- for example, a coach may choose to substitute a defender on for a forward when they have a lead they wish to protect.

References

For a bibliography and source code, please scan the following QR code, or click here.

Acknowledgements

The authors wish to acknowledge the support of the Pawsey SRC, CSIRO, and A. Backhaus, J. Makins, M. Speer and E. Matwiejew for providing support and mentorship during this research project.