Multi-Class ECG Feature Importance Rankings: Cardiologists vs. Algorithms

Philip J Aston[1], Temesgen Mehari[2], Alen Bosnjakovic[3], Peter M Harris[1], Ashish Sundar[1], Steven E Williams[4], Olaf Dössel[5], Axel Loewe[5], Claudia Nagel[5], Nils Strodthoff[6]

[1] National Physical Laboratory/University of Surrey, Teddington/Guildford, UK
[2] Heinrich Hertz Institute, Berlin, Germany
[3] Institute of Metrology of Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina
[4] University of Edinburgh/King's College London, Edinburgh/London, UK
[5] Karlsruhe Institute of Technology, Karlsruhe, Germany
[6] University of Oldenburg, Oldenburg, Germany

Abstract

Cardiologists have been using electrocardiogram (ECG) to diagnose a wide variety of heart conditions for many decades. Conversely, there are numerous algorithms that rank feature importance for a particular classification task. However, different algorithms often give quite different feature rankings. Therefore, we compared the feature importance rankings obtained by various algorithms with the features that cardiologists use for diagnosis.

1. Introduction

Cardiologists diagnose over 150 conditions from an electrocardiogram (ECG) based on interval, amplitude and timing features [1]. For each pathology, conditions on specific features are well documented [2]. Conversely, there are numerous algorithms that rank feature importance for a particular classification task. However, different algorithms often give quite different rankings and it is not clear whether one ranking is better than another. Therefore, we compared the feature importance ranking obtained by various algorithms with the features that cardiologists use for diagnosis. The advantage of this approach is that the important features for diagnosis have been derived from many years of clinical experience of each condition and provide a gold standard which the feature rankings of the algorithms can be compared with.

In a previous paper [3], we considered the feature rankings for binary classifications of Normal vs. First degree atrioventricular block (AV block), Normal vs. Complete right bundle branch block (RBBB) and Normal vs. Complete left bundle branch block (LBBB). However, such binary classifications are not very realistic since a cardiologist has to diagnose a condition out of the long list of possible conditions, which is a much more complex task. Also, it is possible that a simple binary classification of healthy vs. a particular pathology could be successfully achieved using only a subset of the full list of conditions. Thus, we now extend our previous work by considering a multi-class classification with all four of these classes combined, which is more realistic, although still not as complicated as the task undertaken by cardiologists.

With a multi-class problem, the aim is to give a positive classification for one class for each signal, which implies a negative classification for the other classes. This is basically a one vs. all other classes binary classification performed four times. So we will consider the feature rankings for these four binary classifications rather than for one classification of all four classes. This also makes comparison with our previous results easier.

2. Methods

2.1. ECG data

We constructed a dataset consisting of 500 12-lead ECGs for each of the classes Normal, AV block, RBBB and LBBB from the PTB-XL dataset [4], giving 2,000 records in total. Records with more than one of these four labels was excluded, so that each record had a single label of interest.

2.2. ECG features

The University of Glasgow 12-lead ECG analysis algorithm [5] was used to extract 772 features from each 10-second 12-lead ECG signal. From these features, we identified 117 which cardiologists would typically consider when making a diagnosis, which are listed in the Appendix of [3]. These features were derived from the 2,000 records described above. The final feature dataset did not contain any NaNs.

2.3. Pathologies

The three pathologies we consider are described in [3] and the important features used for a clinical diagnosis are:

- **AV block** (1 feature): PR interval
- **Right bundle branch block** (7 features): QRS duration, R amplitude in lead V1, R' amplitude in lead V1, S amplitude in leads I, aVL, V1 and V6
- **Left bundle branch block** (14 features): Q amplitude in lead V1, QRS duration, R amplitude in leads I, aVL, V5 and V6, R' amplitude in leads I, aVL, V5 and V6, S amplitude in leads I, aVL, V5 and V6

Note R' is the second positive wave in the QRS complex.

In this work, we are also considering Normal as a class to be identified, which occurs if a signal is not in one of the other three pathology classes, and so we use the union of the features for the three pathologies to be the important features for this class, giving 18 features in total.

An algorithm might rank a feature highly that correlates with an important feature. We looked for features that correlate with one of the important features with absolute value of the correlation coefficient ≥ 0.7. There are no correlating features for AV block, 5 correlating features for RBBB, 17 correlating features for LBBB and 17 correlating features for Normal.

2.4. Feature importance algorithms

We considered the same model-dependent and model-independent feature importance algorithms that we used in [3]. The model-dependent methods used are Random Forest (RF), Random Forest (permutation), Gaussian Processes (GP); in addition, SHAP [6] and LIME [7] were used with the models Random Forest (RF), XG Boost (XGB), Logistic Regression (LR) and Deep Networks (DN). The model-independent methods used are Chi-square test, Maximum Relevance - Minimum Redundancy (MRMR), Neighbourhood Component Analysis (NCA), ReliefF, and ROC AUC.

For the multi-class data, we want to positively identify a particular class out of the four and so we consider the binary classification problem of one of the four classes vs. the other three classes. A feature ranking is obtained for each class being singled out vs. the other classes, giving four rankings in all.

2.5. Scoring

To compare feature rankings with the set of important features for diagnosis of a pathology, we defined a score as a weighted sum of the positions of the top 5 features that are important features, and then normalised so that if all 5 top features are important then the score is 100 [3].

Table 1. The most common features in the top 5 for AV block for all 12 methods, which excludes LR and DN, and their ROC AUC values.

Feature	Frequency in the top 5 features	ROC AUC
PR interval	12	0.8915
QRS duration	12	0.7384
T' amplitude, lead I	5	0.5976
ST slope, lead I	3	0.5588
ST slope, lead V1	3	0.5275

3. Results

3.1. AV block

First degree AV block is defined by PR interval > 200ms [1] and so clearly there is only one important feature for diagnosis, namely the PR interval. Therefore we would naturally expect this to be high in the feature ranking.

All algorithms ranked the PR interval as the most important except for LR (SHAP) (5), LR (LIME) (8), DN (SHAP) (33), DN (LIME) (36) and GP (2). As we found previously [3], LR and DN do not rank the PR interval highly, with DN performing particularly poorly.

Next, we found the top 5 features for each method, but excluding LR and DN because of their poor results, to see if there is any commonality between them. For the 12 remaining methods, the frequency of the features in the top 5 is shown in Table 1 which of course includes the PR interval as the most common. Interestingly, the QRS duration was in the top 5 features for all methods. Also, we note that the last three features listed have ROC AUC values close to 0.5, and so are very poor discriminators on their own.

The only method with top 5 features matching those in Table 1 is RF, while RF (SHAP, LIME), XGB (SHAP, LIME) and Chi-square test all had 4 out of the 5 in their top 5. On the other hand, GP and MRMR only had the PR interval out of those listed in Table 1 in their top 5 features.

3.2. Right bundle branch block

There are 7 important features for RBBB and a further 5 features that correlate with one of these. We found the score for each method using the scoring algorithm described in Section 2.5 using the top 5 features of each ranking only. In Table 2, scores comparing the top 5 features for each method with both the important features and the important and correlating features are given. We see that four methods have all top 5 features either as important or correlating features, while Random forest (permutation) gives the best result when considering only the important features. LR does not have any of the important or correlating features in the top 5.

Table 2. RBBB top 5 scores for the important features only and the important features and correlating features.

Method	Top 5 score (important features only)	Top 5 score (incl. correlating features)
RF	60	**100**
RF (permutation)	**93**	**100**
RF (SHAP)	47	87
RF (LIME)	60	87
XGB (SHAP)	60	**100**
XGB (LIME)	60	**100**
LR (SHAP)	0	0
LR (LIME)	0	0
DN (SHAP)	40	67
DN (LIME)	47	73
GP	53	53
Chi-square test	53	87
MRMR	67	67
NCA	67	80
ReliefF	47	47
ROC AUC	47	87

Table 3. The most common features in the top 5 for RBBB for all 16 methods and their ROC AUC values.

Feature (Important/Correlating)	Frequency	ROC AUC
R' amplitude, lead V1 (I)	13	0.8009
S amplitude, lead I (I)	10	0.9403
ST slope, lead V1 (C)	7	0.9652
QRS duration (I)	6	0.7720
S amplitude, lead V2 (C)	6	0.8656

We next found the top 5 features for each method, and the 5 most common are shown in Table 3, which are all either important or correlating features and have high ROC AUC values. XGB (LIME) is the only method that had top 5 features matching those in Table 3, while RF, RF (SHAP) and XGB (SHAP) had 4 out of the 5 in their top 5.

3.3. Left bundle branch block

There are 14 important features for LBBB and a further 17 correlating features. We again found the score for each method using only the top 5 features of each ranking, which are shown in Table 4 for both the important features and the important and correlating features. In this case, the scores using only the important features were very poor for all methods. With the important and correlating features, the scores were much better with DN (LIME) giving the best result, while LR (LIME) was again the worst.

The 5 most common features in the top 5 features for all methods are listed in Table 5, which all have a high ROC AUC value. The ST slope in leads I and V6 were frequently highly ranked but are not important or correlating features. Only RF (LIME) had top 5 features that matched those in Table 5, while RF (SHAP), XGB (SHAP, LIME), Chi-square and ROC AUC had 4 of these in their top 5.

Table 4. LBBB top 5 scores for the important features only and the important features and correlating features.

Method	Top 5 score (important features only)	Top 5 score (incl. correlating features)
RF	0	80
RF (permutation)	27	80
RF (SHAP)	0	60
RF (LIME)	13	67
XGB (SHAP)	33	67
XGB (LIME)	**40**	73
LR (SHAP)	0	7
LR (LIME)	0	0
DN (SHAP)	27	87
DN (LIME)	13	**93**
GP	27	27
Chi-square test	0	60
MRMR	27	67
NCA	33	40
ReliefF	13	13
ROC AUC	0	67

Table 5. The most common features in the top 5 for LBBB for all 16 methods and their ROC AUC values.

Feature (Important/Correlating)	Frequency	ROC AUC
T' amplitude, lead V1 (C)	13	0.9747
ST slope, lead V6	8	0.9141
ST slope, lead I	7	0.9372
ST slope, lead V1 (C)	6	0.9584
Q amplitude, lead V1 (I)	6	0.7747

3.4. Normal

There are 18 important features for Normal and a further 17 correlating features. The scores for each method using the top 5 features of each ranking are shown in Table 6 for both the important features and the important and correlating features. In this case, RF and RF (SHAP) gave the best score for the important features, and Chi-square test was the best when the correlating features were included. ReliefF performed particularly poorly in this case.

Only the 4 most common features when considering the top 5 features for all methods are listed in Table 7, as three features all had a frequency of 4. Two of these are important features, with T' amplitude in leads I and V6 being neither important or correlating. RF (SHAP) and XGB (SHAP, LIME) had all of these features in their top 5.

Table 6. Normal top 5 scores for the important features only and the important features and correlating features.

Method	Top 5 score (important features only)	Top 5 score (incl. correlating features)
RF	**80**	87
RF (permutation)	60	60
RF (SHAP)	**80**	80
RF (LIME)	40	40
XGB (SHAP)	60	60
XGB (LIME)	60	60
LR (SHAP)	7	67
LR (LIME)	33	80
DN (SHAP)	33	40
DN (LIME)	47	80
GP	40	40
Chi-square test	60	**93**
MRMR	53	53
NCA	60	60
ReliefF	0	0
ROC AUC	33	33

Table 7. The most common features in the top 5 for Normal for all 16 methods and their ROC AUC values.

Feature	Frequency	ROC AUC
QRS duration (I)	11	0.8920
PR interval (I)	8	0.7494
T' amplitude, lead I	6	0.7896
T' amplitude, lead V6	6	0.7999

4. Conclusions

The best performing method was different for all the pathologies considered. However, RF performed well for all pathologies when correlating features were taken into account. Also RF (permutation) and RF (SHAP) did well for two out of RBBB, LBBB and Normal. DN (LIME) did well for LBBB and Normal, and reasonably well for RBBB, but ranked the PR interval as 36 for AV block, and so seems to perform either very well or very poorly. However, LR (SHAP, LIME) performed very poorly in all cases except for Normal, for which it did quite well.

For the model-independent methods, Chi-square test performed well for RBBB and Normal and reasonably well for LBBB, while ReliefF performed particularly poorly.

For AV block, the QRS duration was a highly ranked feature for all methods, even though it is not used in the diagnosis. For RBBB, the important features R' amplitude in lead V1 and the S amplitude in lead I were in the top 5 features for many methods and so seem particularly significant. LBBB is diagnosed entirely by changes in the QRS complex but four of the features in Table 5 are not associated with the QRS complex, However, the T' amplitude in lead V1 and ST slope in lead V1 both correlate with the Q amplitude in lead V1 (with correlation coefficients -0.8203 and -0.7700 respectively). Also, the S amplitude in leads I and V6 are important features, and changes in these amplitudes could affect the ST slope in leads I and V6. For Normal, QRS duration, PR interval and T' amplitude in lead V6 were all frequently highly ranked, and these are all significant for diagnosis of at least one of the pathologies. As Normal is diagnosed in this context as not a pathology, it makes sense that features from the other three cases are significant.

Acknowledgements

This project 18HLT07 MedalCare has received funding from the EMPIR programme co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme.

The authors acknowledge the support of the British Heart Foundation Centre for Research Excellence Award III (RE/18/5/34216). This research is supported by the British Heart Foundation (RE/18/5/34216). This research is part of the British Heart Foundation Centre for Research Excellence at The University of Edinburgh (RE/18/5/34216). SEW is supported by the British Heart Foundation (FS/20/26/34952).

References

[1] ECG Clinical Interpretation: A-Z by diagnosis. https://litfl.com/ecg-library/diagnosis/.

[2] Schuster HP, Trappe HJ. EKG-Kurs für Isabel. 7th edition edition. Georg Thieme Verlag, 2017.

[3] Mehari T, et al. ECG feature importance rankings: Cardiologists vs. algorithms. In preparation 2022;.

[4] Wagner P, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data 2020;7(1):154.

[5] Macfarlane P, Devine B, Clark E. The university of Glasgow (Uni-G) ECG analysis program. In Computers in Cardiology, 2005. IEEE, 2005; 451–454.

[6] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017; 4768–4777.

[7] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016; 1135–1144.

Address for correspondence:

Prof Philip J. Aston
National Physical Laboratory, Hampton Road
Teddington TW11 0LW, UK
Philip.Aston@npl.co.uk