Explaining Black-Box Automated Electrocardiogram Classification to Cardiologists

Derick M Oliveira, Antônio H Ribeiro, João A O Pedrosa, Gabriela M M Paixão, Antonio L Ribeiro, Wagner Meira Jr

Universidade Federal de Minas Gerais, Brazil

Abstract

In this work, we present a method to explain "end-to-end" electrocardiogram (ECG) signal classifiers, where the explanations were built along with seniors cardiologist to provide meaningful features to the final users. Our method focuses exclusively on automated ECG diagnosis and analyzes the explanation in terms of clinical accuracy for interpretability and robustness. The proposed method uses a noise-insertion strategy to quantify the impact of intervals and segments of the ECG signals on the automated classification outcome. An ECG segmentation method was applied to ECG tracings, to obtain: (1) Intervals, Segments and Axis; (2) Rate, and (3) Rhythm. Noise was added to the signal to disturb the ECG features in a realistic way. The method was tested using Monte Carlo simulation and the feature impact is estimated by the change in the model prediction averaged over 499 executions and a feature is defined as important if its mean value changes the result of the classifier. We demonstrate our method by explaining diagnoses generated by a deep convolutional neural network. The proposed method is particularly effective and useful for modern deep learning models that take raw data as input.

1. Introduction

Deep neural networks trained on large datasets have demonstrated the ability to provide accurate automated analysis of the electrocardiogram (ECG) [1, 2]. These models use the raw signal, a 12-lead ECG, as an input to the classifiers, being called "end-to-end" approaches. In such approaches, the model has the ability to learn complex patterns directly from the signal.

Classical methods for automated ECG analysis, such as the University of Glasgow ECG analysis program [3], employ a two-step approach: (1) Extract the main features of the ECG signal using traditional signal processing techniques; and (2) Uses these features as inputs to a classifier. In this approach, the models are built based on measures and features that are known by the cardiologists, making it easier to verify and to understand the algorithm decisions and, also, to identify sources of algorithmic mistakes. In "end-to-end" deep learning approaches such transparency is not possible.

There are still significant challenges using deep neural networks for ECG interpretation,including several case studies where the neural network learns to solve the task in unwanted ways or for which small perturbations may have a huge impact on the model prediction outcome [4]. In this work, we develop a method to understand the ECG "end-to-end" classifiers and report a close-to-cardiologist interpretation of the model output.

2. Methods

Our method quantifies the feature importance used in an "end-to-end" approach. In Figure 1 we illustrate each step of our method. As a case study, we applied it to analyze a deep learning ECG classification model [1]. We extract the features segmenting the ECG through the method Neurokit [5].

The deep convolutional neural network under analysis contains 9 convolutional layers and more than 6 million trainable parameters and is depicted in Figure 2. This neural network was trained using a dataset that consists of 2,322,513 ECG records from 1,676,384 different patients from 811 counties in the state of Minas Gerais/Brazil, acquired through the Telehealth Network of Minas Gerais (TNMG) [6]. Ninety-eight percent of the dataset was used for training and 2% for hyperparameter tuning. The resulting model is capable of classifying the 6 abnormalities in Figure 3: (1) 1st degree AtrioVentricular block (1dAVb); (2) Right Bundle Branch Block (RBBB); (3) Left Bundle Branch Block (LBBB); (4) Sinus Bradycardia (SB); (5) Atrial Fibrillation (AF); and (6) Sinus Tachycardia (ST).

The proposed method is depicted in Figure 1 and can be used to determine the interpretable features used by the model. It inserts noise into the raw signal that is fed into the model and computes the impact of each one of the features on the classifier outcome through a Monte Carlo ap-

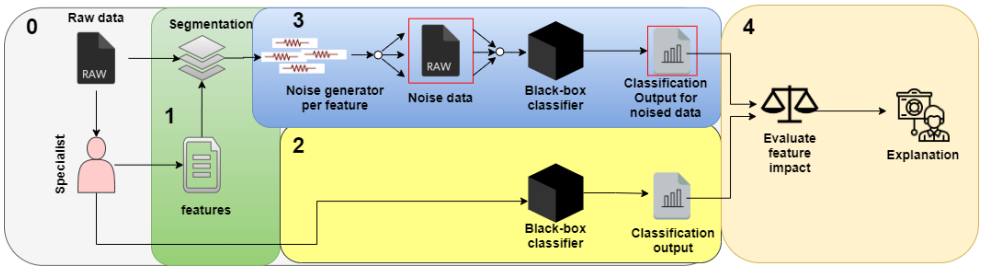

Figure 1. General procedure of the proposed explanation method. In this figure we assign each step a different color and number them to highlight their order.

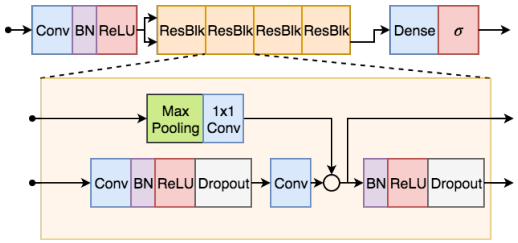

Figure 2. The uni-dimensional residual neural network architecture used for ECG classification. Reprinted from [1].

proximation, while assuring that the method did not introduce an outlier that may impact the outcome significantly. The method is summarized next:

1. Identification of the electrocardiogram waves. This step is known as ECG segmentation, several previous works [5, 8] addressed this issue. Our empirical analysis identified the algorithm [5] more suitable for our purposes.
2. Determination of the actual output of the sample given by the ECG automated classifier [1].
3. Noise insertion. For each segment of the ECG, we insert noise by changing its shape. The noise criteria were defined together with cardiologists.
4. Impact assessment. We evaluate the impact of each feature to the real outcome and the simulation [9].

The noise insertion procedure was designed along with a cardiologist to avoid creating an infeasible ECG signal. All perturbations have zero mean except the Axis feature, the standard deviations used for noise generation are:

- for derivations DI, DII, DIII, AVL, AVF:
  - QRS = 1.55 mV
  - T = 0.95 mV
- for derivations V1 - V6:
  - QRS = 1.70 mV
  - T = 1.10 mV
- for all derivations:
  - P = 0.30 mV
  - Duration = 400(ms)
  - Axis mean = 90° and std = 30°

In order to analyze our method efficacy, we assess the

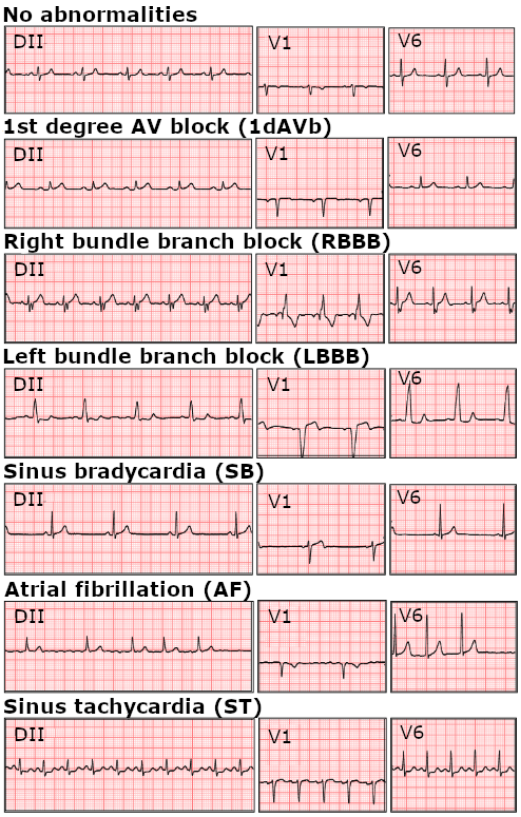

Figure 3. A list of all the abnormalities the model classifies. We show only 3 representative leads (DII, V1 and V6). Reprinted from [7].

explanation robustness. We define that a robust explanation must report similar feature impacts for all samples from the same class, that is, we did not expect to find significantly different explanations for different occurrences of the same disease. In order to assess robustness, we compute the frequency of each feature composing an explanation in the test set. We also tested five levels of noise in random intervals, the level of noise is generated from a normal distribution and variable variance, in this case 15%, 30%, 50%, 100%, and 200% of the signal variance. Finally we

Table 1. Relevance of interpretable features modified by random noise for the automatic diagnosis of each ECG class using the deep learning classifier.

		Measurements						
	#Exams	P Wave	PR Interval	QRS Complex	QT Interval	T Wave	Heart Rate	Rhythm
1dAVb	30	0.00	0.30	0.78	0.26	0.00	0.00	0.26
RBBB	38	0.00	0.00	0.52	0.00	0.21	0.48	0.64
LBBB	30	0.00	0.00	0.86	0.00	0.07	0.14	0.25
SB	18	0.00	0.00	0.00	0.00	0.00	0.86	1.00
AF	10	0.00	0.30	0.20	0.00	0.00	0.70	1.00
ST	38	0.00	0.00	0.03	0.00	0.00	0.69	1.00

analyze the correct clinical feature reported in the model, we evaluated with the multilabel AUC score [10, 11].

3. Results

In Table 1, we show how frequently features impact significantly each class. We observe that a common set of features explains each class, confirming the robustness of our explanations. We may also observe that there is a small variation on the results, e.g., the *QRS* complex explains the *1dAVb* in some cases. These errors are associated with segmentation inaccuracies and correlations among features, e.g., Heart rate affects the duration of the *QRS* complex. It is important to highlight that the feature *Heart rate* is a cofactor to the other features, and the variation on the ECG frequency modifies the duration of all segments. Thus, *Heart rate* can be used as duration criteria for other features.

As we may expect, the random features did not impact significantly the classifier performance, regardless the impact level employed. As a consequence, we did not show the random features results into Table 1. Notice that a feature does impact explanations if it affects a large number of tests, then being an explanation for the classifier as reported by our method.

The result presented by our model to the end user is the impact that each interpretable feature has on the classifier, depicted in Figure 4, as discussed in [9]. An explanation consists of both a visual and a textual explanation. Each visual explanation is a horizontal bar graph where each bar is associated with a feature, its length represents the impact, and the error bar at the right end of the colored bar represents the impact standard deviation considering samples from a given disease. Features that do not impact any of the samples are omitted. The red dotted line is the usage threshold of the feature, as proposed by [7]. The textual explanation is an automatically generated text that *reads* the visual explanation for the end user.

Figure 5 depicts the correct clinical features reported in our model, in particular that every class are above the 0.7 AUC score in the table, except for the *AF* diagnosis.

4. Conclusions and future works

In this work we propose and evaluate a method specialized for ECG end-to-end ECG classifiers, designed with features easily understandable by any cardiologist. Interpretable methods for automated classification for healthcare give to the doctor tools to be applied in real contexts, specially for cardiology, since any mistake may be fatal. Several works show how models may be biased [12, 13] and consistently make mistakes, even with high precision in test sets. Our main premise is that an interpretation must consist of features that are understandable by a specialist. As far as the authors known, this is the first work that proposes a method to generate explanations for an ECG classifier, using contextual features, that is, features that are understandable by any cardiologist. Our presented model is based on contextual features that support better explanations of the results of a black-box classifier to a physician.

In order to improve our method, we need to enhance the segmentation method. In particular, we expect that better segmentation will support more precise explanations and eliminate cases such as the *AF* classification example, where the segmentation found a *P* wave even if one criteria for AF is its absence.

We also want to perform a larger-scale experiment with cardiologists, by providing the correct ECG along with the classification explanation and measure the aggregated value of our method to real life applications.

Acknowledgement

The authors would like to thank FAPEMIG, CNPq and CAPES for their financial support. This work was also partially funded by projects MASWeb, EUBra-BIGSEA, INCT-IATS, INCT-Cyber, ATMOSPHERE and by the Google Research Awards for Latin America program.

References

[1] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MP, Andersson CR, Macfarlane PW, Wagner Jr M, et al. Automatic diagnosis of the 12-lead

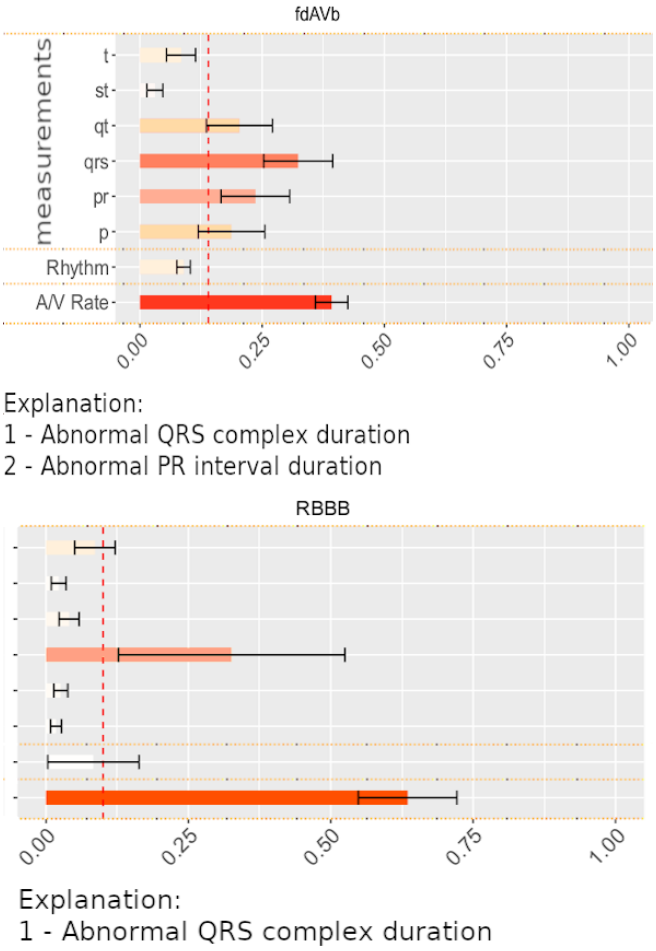

Figure 4. Explanations examples for 2 ECG disease. Each explanation has a visual and textual component.

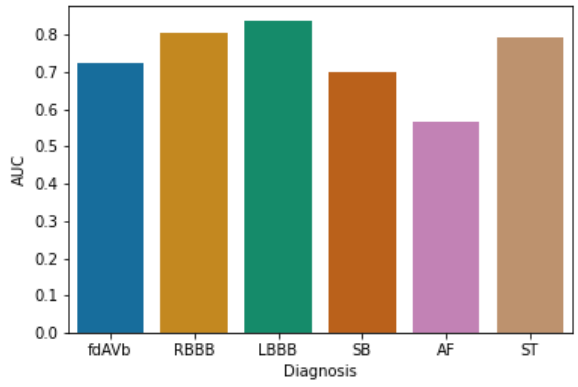

Figure 5. Multilabel AUC for the correct clinical feature usage in the model.

ecg using a deep neural network. Nature communications 2020;11(1):1–9.

[2] Smith SW, Walsh B, Grauer K, Wang K, Rapin J, Li J, Fennell W, Taboulet P. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. Journal of electrocardiology 2019;52:88–95.

[3] Macfarlane PW, Devine B, Clark E. The university of glasgow (Uni-G) ECG analysis program. In Computers in Cardiology. ISBN 0276-6574, 2005; 451–454.

[4] Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. arXiv14126572 December 2014;.

[5] Makowski D. Neurokit: A python toolbox for statistics and neurophysiological signal processing (eeg, eda, ecg, emg...). Memory and Cognition Lab Day 01 November Paris France 2016;.

[6] Alkmim MB, Figueira RM, Marcolino MS, Cardoso CS, Pena de Abreu M, Cunha LR, da Cunha DF, Antunes AP, Resende AGdA, Resende ES, Ribeiro ALP. Improving patient access to specialized health care: The Telehealth Network of Minas Gerais, Brazil. Bulletin of the World Health Organization May 2012;90(5):373–378. ISSN 1564-0604.

[7] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Macfarlane PW, Meira Jr. W, Schön TB, Ribeiro ALP. Automatic Diagnosis of the Short-Duration 12-Lead ECG using a Deep Neural Network: The CODE Study. arXiv April 2019;.

[8] Macfarlane P, van Oosterom A, Pahlm O, Kligfield P, Janse M, Camm J. Comprehensive electrocardiology. 978-1-84882-046-3. Springer, 2010.

[9] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems. 2017; 4765–4774.

[10] Hand DJ, Till RJ. A simple generalisation of the area under the roc curve for multiple class classification problems. Machine learning 2001;45(2):171–186.

[11] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research 2011;12:2825–2830.

[12] Lipton ZC. The doctor just won't accept that! arXiv preprint arXiv171108037 2017;.

[13] 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. ACM, 2016; 1135–1144.

Address for correspondence:

Derick M. Oliveira

Computer Science Department - UFMG, Rua Reitor Píres Albuquerque 190, Belo Horizonte, Minas Gerais, Brazil

derickmath@dcc.ufmg.br