Interpretability Analysis of Machine Learning Algorithms in the Detection of ST-Elevation Myocardial Infarction

Matteo Bodini, Massimo W Rivolta, Roberto Sassi

Dipartimento di Informatica “Giovanni Degli Antoni”, Università degli Studi di Milano, Milan, Italy

Abstract

Recent studies suggested that ST-Elevation Myocardial Infarction (STEMI) can be detected in the ECG relying on machine learning (ML) algorithms. However, most of ML algorithms lack of an interpretability analysis, since they do not provide any justification for their decisions.

In this study, we trained a Random Forest (RF) on the Physionet PTB database to automatically detect STEMI patients, considering 12-lead average templates as input. Then, we used the Local Interpretable Model-agnostic Explanations (LIME) method to highlight the input parts that mostly contributed to the detection. LIME interpretations were validated with the anatomical position of the myocardial infarction available within the dataset.

Experimental results showed that RF achieved a high test set accuracy (ranging from 0.84 to 0.92). However, LIME identified areas within QRS complexes as the most relevant ones for the RF decision, rather than in the ST segment as expected.

Our study suggests that, despite the test set accuracy, ML algorithms for STEMI classification, trained on small or unbalanced/biased populations, may rely on features which are not clinically significant. In this regard, interpretability algorithms like LIME may help in understanding possible pitfalls.

1. Introduction

ST-Elevation Myocardial Infarction (STEMI) is one of the leading causes of death for humans. Indeed, according to World Health Organization, an estimate of 7.3 million people died from heart attack [1]. Therefore, an accurate and early detection of STEMI is fundamental to increase the life expectancy and to improve the life quality.

The electrocardiogram (ECG) analysis is a crucial step in the diagnostic triage of patients suspected with STEMI. Clinical 12 leads ECG is acquired, and ST-segment Elevation (STE) is the marker most commonly linked to coronary occlusion. Further, STE persists on the ECG for several weeks after an acute infarct [2].

Being the ECG the most effective tool for prompt diagnosis of STEMI, as it is inexpensive, quickly performed, and rapidly available [2], to complement the role of physicians, computer-aided diagnosis (CAD) systems have been widely developed and have been gaining high attention worldwide.

Focusing on STEMI, researchers proposed CAD ECG classification systems based on machine learning (ML) algorithms. Standard ML algorithms make use of features extracted according to the medical expertise [3, 4]. However, since in other fields avoiding the step of feature engineering provided remarkable results, algorithms that automatically learn useful features from the ECG signal have been recently introduced [5–7].

Despite latest ML models for ECG classification seem to have reached the highest performance [7], they often lack of interperability. Thus, the goal of this study is to investigate on methodologies capable of providing interpretations of the model’s decisions.

In this study, we investigated on the interpretation of ML models trained to detect STEMI, using the Local Interpretable Model-agnostic Explanations (LIME) method [8]. In order to validate the interpretations provided by LIME, we compared them with the anatomical position of the myocardial infarction, known as part of the diagnostic report of the patient.

2. Materials and Methods

2.1. Dataset

ECG signals were collected from the Physikalisch Technische Bundesanstalt (PTB) database [9]. The database contained 549 acquisitions from 290 subjects (aged 17 to 87, mean 57.2; 81 women). ECG signals were sampled at 1 kHz, 16 bit resolution, and had variable length (the typical duration was two minutes). We considered only the 12 standard leads. For each ECG, diagnostic information were available: the PTB database contained 368 traces for 148 STEMI patients and 80 traces for 52 Healthy Control (HC) subjects. For STEMI, we selected only the 341 traces whose anatomical infarct location was annotated.

ISSN: 2325-887X DOI: 10.22489/CinC.2020.403

Table 1: Values of accuracy, precision, and recall for each infarct location: inputs are average QRST template (top), and average ST segment (bottom). The three highest RV measures are reported along with their respective lead.

Average QRST template

	Accuracy	Precision	Recall	1st lead / RV	2nd lead / RV	3rd lead / RV
Anterior	0.85	0.89	0.84	V1 / 0.24	V2 / 0.14	V4 / 0.14
Antero-lateral	0.84	0.81	0.77	V1 / 0.31	I / 0.29	V2 / 0.09
Antero-septal	0.92	0.89	0.90	I / 0.22	aVF / 0.22	V1 / 0.13
Inferior	0.88	0.87	0.85	II / 0.14	V1 / 0.11	V2 / 0.11
Infero-lateral	0.89	0.88	0.79	I / 0.19	II / 0.18	V1 / 0.15

Average ST segment

	Accuracy	Precision	Recall	1st lead / RV	2nd lead / RV	3rd lead / RV
Anterior	0.91	0.82	0.81	V1 / 0.29	V2 / 0.25	V3 / 0.17
Antero-lateral	0.89	0.83	0.79	I / 0.19	V1 / 0.17	V2 / 0.12
Antero-septal	0.86	0.80	0.90	V3 / 0.29	V1 / 0.21	V2 / 0.14
Inferior	0.87	0.81	0.82	II / 0.44	aVF / 0.17	III / 0.11
Infero-lateral	0.85	0.82	0.78	I / 0.28	V1 / 0.24	II / 0.09

2.2. Preprocessing and feature extraction

Selected ECG signals were filtered with a bandpass Butterworth filter (3rd order, zero phase, and pass-band: 0.67–30 Hz) to reduce powerline interference, baseline wandering and high frequency noise.

Beats were detected on the vector magnitude signal using the *gqrs* algorithm [9]. Beat positions were aligned on the R peak using the Woody algorithm applied to the vector magnitude [10]. Quality of signals was assessed computing the mean crosscorrelation with an average QRS template. An ECG trace was considered of good quality when such crosscorrelation was higher than 0.9 for every lead. After quality assessment, we obtained 44 HC traces, and for STEMI: 18 anterior, 15 antero-lateral, 34 antero-septal, 54 inferior, and 29 infero-lateral infarct traces. Other infarct locations were not considered since less than 10 traces were of good quality.

For each ECG, the average beat was computed for any lead. Then, two configurations were considered. First, we concatenated the average QRST segment of each lead in a single vector. The considered QRST segment spanned from 50 ms before the R peak to 150 ms after it, obtaining a feature vector of 2400 elements. Second, we concatenated the average ST segments only. Specifically, we considered segments from 50 ms after the R peak up to 150 ms after it, with a resulting feature vector of 1200 elements.

2.3. Random Forest training

We considered the Random Forest (RF) algorithm for our proof of concept. A different RF was trained for each of the two feature vectors (the concatenations of average beats or of average ST-segments) and for each of five specific infarct positions. The binary classification approach distinguished HC from STEMI subjects. For each RF, a dataset with features from HC and STEMI subjects was built. Then, a 70/30 training/test split was sampled with stratification (same proportion of classes was preserved).

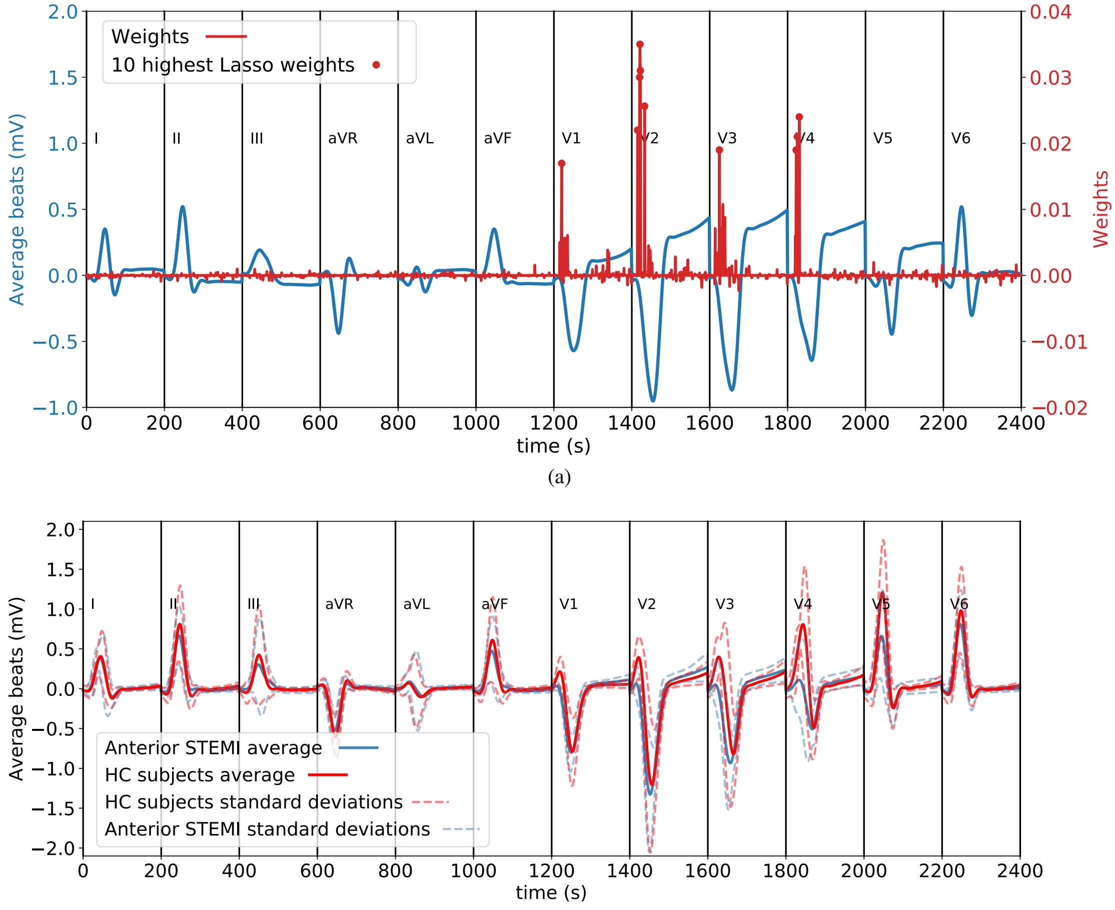
The hyperparameters of the models (*i.e.*, number of estimators, maximum number of leafs, maximum depth, minimum number of samples required to split nodes, and minimum number of samples required to be at a leaf) were tuned using a 10-fold cross validation applied to the training set. Specifically, we performed a random search by uniformly sampling 10^3 combinations in the range from 1 to 50 with a step of 10 for each parameter. In addition, Gini and Shannon entropies were tested as splitting criterion. The combination of hyperparameters that maximized the validation accuracy was then retained for the final training of the RF on the entire training set. Accuracy, precision and recall were finally evaluated on the test set.

2.4. LIME algorithm interpretations

LIME is a local surrogate explanation model, *i.e.*, it approximates the prediction about a new instance by using a simpler model. This simplified model is fitted on an artificial dataset created by probing the model "locally" on the new instance. LIME defines the explanation model as follows

$$\text{explanation}(\boldsymbol{x}) = \underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}(f, g, \pi), \tag{1}$$

where $\boldsymbol{x}$ is the new instance, g is a model within the family of possible explanation models G, and $\mathcal{L}$ is the loss function (for instance, the mean square error). The simpler model g is fitted by minimizing the loss $\mathcal{L}$ using an artifi-



(a)

(b)

Figure 1: (a) Lasso weights and average beat for record patient005/s0025lre. Red dots point to the 10 largest weights. (b) Population averaged QRST template ± standard deviation for HC and anterior STEMI, computed over the PTB dataset.

cial dataset created by sampling in a neighbourhood of the instance $\boldsymbol{x}$. A kernel function π defines the weight of each instance of the artificial dataset based on the distance with $\boldsymbol{x}$ (higher weights for lower distances).

For each of the trained RF f, we ran LIME as follows. Given an instance $\boldsymbol{x}$ belonging to the training set used to train f, we generated an artificial dataset by adding to $\boldsymbol{x}$ a white Gaussian noise, with zero mean and a standard deviation of 0.5 mV, to obtain $\lfloor 10^3/\text{training set size} \rfloor$ "artificial" samples ($\lfloor \cdot \rfloor$ is the floor function). Such artificial samples were weighted according to their distance to the instance $\boldsymbol{x}$ using an isotropic Gaussian kernel π, with 0.5 width. A linear model g was trained on the artificial training set with a loss function $\mathcal{L}$ defined as kernel weighted least square with L1 norm penalizer (Lasso). The λ parameter of the Lasso method was set to 10^{-4}. We repeated the procedure for each sample of the training set.

At the end of the procedure, a Lasso weight was available for each ECG sample in the feature vector. A large weight indicated high relevance of that sample for the classification of that subject. In order to also have a relevance measure (RV) for each lead, we computed the sum of the absolute value of the weights belonging to that lead, and normalized these 12 values with their sum. Finally, the average RV across the training set instances was computed.

3. Results

In Table 1, we report the accuracy, precision and recall quantified on the test set for the five RF models and two feature vectors considered. All metrics ranged from 0.77 to 0.92, hinting to a robust training of the RFs.

Regarding LIME explanations, and the relevance measure RV, we noticed that: 1) in the case features are average ST-segments, the highest RV value always refers to leads that anatomically pertains to the considered infarcts, for any RF model. 2) In the case features are average beats, for antero-septal (noticeable in V1 - V3 leads), and inferior (noticeable in II, III, and aVF leads) infarcts, the three highest RV values are instead referred to leads that are not anatomically related to the considered infarcts.

4. Discussion

Even if the latest ECG-based STEMI ML detection algorithms usually take raw, or almost raw, ECG signals as input, we performed a preprocessing phase and computed an average template representation. This procedure preserved the STE marker and reduced the noise, and proved to be efficient in terms of performance (Table 1). Further, we followed the recommendation of the International Guidelines for myocardial infarction identification [11] by using the standard twelve lead ECG, despite the majority of ML methods applied in this context did not rely on this standard setting [7].

While the two considered average template representations reached comparable performance on the test set, our analysis showed that in the case of antero-septal and inferior infarcts, the RF models using the QRST average template relied on leads which were not anatomically related to the considered infarct according to the guidelines (RV in Table 1). On the contrary, in the case of anterior, antero-lateral, and infero-lateral infarcts, the RF models relied on relevant leads for both feature representations.

Focusing on the anterior infarct, RV showed the highest relevance for the leads anatomically involved in STEMI for both feature representations. However, LIME showed that, for the QRST average template case, ECG samples mostly relevant for the classification were located on the QRS complex (Fig. 1a), rather than on the ST segments as recommended by the guidelines. This result might be explained by observing the high variability of the QRS complex between HC and STEMI (V1, V2 and V3 in Fig. 1b), and implicitly suggests a low inter-subject variability in the PTB dataset. LIME hinted that the considered RF might be unreliable when used in real scenarios. Similar results were obtained for the other kinds of infarcts. Another possible explanation for the relevance of the QRS complex might be due to the age difference between the HC subjects and STEMI patients (HC: 53 ± 17 vs STEMI: 67 ± 14), as QRS narrows while ageing [12].

To the best of our knowledge, only Strodthoff *et al.* [13] studied the interpretability of ML algorithms with the "gradient × input" method to explain the decisions of a Convolutional Neural Network for STEMI detection. Similarly to our results, they noticed that the most relevant segments for classification were located on the QRS complex.

To conclude, LIME may be considered a good ally in supporting researchers aiming to create automatic classifiers.

References

[1] Mendis S, Puska P, Norrving B, World Health Organization, World Heart Federation, World Stroke Organization (eds.). Global atlas on cardiovascular disease prevention and control. Geneva: World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization, 2011. ISBN 978-92-4-156437-3.

[2] Stellpflug SJ, Holger JS, Smith SW. What is the role of the ECG in ACS? In Brady WJ, Truwit JD (eds.), Critical Decisions in Emergency and Acute Care Electrocardiography. Wiley-Blackwell, 2009; 83–91.

[3] Arif M, Malagore IA, Afsar FA. Detection and localization of myocardial infarction using k-nearest neighbor classifier. J Med Syst 2010;36(1):279–289.

[4] Sharma LN, Tripathy RK, Dandapat S. Multiscale energy and eigenspace approach to detection and localization of myocardial infarction. IEEE Trans Biomed Eng 2015; 62(7):1827–1837.

[5] Rajan D, Thiagarajan JJ. A generative modeling approach to limited channel ecg classification. In Conf Proc IEEE Eng Med Biol Soc. 2018; 2571–2574.

[6] Liu W, Huang Q, Chang S, et al. Multiple-feature-branch convolutional neural network for myocardial infarction diagnosis using electrocardiogram. Biomed Signal Process Control 2018;45:22–32.

[7] Liu W, Wang F, Huang Q, et al. MFB-CBRNWN: A hybrid network for MI detection using 12-lead ECGs. IEEE J Biomed Health Inform 2020;24(2):503–514.

[8] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the NAACL-HLT 2016. Association for Computational Linguistics, 2016; 97–101.

[9] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[10] Woody CD. Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. Med Biol Eng 1967;5(6):539–554.

[11] Thygesen K, Alpert JS, Jaffe AS, et al. Fourth universal definition of myocardial infarction (2018). J Am Coll Cardiol 2018;72(18):2231–64.

[12] Levy D, Bailey JJ, Garrison RJ, et al. Electrocardiographic changes with advancing age. a cross-sectional study of the association of age with qrs axis, duration and voltage. J Electrocardiol 1987;20:44–47.

[13] Strodthoff N, Strodthoff C. Detecting and interpreting myocardial infarction using fully convolutional neural networks. Physiol Meas 2019;40(1):015001.

Address for correspondence:

Matteo Bodini
Dipartimento di Informatica "Giovanni Degli Antoni", Università degli Studi di Milano, Via Celoria 18, Milan 20133, Italy
matteo.bodini@unimi.it