Opening the black box: interpretability of machine learning algorithms in electrocardiography

Matteo Bodini†, Massimo W. Rivolta, Roberto Sassi

Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
† Corresponding Author: Matteo Bodini, Dipartimento di Informatica, Università degli Studi di Milano, Via Celoria 18, 20133, Milan, Italy. E-mail: matteo.bodini@unimi.it

Abstract

Recent studies suggested that cardiac abnormalities can be detected from the electrocardiogram (ECG) using deep machine learning (DL) models. However, most DL algorithms lacks of interpretability, since they do not provide any justification for their decisions.

In this study, we designed two new frameworks to interpret the classification results of DL algorithms trained for 12-lead ECG classification. The frameworks allow us to highlight not only the ECG samples that mostly contributed to the classification, but also which between the P-wave, QRS complex and T-wave, hereafter simply called "waves", were the most relevant for the diagnosis. The frameworks were designed to be compatible with any DL model, including the ones already trained. The frameworks were tested on a selected Deep Neural Network, trained on a publicly available dataset, to automatically classify 24 cardiac abnormalities from 12-lead ECG signals.

Experimental results showed that the frameworks were able to detect the most relevant ECG waves contributing to the classification. Often the network relied on portions of the ECG which are also considered by cardiologist to detect the same cardiac abnormalities, but this was not always the case.

In conclusion, the proposed frameworks may unveil whether the network relies on features which are clinically significant for the detection of cardiac abnormalities from 12-lead ECG signals, and thus increasing the trust in the DL models.

1 Introduction

Cardiovascular diseases (CVDs) stand as the main cause of mortality worldwide. An estimated 17.9 million people die each year from CVDs: 31% of all deaths globally [1]. Early diagnosis and prompt treatment are of paramount importance for people with CVDs or who are at high cardiovascular risk [2]. The electrocardiogram (ECG) is considered as one of the most important clinical tools in the detection and diagnosis of CVDs since it is non-expensive, non-invasive, and it can be quickly performed [3]. Usually, a physician provides the interpretation of the ECG. However, the interpretation process is time-consuming and it requires a high degree of training [4]. It is then not surprising that the first attempts to build computer programs for the automatic ECG interpretation are dated back at the end of the 1950s [5]. Such computer programs have had a large impact on electrocardiography itself. Worldwide, millions of ECGs are recorded every year, where the majority is automatically analyzed and interpreted [6]. Traditional ECG analysis programs are often built *translating* into the machine the interpretation rules developed within the standard practice of physicians [6].

In the last decade, we saw the development of a completely new approach, called deep learning (DL), a research field belonging to the machine learning (ML) domain, where computers efficiently learn how to make automatic decisions. Its success was mainly due to the availability of large databases and new high-performance computing methods [7]. DL obtained stunning results in speech recognition, image classification, and language translation, some of them at human-level performance [7]. The large impact that DL had in these domains has motivated the investigation of such methodologies for automatic classification of ECGs [8].

1.1 Machine learning in electrocardiography

As in other applicative fields, supervised and unsupervised ML methods were applied in electrocardiography [8]. Under the supervised ML domain, researchers focused on cardiac abnormality identification by training algorithms to learn a classification function from a set of labeled data (*e.g.* [9, 10]). Most typical algorithms include Support Vector Machine (SVM), k-Nearest Neighbour (KNN), Decision Trees (DT), Random Forests (RF), and Artificial Neural Networks (ANNs), including DL models such as Convolutional Neural Networks (CNNs) [11].

A brief survey on automatic ECG classification can be found in [8]. On the other hand, unsupervised learning was leveraged to uncover hidden relationships in the data without any clinical knowledge (*e.g.* the labels) [11]. For example, algorithms such as k-means, clustering trees and distance metric learning aimed to cluster data, *i.e.* to group data (typically subjects) with a certain degree of similarity in the same set. Another common task in unsupervised learning was dimensionality reduction. In this context, Principal Component Analysis was among the most frequently used algorithms, where data was projected on a feature space whose dimension was less than the original data space, retaining the maximum variance within it [12].

Most of supervised and unsupervised ML algorithms are fed with handcrafted feature vectors for the classification task to accomplish. For example, QRS duration, QT intervals, RR intervals, morphological features and frequency domain features are among the most common ones [11].

Recently, with the advent of DL, the approach for tackling automatic classification problems moved from the calculation of handcrafted features to an innovative learning strategy called end-to-end. Such strategy lets the classifier to learn the relevant features for the classification directly from the raw data (or slightly preprocessed). The most common DL algorithm is represented by an ANN with multiple layers performing specific operations (such as convolutional filters in CNN and subsampling). This class of models is called Deep Neural Network (DNN).

DNNs models may learn the optimal features for the specific classification task, thus likely outperforming ML algorithms that are fed with handcrafted features (however, it is worth noting that the superiority of DNNs with respect to ML algorithms for ECG classification has not been proved yet). These models have already shown promising results in identifying abnormal heart rhythms. For instance, Hannun *et al.* [9] implemented a deep CNN and claimed cardiologist-level performance, even if results were provided on a limited set of classes (representing rhythm-based and morphology-based related cardiac abnormalities). Ribeiro *et al.* [10] implemented a residual neural network with the same aim of classifying cardiac abnormalities among six classes, and it showed remarkable performance.

1.2 The black box of deep learning and objective of the study

The main advantage of DNN models is represented by the optimal feature representation achieved after the training phase. Indeed, the capability of automatic learning the relevant features is due to the large amount of parameters that these models contain (in the order of millions). However, with such large amount of parameters, the decisions taken by the models become difficult (even impossible in certain situations) to interpret. Consequently, automatic ECG classifications performed using DNNs result difficult to be associated with a physiological interpretation. The perceived lack of interpretability gives the feeling of dealing with *black boxes*: even if computer scientists can understand the architecture of the networks, the process by which the models perform the classification can be inscrutable to humans, limiting the trust in them, and thus hindering the acceptance in the healthcare community.

In practice, DNNs are only highly capable to catch patterns and regularities from the data to perform their classifications. In case of good performance, we may expect that, after the training phase, the network learned patterns meaningful for the underlying domain, *e.g.* the network learned patterns correlated with typical ECG markers that physicians search on the ECG to provide an interpretation. However, there is no guarantee that the network makes use of such meaningful patterns to achieve the classification. For example, in our previous study, we found that even with a simpler model, a RF, it was possible to achieve a high classification accuracy for automatic detection of myocardial infarction, even though the RF was not relying its decisions on the ECG segments reported in the international guidelines for ECG interpretation [13].

In order to *open* the black box and understand how DL methods take decisions, researchers introduced different approaches to interpret and explain the model outcomes, thus creating a new line of scientific research usually called "Explainable AI". These approaches were mostly developed in the Computer Vision domain, where researchers needed to understand which pixels of the input image were utilized for the classification [14]. A brief overview on such methodologies is reported in sec. 11.3. It is worth mentioning that there is still no agreement within the ML community on the definition of the terms "interpretability" and "explainability" [15, 16]. Even if several authors attempted to distinguish between them, most use the two terms interchangeably. Although a clear universal definition is still not available, for the sake of convenience, we will refer here to the distinction reported in Guidotti et al. [15], where interpretability is defined as the ability to explain, or to present, the reason behind the decision of a ML model in an understandable terms to a human.

In this study, we introduce two interpretability frameworks, specifically designed for DNNs trained for the ECG classification task, that let us to inspect the decision of DNNs, unveiling which waves of the input ECG were most relevant to the final classification. In this work, we refer to the P-wave, QRS complex and T-wave composing the ECG beat as simply "waves". The rationale behind the development of new interpretability frameworks in this context relies on the fact that the evaluation of the interpretation itself results challenging with the current interpretability methodologies. Indeed, such methodologies can highlight the most important samples of the ECG contributing to the classification. However, understanding whether such samples result

meaningful for the cardiac abnormality to detect is not currently handled. For example, a single heartbeat on 1-lead ECG sampled at 1000 Hz has approximately 600 samples, and each sample has a weight on the classification. However, for a given cardiac abnormality, only some ECG samples must result useful for the classification, based on prior knowledge from electrocardiography. In order to address such issue, we propose to combine two modules. The first one will provide interpretability by using two state-of-the-art techniques. The second one will assess whether the most important samples are matching those expected to be affected by the cardiac abnormality, thus including the domain knowledge. Differently from the Computer Vision domain, this approach results feasible because segmenting ECG signals is rather easy in the considered context, with very well-established and validated algorithms already available.

1.3 Background on interpretability methods and their application to automatic ECG classification

A taxonomy for methods to open "black box" ML algorithms is provided in [15]. At a high level, they differentiate between two strategies: i) *post hoc* interpretability; and ii) "transparent box design". The former provides an interpretation of a choice taken by a black box decision maker. On the other hand, the latter aims to develop an interpretable predictor model. This is achieved by considering that a small set of existing ML models is recognized to be interpretable, *i.e.* decision tree, decision rules, and linear models [15], because they are usually considered easily understandable and interpretable for humans [15, 14]. We will focus this background section only on the first strategy since DNNs do not fall in the transparent box design.

According to [15], *post hoc* interpretability methods can be divided in the three following classes:

- "Model explanation" methods aim to understand the overall logic within the black box. These approaches usually build a globally interpretable predictor that is able to mimic the behavior of a black box.
- "Outcome explanation" methods consist in providing an explanation for the outcome of the black box on a specific instance.
- "Model inspection" methods do not provide a comprehensible predictor, but a representation (for instance, a visual one) for understanding certain specific property of the black box model.

Another taxomony for *post hoc* interpretability methods is based on the following characteristics [15, 14]:

- Model-specific or model-agnostic. Model-specific methods are specifically designed for a class of models, *e.g.* for neural networks. On the other hand, model-agnostic methods can be used for any ML model, and they are applied after the training step. These methods do not access to model internals, *e.g.* the model architecture and its weights for neural networks.
- Local or global. A global method is capable to interpret the entire behavior of the model while a local one only interprets an individual prediction at a time. Most of the available methods are local, since global interpretability usually implies a simple structure in the design of the ML model.
- Type of explanator used. It refers to the way the explanator provides its interpretations. For instance, the feature importance method returns the weight of the features used by the black box, decision trees provide rules, and activation maximization techniques inspect DNNs to find the artificial neurons fundamental for the classification.

In the context of automatic ECG classification, several works investigated the interpretability of ML algorithms for the classification of arrhythmias. The works reported below limit their scope to the *post hoc* methods, and they all provide local and model-specific interpretability techniques, while the most can be framed within the outcome explanation methods.

Vijayarangan *et al.* [17] proposed two methods to provide interpretability for ECG classification. In the first approach, they applied Gradient-weighted Class Activation Maps (Grad-CAM), a type of outcome explanation method to visualize the saliency maps of a CNN model. In the second approach, saliency maps were derived by learning the input deletion mask for the LSTM model. Yao *et al.* [18] introduced a CNN with an outcome explanation method based on attention mechanism with the aim of adding interpretability to the model. The attention mechanism outputs the signal segment of interest along with classification result. Mousavi *et al.* [19] proposed an end-to-end hierarchical attention mechanism model based on attention neural networks: the model is composed of three parts in which each part contains a stacked bidirectional recurrent neural networks, followed by an attention model, capable of providing multi-level interpretability considering segments within the heartbeat, the whole heartbeat and the combination of all the heartbeats. Baalman et al. [20] built a feedforward neural network to classify ECGs within atrial fibrillation and sinus rhythm along with an attention mechanism. Through the attention mechanism they built a heat map on the input signal to show the areas of

the ECG used by the classifier to come to the correct classification. Hong *et al.* [21] developed a similar multi-level attention model by extracting multilevel (beat-, rhythm- and frequency-level) domain knowledge features. Goodfellow *et al.* [22] built a CNN with an outcome explanation method based on class activation mappings to understand which areas of the waveform the model was focusing on when making a classification. Han *et al.* [23] developed a method based on model inspection to build adversarial examples for ECG tracings that are invisible to human expert evaluation and showed that a DNN for arrhythmia detection from single-lead ECG is vulnerable to this type of attack. Finally, Strodthoff *et al.* [24] designed a CNN and applied the outcome explanation interpretability method named "gradient × input" to identify which part of the input were the most relevant to the final classification.

2 Methods

2.1 Interpretability framework

We introduce two new frameworks to interpret the classification results of DL algorithms trained for the task of multi-label 12-lead ECG classification, that takes ECG signals in input. Both the frameworks comprises of two modules. The first one relies on two *post hoc*, local, and model specific interpretability algorithms, whose output show the contribution of each ECG sample to the final classification. The second one segments the ECG signal using validated algorithms (see sec. 22.4) and quantifies whether the ECG samples most relevant for classification belong to the ECG waves which the domain knowledge links to the cardiac abnormality.

2.1.1 Framework 1: occlusion method

The first framework uses an occlusion-based methodology [25]. It is an inspection technique originally designed to interpret DNN for image classification. We adapted this technique to interpret ECG signals. Given a 12-lead ECG signal, an occlusion is performed by setting to zero a specific interval of the signal (*e.g.* all the samples in the T-wave are set to zero). The classifier is then run to compute the output class after applying the occlusion. The occlusion of the segments that leads to a relevant change in the classification, with respect to the ground truth labels, points to those segments which are important for the final classification.

In particular, in this work, the occlusion was performed by setting to zero, for each beat and lead, all the samples relative to either the P-wave, QRS complex, or T-wave. Then, we calculated the percentage variation in the model output of a given class after the occlusion of the three waves. Finally, we normalized these three values for their sum. We termed these three normalized quantities as "relevance measures" (RV) [13], that we formally defined for each ECG signal $\mathbf{x}$ as

$$\mathrm{RV}^{\mathrm{F1}}_{c,w}(\mathbf{x}) = \frac{|P_c(\mathbf{z}_w \circ \mathbf{x}) - P_c(\mathbf{x})|}{\sum_{i\in\{\mathrm{P,QRS,T}\}} |P_c(\mathbf{z}_i \circ \mathbf{x}) - P_c(\mathbf{x})|} \tag{1}$$

where $\mathbf{z}_w$ is a mask vector containing ones in the position of the indices belonging to the wave w in the ECG signal (*i.e.* P, QRS or T-wave), $\circ$ is the element-wise product, and $P_c(\mathbf{x})$ is the probability estimated by the network for the cardiac abnormality (or class).

2.1.2 Framework 2: saliency maps

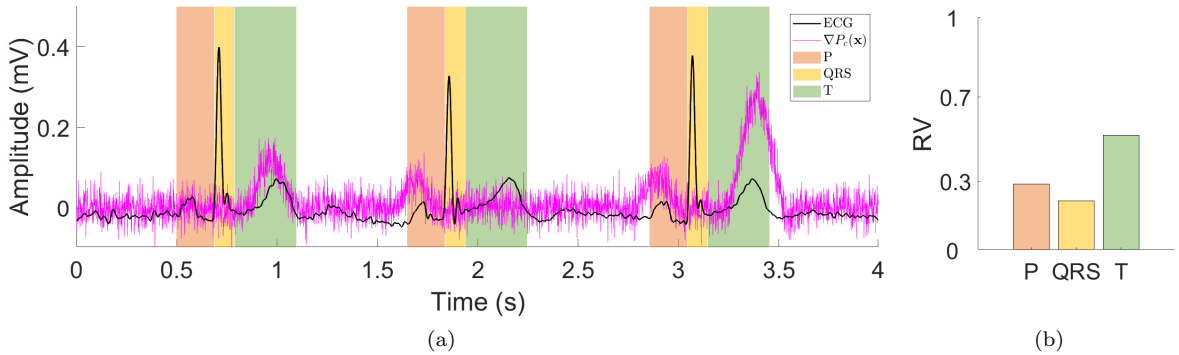
The second framework implements saliency maps in the first module. The saliency map is an outcome explanation approach in which the DNN output $P_c(\mathbf{x})$ for the class c and the ECG input $\mathbf{x}$ is approximated with a first-order Taylor expansion in the form $P_c(\mathbf{x}) \approx \mathbf{m}^\intercal\mathbf{x} + q$, where q is a scalar quantity and $\mathbf{m} = \nabla P_c(\mathbf{x})$ is the weight vector [26]. The latter stands as the explanation for the classification of the underlying DNN: the largest entries of $\mathbf{m}$ are associated to the samples that are the most relevant in the final classification.

The second framework quantifies the RV as follows. For each segmented beat, we computed the sum of the absolute value of the weights belonging to any of the three ECG waves. Then, we averaged these three values across the beats of a given signal. The three values, obtained for any ECG signal, were normalized on the length of the ECG waves (190 ms, 100 ms and 310 ms for the P, QRS and T, respectively) and to have unit sum. Differently from the first framework, this approach weights the RV based on value of the entries of the vector $\mathbf{m}$. The formulation of RV for the input signal $\mathbf{x}$ in this second framework was

$$\mathrm{RV}^{\mathrm{F2}}_{c,w}(\mathbf{x}) = \frac{(\mathbf{z}_w^\intercal\mathbf{z}_w)^{-1}\mathbf{z}_w^\intercal|\nabla P_c(\mathbf{x})|}{\sum_{i\in\{\mathrm{P,QRS,T}\}}(\mathbf{z}_i^\intercal\mathbf{z}_i)^{-1}\mathbf{z}_i^\intercal|\nabla P_c(\mathbf{x})|} \tag{2}$$

where $\mathbf{z}_w$ is the same mask vector of eq. (1) and $|\nabla P_c(\mathbf{x})|$ is the absolute value of the gradient of $P_c(\mathbf{x})$ (column vector).

Figure 1 reports an example of interpretation provided by this second framework on a single-lead ECG signal. In particular, it shows the output of both modules to finally produce the RV value.



(a) (b)

Figure 1: Example of interpretation for a single-lead 5 s ECG signal by means of Framework 2. For a given ECG signal (black line), the output of the first module of the framework is reported, for each ECG sample, in magenta (a). The shaded boxes represent the segmentation of each ECG wave. Then, the second module computes the RV of the P-wave, QRS complex and T-wave (b). In this example, the T-wave was linked to the largest relevance measure, as shown in (b).

2.2 The experiments

The two versions of our framework were tested on a ML model trained for the multi-class classification of 12-lead ECGs, specifically developed for the Physionet/Computing in Cardiology Challenge 2020. In particular, we selected the deep machine learning model that won the official phase of the challenge, namely the BUTTeam network [27]. The main rationale behind the selection of this model was that the good performance of the network was certified by the official ranking of all submitted models, and such ranking was built considering an unseen test set.

The model is a CNN architecture based on a residual neural network. We did not retrain the model from scratch, but we instead used the pre-trained model provided by the authors (`https://github.com/tomasvicar/BUTTeam`). The performance of the network on the entire dataset (see sec. 22.3) was quantified using the macro measures AUC, AUPRC, F1 and Accuracy. Macro measures refer to the average of all values of a given measure determined for each class. The AUC is computed as the area under the curve defined by true positive rate (TPR)/sensitivity and true negative rate (TNR)/specificity. Similarly, the AUPRC is computed as the area under the curve depicted by TPR/sensitivity and PPV/precision. F1 is computed as the product of twice PPV/precision and TPR/sensitivity normalized by their sum. Accuracy is the proportion of correctly classified ECG over the total number of ECGs. The performance of the model are 0.67, 0.44, 0.50 and 0.42 for AUC, AUPRC, F1 and Accuracy, respectively.

In these experiments, we quantified the average RV across all signals of a given class and the confidence interval of the mean at $1 - \alpha = 0.95$ confidence level. It is worth recalling that, differently from other interpretability methodologies, such quantification was possible because of the beat segmentation performed on the ECG.

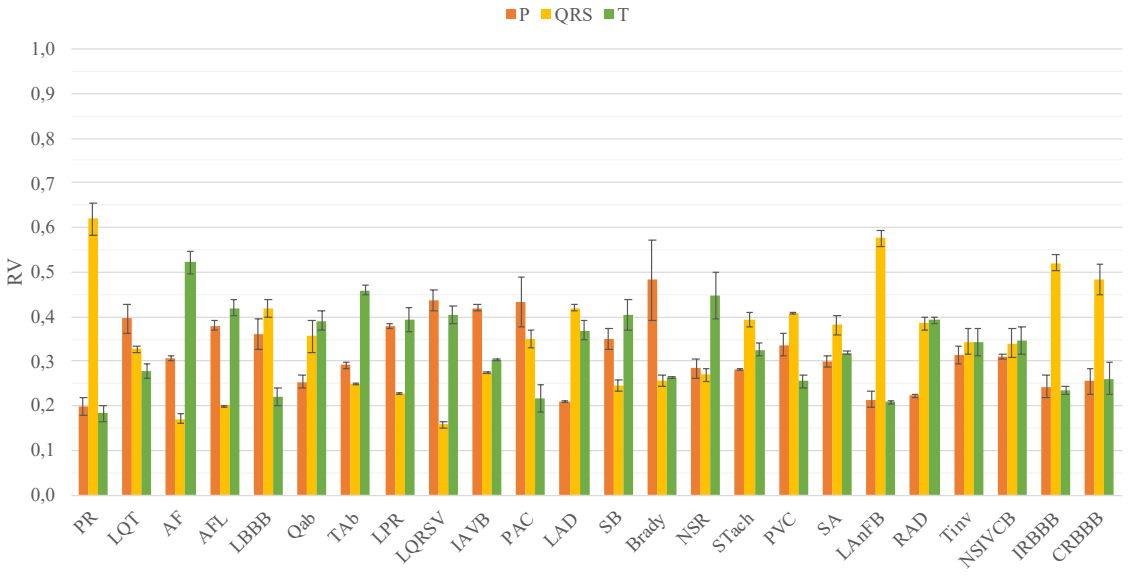
In addition, we determined whether the two interpretability frameworks were in agreement between each other by means of the Hellinger distance [28], quantifying the agreement as follows

$$a_c = 1 - \frac{1}{\sqrt{2}} \left[\sum_{i \in \{\mathrm{P,QRS,T}\}} \left(\sqrt{\overline{\mathrm{RV}}_{c,i}^{\mathrm{F1}}} - \sqrt{\overline{\mathrm{RV}}_{c,i}^{\mathrm{F2}}} \right)^2 \right]^{\frac{1}{2}}, \tag{3}$$

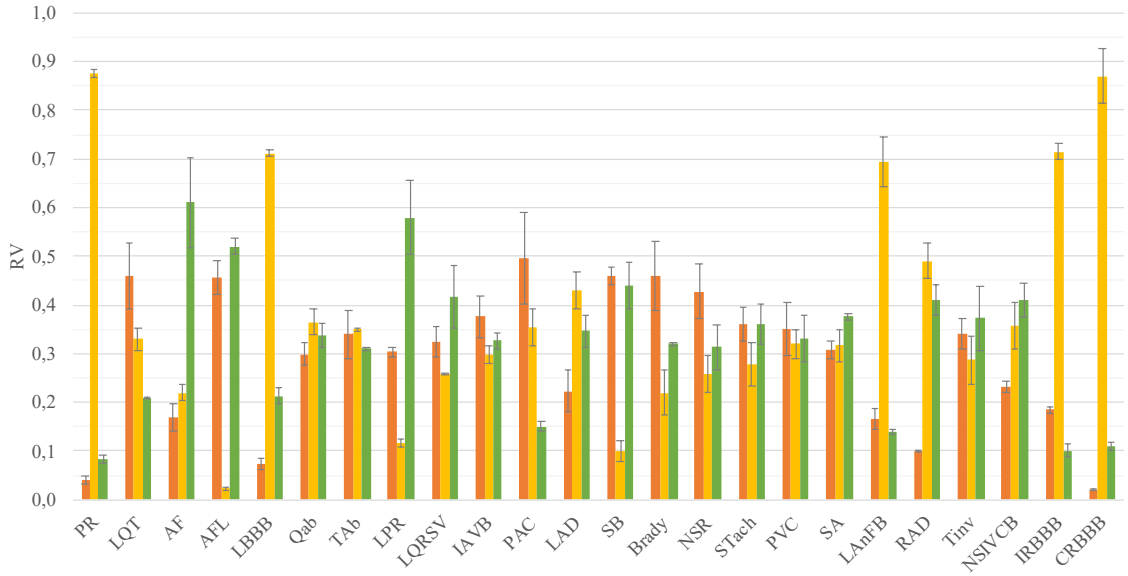
where $\overline{\mathrm{RV}}_{c,i}^{\mathrm{F1}}$ and $\overline{\mathrm{RV}}_{c,i}^{\mathrm{F2}}$ were the average RV over signals for class c and wave i, for framework 1 and 2, respectively. The agreement for the class c is maximum when the two triplets of RV values are equal, whereas it is minimum when the frameworks point to different waves with maximum RV. The agreement was calculated for each class.

2.3 The dataset

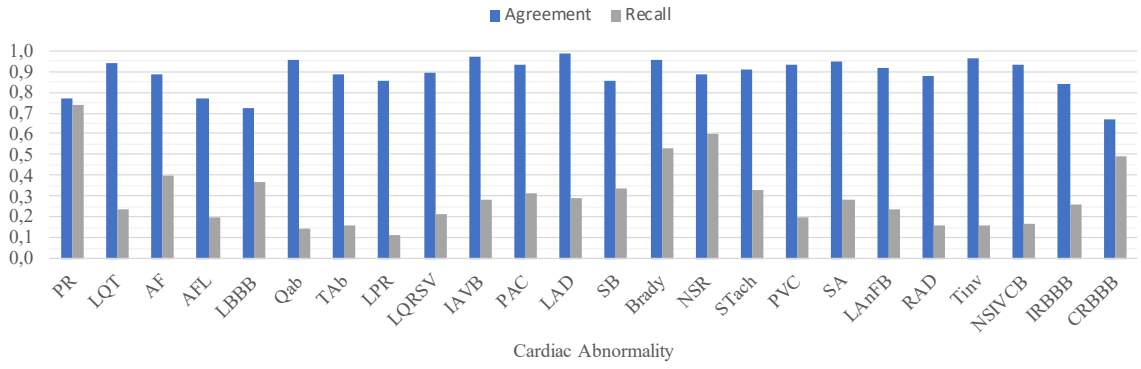
We used the dataset prepared for the PhysioNet/Computing in Cardiology Challenge 2020. The challenge was focused on automatic classification of cardiac abnormalities from 12-lead ECGs [29]. The dataset contains 12-lead clinical ECGs, labeled with one or more cardiac abnormalities, among 111 possible ones, in SNOMED-CT codes. The dataset was obtained merging data from the following sources: i) Southeast University, including the data from the China Physiological Signal Challenge 2018; ii) St. Petersburg Institute of Cardiological Technics; iii) Physikalisch Technische Bundesanstalt; and iv) Georgia 12-Lead ECG Challenge Database. ECG recordings lasted from 6 seconds to 30 minutes and sampling rates ranged from 257 Hz to 1000 Hz, where the majority was



(a)



(b)



(c)

Figure 2: Average and confidence interval of RV, for each class and ECG wave, of both versions of the framework, *i.e.* occlusion (a) and saliency maps (b). Agreement between the two methodologies and recall for each cardiac abnormality (c).

sampled at 500 Hz. A total number of 43101 ECG signals was available, each labeled by 111 possible classes. The challenge organizers restricted the problem to only 24 classes among the 111 ones. Classes referred to cardiac abnormalities related to either ECG morphology or heart rate rhythm.

2.4 Preprocessing of ECG signals for ECG segmentation

The 12-lead ECG signals were downsampled or upsampled to 500 Hz according to their actual sampling rate and 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. Beat detection was performed on the vector magnitude signal (VM) (square root of the sum of the squared ECG leads) using the *gqrs* algorithm [30] and beat positions, *i.e.* the R-peaks, were refined using the Woody algorithm applied to the VM [31]. Within each signal, we segmented the P-wave, QRS complex and T-wave for all beats. Segments relied on the R peaks previously identified and were defined as follows: 1) P-wave: R−240 ms to R−50 ms; 2) QRS complex: R ± 50 ms; 3) T-wave: R+50 ms to R+360 ms.

3 Results

In Figure 2a and 2b, we report the average RV and their confidence intervals for each class (cardiac abnormality) considered by the classifier, in Framework 1 (panel a), and 2 (panel b), computed by using eq. (1) and (2), respectively. The acronyms related to the classes are reported in the Appendix A. Focusing on the comparison between the confidence intervals, for the first framework, classes PR, AF, PAC, Brady, NSR, LAnFB, IRBBB, and CRBBB displayed average values of maximum RV which were significantly differed between QRS, T and P (hinting that one of these regions was significantly linked to the classifier's output). Similarly, PR, LQT, AF, LBBB, LPR, Brady, LAnFB, RAD, IRBBB, CRBBB were, in the second framework. For eight classes, both frameworks considered the same ECG wave as the most relevant for the classification.

In Figure 2c, we report the agreement plot between Framework 1 and 2, which is in general very high. The three classes with the highest agreement are IAVB, LAD, and Tinv. In such cases, the interpretability methods agree that the most relevant part of the ECG to take into consideration are the P-wave (IAVB), QRS complex (LAD) and T-wave (Tinv). The three classes with the lowest agreements are PR, LBBB and CRBBB. For only 4 out of 24 cardiac abnormalities (classes) the agreement is below 80%.

4 Discussion

The main contribution of our study is threefold. First, both versions of the framework were capable to provide a local interpretation for a given ECG signal classified by the network, by suggesting which ECG wave was involved in the decision. Depending on the class-wise recognition accuracy (recall) achieved and the ECG wave highlighted by the framework, our understanding of the decision process in place by the network may change. In fact, when the recall is low and the framework points to the right ECG wave for the diagnosis of the cardiac abnormality, the network is likely confounding its decisions with other abnormalities involving the same wave. For example, high RV values regarding the T-wave may mean that the network did not learn the correct pattern to distinguish between T-wave abnormality or T-wave inversion, but understood to focus on the correct wave. On the other hand, when the recall is high and the RV value is low, the network might be overfitting on the given dataset, or the cardiac abnormality is not uniquely related to a specific ECG wave (*e.g.* AF). In our preliminary work [13], using the LIME methodology [32], we found that a RF classifier was mostly relying on the QRS peak amplitude for providing its classification between normal ECG vs myocardial infarction. Such feature is not related with the considered cardiac abnormality, hence the ML algorithm was overfitting on the current dataset (which was then confirmed by further analysis). For the two remaining cases, *i.e.* high recall with high RV and low recall with low RV, the interpretations are straightforward: the network has either learnt to properly distinguish the cardiac abnormality from the others, or simply not.

Second, comparing the results of the framework with the domain knowledge, it became possible to determine whether the DL model selected was, on average, focusing on the right ECG wave. For example, a case when the ML method agreed with the cardiology domain knowledge was CRBBB (complete right bundle branch block) which affects the QRS complex morphology: both frameworks reported a maximum RV in correspondence with this ECG wave (Fig. 2a and 2b). On the contrary, for other classes the agreement with the domain knowledge was minimal. It is the case of Qab (abnormal Q point), where the Q point of the QRS complex is not within normality range: the frameworks pointed out that the model did not focus on any particular portion of the ECG, resulting in similar RV values for QRS, T and P waves. Similar results were achieved for Tinv (T-wave inversion).

Third, the frameworks agreed with each other for most of the classes (Fig. 2c), while differences were found for a few cardiac abnormalities. One may expect that the differences in the agreement between frameworks may

be connected with the recall of the specific class. In other words, when the recognition of the network is low, the agreement is also low, and viceversa. However, such expectation seems not to be supported by our findings (Fig. 2c). Indeed, we found that the agreement was unrelated with the class-wise performance of the network (Pearson's correlation coefficient between a_c and recall for each class was -0.4; $p > 0.05$). A similar result was obtained in a recent contribution from the Computer Vision domain, where several Explainable AI techniques were found to have significantly different performance (with the occlusion method as ground truth), even with a state-of-the-art DNN trained on millions of images [33].

Other recent studies proposed algorithms for interpreting the decisions of DNNs for automatic ECG classification. For instance, Baalman et al. [20] and Mousavi et al. [19] implemented attention mechanisms as multi layer feed-forward neural networks. Baalman et al. included such mechanism within a DNN to highlight the samples belonging to a single ECG beat that mostly contributed to the classification. In a similar fashion, more recently, Mousavi et al. proposed an attention mechanism working on three hierarchical levels. Their method was able to point to which wave (P, QRS, T), beat, or combination of beats were important for the classification. There are mainly three differences between these works and our approach. First, to the best of our knowledge, we are the first to systematically evaluate the performance of a DNN against the domain knowledge of ECG interpretation. Our frameworks indeed not only provide the ECG samples important for the classification, but also the importance of each wave in the final decision by means of the measure RV. Second, the evaluation was performed on 24 different cardiac abnormalities, while both papers only considered two classes. Third, our frameworks are also suitable for models already trained, and thus they do not necessarily require a dataset.

Both frameworks can be considered from two different perspectives: the one of the computer scientist (who creates the classifier) and the one of the physician (who uses the tool in the clinical practice). From the perspective of the computer scientist, the frameworks allow to inspect if the network relies on the ECG segments expected for the classification according to the clinical standard practice. Otherwise, the user can try to address the issue by understanding the reasons behind it, and thus guiding the architecture towards the domain knowledge. From the perspective of the physician, the frameworks allow to understand whether the decision taken relied on a known domain knowledge by highlighting the corresponding ECG wave, thus the trust in the deep learning model can be increased.

Under a supervised classification framework, a ML algorithm take a decision that is represented by the classification output itself. As a consequence, ML decision-makers can be trusted relying only on their predictive performance evaluated on the dataset available. Our effort in the direction of developing an interpretability methodology relies on the fact that we believe necessary that future advancement in automatic processing of ECG progresses together with our capability of understanding the decisions taken by a ML model. In this way, the large accuracy which ML algorithms might obtain in the future will also contribute in progressing the understanding of the underlying physiology.

Similar considerations were already present in the thinking of the ancient Greeks to obtain what Aristotle defined as τέχνη [téchne]: a real productive science [34]. He noted that the technological advancement can be obtained with different means, that could be achieved by either scientists or empirics. However, Aristotle set apart scientists from the empirics: people with a high degree of expertise in a specific domain, but who lack of any theory to justify their results. The empirics can even often achieve outstanding results, but what clearly divide science from empiricism is that it comes with a theory whose domain principles justify why certain decisions are correct in specific circumstances [35, 34]. ML algorithms are like empirics to a certain extents, but complemented with means of understanding (interpreting) their decision process may lead to scientific knowledge.

In our opinion, we do not believe that the current state-of-the-art ML algorithms might outperform significantly the human capability of detecting cardiac abnormalities. In fact, the number of possible confounding factors, co-morbidities, number of rare conditions, and evolution in time of the diseases may all increase the amount of data necessary for ML models to be trained. In order to mitigate such issues, the creation of innovative ML algorithms, capable of incorporating the domain knowledge, would facilitate the development of these models (*e.g.* less data-hungry algorithms, faster training, larger interpretability) and the introduction of such methodologies in the clinical practice, thus fostering trust for their use.

4.1 Limitations of the study

The frameworks presented some limitations. First, rhythm-based cardiac abnormalities were not properly handled by our methodology. Given the fact that an altered rhythm may or may not affect the regularity of the occurrence of any waves, it was not possible to define the one-to-one match between the cardiac abnormality and a specific ECG wave. For example, Brady (bradycardia), one of the class detected correctly by the network (recall of 0.53), refers to a very low heart rate. Typically, heart-rate alterations are quantified looking at the time intervals between consecutive R peaks because of their ease of detection. However, the frameworks found the P-wave very important for this class (Fig. 2a). Given the fact that the P-wave does not change during bradycardia, the assessment of the low heart rate might have been performed by the network "looking" at the

rate of both P-wave and QRS complex. Similarly, the frameworks found the T-wave relevant for the detection of AF (atrial fibrillation), which is characterized by the absence of the P-wave, an oscillatory pattern on the ECG baseline and irregular heart rate. In this case, the frameworks hint that the network may use samples between consecutive beats where the T-wave, the isoelectric line, and part of the P-wave are located to detect AF. The same observation is shared by the work of Mousavi et al. [19], where an attention mechanism was used to show that the network relied on samples between consecutive beats to detect AF. Third, PVC (premature ventricular contraction) have a morphology which is largely different from a normal sinus beat. Therefore, considering portion of the ECG where the P, QRS and T waves are usually located was not relevant (the different ECG waves had a similar low RV value).

Even if the algorithms for ECG segmentation have become well-established and validated in the recent years, it must be noted that the performance of the proposed framework is potentially dependent on the algorithm used. Usually, segmentation is performed after beat detection which is dependent on the quality of the ECG signal [36]. In the current work, we did not focus on the selection of the most robust segmentation algorithm for the clinical 12-lead ECG. However, given the fact that this type of ECG can be quickly acquired at low cost, physicians usually recollect the measurements in case of low quality. We therefore assumed that the ECG within the dataset were of sufficient quality (but we leave this investigation for the future). On the other hand, when the ECG signal is acquired in different contexts, *e.g.* sport activities and Holter acquisitions, the quality could be lower. In such cases, a careful preprocessing should be applied before running our frameworks.

5 Conclusion

In conclusion, by leveraging the output of interpretability methodologies and the easy segmentation of ECG signals, we proposed to evaluate the DL models for automatic ECG classification not only with the performance metrics typical of classification problems (*e.g.* accuracy, recall), but also with the domain knowledge of the clinical context. Therefore, the frameworks may be useful for the computer scientist as a "debugging" tool of DL models, and for the physician to increase their trust in these methodologies.

Appendix A List of Acronyms for Cardiac Abnormalities

Here, we report the list of acronyms referring to the considered cardiac abnormalities: pacing rhythm (PR), prolonged QT interval (LQT), atrial fibrillation (AF), atrial flutter (AFL), left bundle branch block (LBBB), Qwave abnormal (Qab), T wave abnormal (TAb), prolonged PR interval (LPR), low QRS voltages (LQRSV), 1st degree av block (IAVB), premature atrial contraction (PAC), left axis deviation (LAD), sinus bradycardia (SB), bradycardia (Brady), sinus rhythm (NSR), sinus tachycardia (STach), premature ventricular contractions (PVC), sinus arrhythmia (SA), left anterior fascicular block (LAnFB), right axis deviation (RAD), T wave inversion (Tinv), nonspecific intraventricular conduction disorder (NSIVCB), incomplete right bundle branch block (IRBBB), complete right bundle branch block (CRBBB).

Data Accessibility: The dataset provided for the PhysioNet/Computing in Cardiology Challenge 2020 can be accessed at: `https://physionetchallenges.github.io/2020/`. The BUTTeam deep learning model can be found at: `https://github.com/tomasvicar/BUTTeam`.

Authors' Contributions: MB designed the study and performed the analysis. All authors equally contributed to the preparation of the manuscript.

Competing Interests: The authors declare that they have no competing interests.

References

[1] Mendis S, Puska P, Norrving B, World Health Organization, World Heart Federation, World Stroke Organization, editors. 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.

[2] Adams HP, del Zoppo G, Alberts MJ, Bhatt DL, Brass L, Furlan A, et al. Guidelines for the Early Management of Adults With Ischemic Stroke. Stroke. 2007;38(5):1655–1711.

[3] Kligfield P, Gettes LS, Bailey JJ, Childers R, Deal BJ, Hancock EW, et al. Recommendations for the Standardization and Interpretation of the Electrocardiogram. J Am Coll Cardiol. 2007;49(10):1109–1127.

[4] Bickerton M, Pooler A. Misplaced ECG electrodes and the need for continuing training. British Journal of Cardiac Nursing. 2019 Mar;14(3):123–132.

[5] Rautaharju PM. Eyewitness to history: Landmarks in the development of computerized electrocardiography. J Electrocardiol. 2016;49(1):1–6.

[6] Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms. J Am Coll Cardiol. 2017;70(9):1183–1192.

[7] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.

[8] Mincholé A, Camps J, Lyon A, Rodríguez B. Machine learning in the electrocardiogram. J Electrocardiol. 2019;57:S61–S64.

[9] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65–69.

[10] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun. 2020;11(1).

[11] Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol. 2017;69(21):2657–2664.

[12] Castells F, Laguna P, Sörnmo L, Bollmann A, Roig JM. Principal Component Analysis in ECG Signal Processing. EURASIP J Adv Signal Process. 2007;2007(1).

[13] Bodini M, Rivolta MW, Sassi R. Interpretability Analysis of Machine Learning Algorithms in the Detection of ST-Elevation Myocardial Infarction. In: 2020 Comput Cardiol (CinC); 2020. p. (in press).

[14] Molnar C. Interpretable Machine Learning. Leanpub; 2020.

[15] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys (CSUR). 2018;51(5):1–42.

[16] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019 May;1(5):206–215.

[17] Vijayarangan S, Murugesan B, Vignesh R, Preejith S, Joseph J, Sivaprakasam M. Interpreting Deep Neural Networks for Single-Lead ECG Arrhythmia Classification. In: 42nd Annu Int Conf IEEE Eng Med Biol Soc; 2020. p. 300–303.

[18] Yao Q, Wang R, Fan X, Liu J, Li Y. Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network. Inform Fusion. 2020;53:174–182.

[19] Mousavi S, Afghah F, Acharya UR. HAN-ECG: An Interpretable Atrial Fibrillation Detection Model Using Hierarchical Attention Networks. Comput Biol Med. 2020:104057.

[20] Baalman SWE, Schroevers FE, Oakley AJ, Brouwer TF, van der Stuijt W, Bleijendaal H, et al. A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples. International Journal of Cardiology. 2020 Oct;316:130–136. Available from: `https://doi.org/10.1016/j.ijcard.2020.04.046`.

[21] Hong S, Xiao C, Ma T, Li H, Sun J. MINA: Multilevel Knowledge-Guided Attention for Modeling Electrocardiography Signals. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; 2019. p. 5888–5894.

[22] Goodfellow SD, Goodwin A, Greer R, Laussen PC, Mazwi M, Eytan D. Towards Understanding ECG Rhythm Classification Using Convolutional Neural Networks and Attention Mappings. In: Machine Learning for Healthcare Conference. vol. 85; 2018. p. 83–101.

[23] Han X, Hu Y, Foschini L, Chinitz L, Jankelson L, Ranganath R. Deep learning models for electrocardiograms are susceptible to adversarial attack. Nat Med. 2020;26(3):360–363.

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

[25] Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Computer Vision – ECCV 2014; 2014. p. 818–833.

[26] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: Workshop at International Conference on Learning Representations; 2014. p. 1–8.

[27] Vicar T, Novotna P, Hejc1 J, Ronzhina M, Janousek O. ECG Abnormalities Recognition Using Convolutional Network with Global Skip Connections and Custom Loss Function. In: 2020 Comput Cardiol (CinC); 2020. p. (in press).

[28] Hellinger E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik. 1909;1909(136):210–271.

[29] Perez Alday EA, Gu A, Shah A, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas. 2020 (Under Review).

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

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

[32] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining; 2016. p. 1135–1144.

[33] Lin ZQ, Shafiee MJ, Bochkarev S, Jules MS, Wang XY, Wong A. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. arXiv preprint arXiv:191007387. 2019.

[34] Barnes J. Complete works of Aristotle, volume 1: The revised Oxford translation. vol. 192. Princeton University Press; 1984.

[35] London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Cent Rep. 2019;49(1):15–21.

[36] Liu F, Liu C, Jiang X, Zhang Z, Zhang Y, Li J, et al. Performance Analysis of Ten Common QRS Detectors on Different ECG Application Cases. Journal of Healthcare Engineering. 2018;2018:1–8. Available from: `https://doi.org/10.1155/2018/9050812`.