Quantified Explainability: Convolutional Neural Network Focus Assessment in Arrhythmia Detection

Rui Varandas
Faculty of Science and Technology: Universidade Nova de Lisboa Faculdade de Ciencias e Tecnologia
Bernardo Brás Gonçalves (bb.goncalves@campus.fct.unl.pt)
Faculty of Science and Technology: Universidade Nova de Lisboa Faculdade de Ciencias e Tecnologia
https://orcid.org/0000-0001-8002-0391
Hugo Gamboa
Faculty of Science and Technology: Universidade Nova de Lisboa Faculdade de Ciencias e Tecnologia
Pedro Vieira
Faculty of Science and Technology: Universidade Nova de Lisboa Faculdade de Ciencias e Tecnologia

Research

Keywords: Deep Learning, Electrocardiogram, Computer Vision, Explainability, Convolutional Neural Network, Focus Quantification, Attribution Maps

Posted Date: July 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-666509/v1

Version of Record: A version of this preprint was published at BioMedInformatics on January 17th, 2022.
See the published version at https://doi.org/10.3390/biomedinformatics2010008.

RESEARCH

Quantified Explainability: Convolutional Neural Network Focus Assessment in Arrhythmia Detection

Rui Varandas[1,3*†], Bernardo Gonçalves[4†], Hugo Gamboa[1] and Pedro Vieira[4]

*Correspondence:
rvarandas@plux.info
[1]LIBPhys-UNL, Departamento de Física, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: Deep Learning (DL) models are able to produce accurate results in various areas. However, the medical field is specially sensitive, because every decision should be reliable and explained to the stakeholders. Thus, the high accuracy of DL models pose a great advantage, but the fact that they function as a black-box hinders their application to sensitive fields, given that they are not explainable *per se*. Hence, the application of explainability methods became important to provide explaination DL models in various problems. In this work, we trained different classifiers and generated explanation of their classification of electrocardiograms (ECG) by applying well-known methods. Finally, we extract quantifiable information to evaluate the explanation of our classifiers.

Methods: In this study two datasets were built consisting of image representation of ECG that were labelled given one specific heartbeat: 1. labelled given the last heartbeat and 2. labelled given the first heartbeat. DL models were trained with each dataset. Three different explainability methods were applied to the DL models to explain their classification. These methods produce attribution maps in which the intensity of the pixels are proportional to their importance to the classification task. Thus, we developed a metric to quantify the focus of the models in the region of interest (ROI) of the ECG representation.

Results: The developed classification models achieved accuracy scores of around 93.66% and 91.72% in the testing set. The explainability methods were successfully applied to these models. The quantification metric developed in this work demonstrated that, in most cases, the models did have a focus around the heartbeat of interest. The results ranged from around 8.8% in the worst case, until 32.4%, where the random focus would mean a value of approximately 10%.

Conclusions: The classification models performed accurately in the two datasets, however, even though their focus is higher in the ROI of the figures compared with the random case, the results allow the interpretation that other regions of the figures might also be important for classification. In the future, it should be investigated the importance of regions outside the ROI and also if specific waves of the ECG signal contribute to the classification.

Keywords: Deep Learning; Electrocardiogram; Computer Vision; Explainability; Convolutional Neural Network; Focus Quantification; Attribution Maps

Background

Deep learning (DL) models have a increasing impact on today's scientific research. These models have achieved state of art results (most of them strictly academic) in many fields, such as image classification or natural language processing. However,

they lack transparency. For that reason they are often referred as black-box models. Also, there are questions about the fairness of the models. For example, lack of fairness in the medical domain may be introduced by bias towards some aspect of the person, such as, gender, ethnicity, sexuality or disability [1]. The inability to explain DL models is a major handicap. It prevents these models to be massively used in sensitive tasks, such as, autonomous driving or medical diagnoses. Government organisations and institutions have shown their concern about this issue. A series of guidelines about methods to improve the transparency of artificial intelligence models have been created [2]. For example, the European Commission redacted a technical report, titled, *Robustness ans Explainability of AI*. The authors of [3] warned about the importance of standardisation and certification tools for AI in order to create more robust and understandable AI applications. All of the above serves as motivation for our study.

To explain a DL model, one can have two different approaches: transform the model until it becomes a self explainable model or apply explainability methods. The first approach relies on the reduction of complexity of the model [4]. The second one relies on the usage of specific methods that extract explanations from complex models [2]. These two approaches are called intrinsic and post-hoc explainability, respectively [5]. Explainability tools can also be categorised as model agnostic or model specific and local or global [6]. Model agnostic tools are also post-hoc and they can be applied to any type of model [5]. These methods do not have access to the network model internal components like weights or structural information [5]. Global interpretation tools focus on the overall understanding of the DL model features and each of the learned components. Local interpretation tools check individual predictions of the model. Local methods are less complex to implement compared to global interpretation tools. Global methods are normally applied to simpler DL models [6].

The transparency and explainability of DL models is key for medical applications, as discussed in [7]. The same work enumerates numerous flaws that black-box models present in this specific domain, namely in the ethical sense and in the disputability sense from the patients perspective. Patients have the right to know the origin of diagnostics, recommended therapeutics or any other medical intervention that may be supported by artificial intelligence models, such as DL models (Chapter 3 of General Data Protection Regulation - GDPR) [8, 9]. Thus, without the creation of tools to explain a DL model it is difficult to apply them in clinical context.

In our study we applied three different methods to explain the classification task. To perform the classification we used a convolution neural network (CNN). Our classification task was the detection of arrhythmia in electrocardiography images. Then, we create three different attribution maps: saliency maps, gradient-weighted Class Activation (grad-CAM) maps and guided backpropagation (GB) grad-CAM maps - these highlight the most important pixels, of an input image, to the resultant classification [2, 6]. The methods that underline these maps are local and post-hoc. In the methods section a detailed description of these methods will be presented.

Electrocardiography Classification

Machine learning and, specifically, deep learning techniques have been applied for ECG signals classification in the case of arrhythmia detection [10].

Traditional machine learning algorithms, even though being transparent regarding the classification process, can also depend on tedious and costly tasks, such as feature engineering. For example, in [11] the authors used optimum-path forest and demonstrated its dependence on the feature representation given as input. In the same work, they presented results using other commonly used classifiers using the same sets of features, namely, Support Vector Machines (SVM), Multi-Layered Perceptron (MLP) and Bayesian expert system classifiers, all of which depended on the input features.

In [12], the authors applied an intermediate approach, in which they used Radial Basis Function (RBF) network to model the ECG heartbeats from the different classes and then, using a deterministic learning algorithm, performed classification on the test set, with accuracy score of around 98%. Thus, the application of the RBF allowed to automatically extract dynamic features from every heartbeat.

Notwithstanding the issues around DL models in practice, they have been applied in academic works on ECG signals for the detection of arrhythmia events [13, 14, 15]. To overcome the feature engineering process that is required for traditional machine learning algorithms, in [16] the authors applied a combination of a 1-D CNN for feature extraction and 3 Fully Connected Network (FCN) layers for classification, achieving an accuracy of 86%. The authors of [13] used a Denoising AutoEncoder (DAE) for unsupervised features extraction and stacked its hidden representation layers with a regression layer, which is capable of assigning scores based on the examples in the training set. The innovation is the usage of Active Learning, in which experts can provide input to the most informative heartbeats given certain criteria, resulting in improved results compared to other works.

Explainability in Electrocardiography Classification

Works addressing the problem of interpretability in an ECG classification problem are scarce. The work in [17] used a hierarchical attention network combined with Bidirectional Recurrent Neural Networks (BiRNN) for the classification task. Explainability was introduced by the hierarchical structure and no specific aforementioned algorithm was used. However, the authors were able to explain the decision process based on the specific results for each hierarchy, that corresponded to windows (set of heartbeats), heartbeat and, finally, waves (P, QRS complex or T).

The authors in [18] developed an explainability framework specifically addressed at the problem of ECG classification, which encompassed three modules that evaluated the features extracted from a 1-D CNN used for the classification of ECG data. In this case, the authors used segments that contained 5 heartbeats and signal synthesis methods as data augmentation for improved results. The internal states of the CNN were not taken into account for the proposed explainability method, which relied only on the features extracted from the input signals.

We propose to expand the knowledge of this specific problem by using computer vision allied with specific explainability algorithms to increase interpretability of the results of DL models in ECG images that contain six heartbeats. We chose to use ECG images to simulate the work of cardiologists when analysing ECG in real-life. We trained two different models to predict the label of specific heartbeats.

One model classifies the first heartbeat while the other classifies the last heartbeat. We aim to verify if the model focus is on the labelled heartbeat. To perform a quantitative analysis, we computed the proportion of attribution between the area with heartbeat of interest and the rest of the image.

Given this, our objectives were to (1) develop two different datasets consisting of ECG figures containing various heartbeats and labelled given a specific heartbeat; (2) train a CNN model for each dataset to detect arrhythmia in ECG figures; (3) apply the aforementioned explainability methods to visually interpret the classification of the CNN with the aim of understanding if the models are able to focus on the heartbeat of interest - the heartbeat that provides the label to each figure; (4) develop a metric to quantify the amount of focus of the models to the heartbeat of interest. Moreover, with this metric we investigated the effect that different labels have on the focus of the classification models and also if accurate classifications are reflected on the focus of the model.

Results

This work can be divided in two distinct parts: one regarding the classification of ECG images and the other focused on the analysis of the created attribution maps using our custom metric.

Starting by the classification results presented in Table 1, model 1 (label corresponding to the last heartbeat of the figure) performs better than model 2 (label corresponding to the first heartbeat of the figure) in the testing set, with an accuracy of 93.66%.

Table 1 Classification results. The models 1, 2, were trained with dataset 1, 2, respectively. All presented values are percentages.

Models	Train	Validation		Test		
	Epochs	Accuracy	F_1 score	Accuracy	F_1 score	Precision
Model 1	7	94.06	96.82	93.66	96.47	74.12
Model 2	5	96.23	98.00	91.72	95.38	63.57

In regards of the explored explainability metrics, we present three different scenarios. The generic scenario, presented in Table 2, where histograms are shown in Figure 1, is the overall comparison of the three explainability methods relative to the classification of the two models. In this case, we see that the method that better focus the heartbeat of interest in both classification models is the Gradients method, with a focus of around 30% in model 1 and 25% in model 2, while the Grad-CAM is the method with worst focus with a focus of around 23% in model 1 and 9% in model 2.

Table 2 Attribution Metric - Generic. Mean value of the metric of all test samples. Each line for a different dataset: 1 - last heartbeat labelled; 2 - first heartbeat labelled. All presented values are percentages.

Set	Gradients	Grad-CAM	GB grad-CAM
1	30.3 ± 9.2	22.7 ± 11.5	27.4 ± 10.9
2	25.0 ± 18.2	8.7 ± 5.2	15.3 ± 9.9

The comparison between correct classification vs incorrect classification, presented in Table 3 shows that, in the case of the Grad-CAM and GB grad-CAM in model

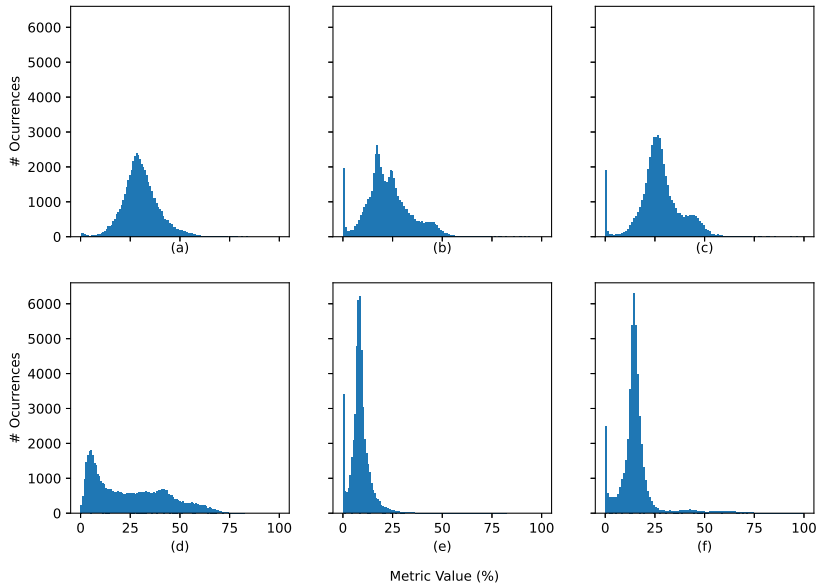

Figure 1 Histograms of the distribution of the custom attribution metric values for the generic case. (a), (b) and (c) correspond to the metrics estimated using model 1 and the methods: gradients, grad-CAM and GB grad-CAM, respectively. (d), (e) and (f) correspond to the metrics estimated using model 2 and the methods: gradients, grad-CAM and GB grad-CAM, respectively.

1, there is a positive relation between correct classification and the focus of the attribution maps. Notwithstanding, in model 2 there is not a clear relation between the classification results and the focus of the attribution maps.

Table 3 Attribution Metric - Correct Classification vs Incorrect Classification. Mean value of the metric of the test samples that were correctly classified and those who were not.Each line for a different dataset: 1 - last heartbeat labelled; 2 - first heartbeat labelled. All presented values are percentages.

Set	Gradients		Grad-CAM		GB grad-CAM	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	30.3 ± 8.9	31.4 ± 12.6	23.6 ± 10.7	10.2 ± 15.2	28.3 ± 9.7	13.3 ± 16.8
2	24.4 ± 18.0	32.4 ± 19.3	8.8 ± 4.3	8.3 ± 10.8	15.5 ± 8.8	12.7 ± 19.4

Finally, the comparison between the two labels of the images, presented in Table 4 and exemplified in Figure 2, shows that there is not a consistent relation between the label of the images and the focus of the attribution maps. For example, the Grad-CAM shows better results in model 1 for the normal label, while for model 2 the opposite happens.

Discussion

The first step in our project was to detect arrhythmia in the last or first heartbeat within ECG images containing six heartbeats. Thus, we created two different models: one to classify the last heartbeat (model 1) and another to classify the first heartbeat (model 2). Table 1 shows that model 1 performs slightly better than

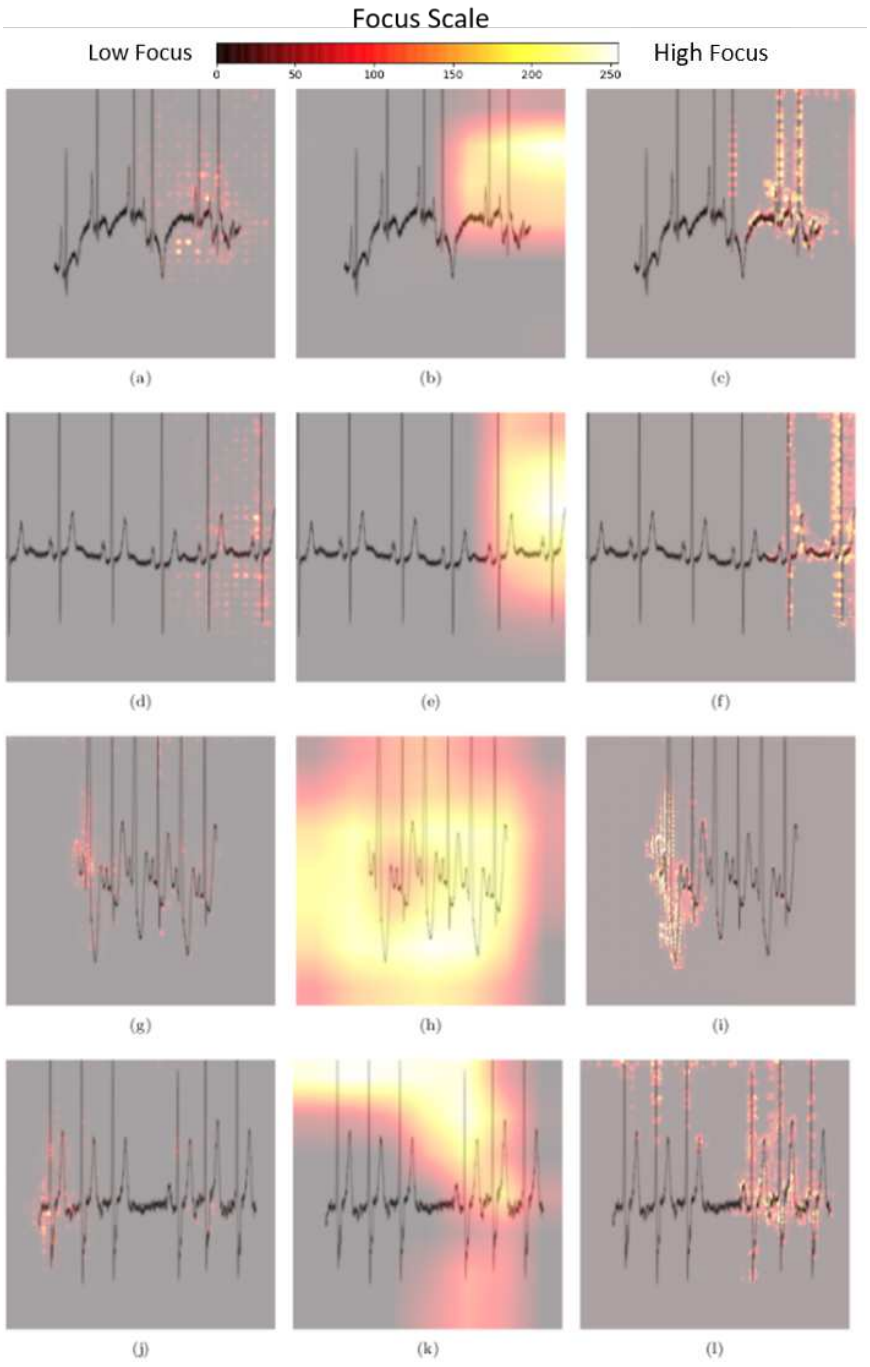

Figure 2 Examples from the obtained attribution maps. The first row of figures corresponds to maps created using model 1 (the heartbeat of interest is the last) and to an abnormal sample: (a) saliency map; (b) grad-CAM map; (c) GB grad-CAM map. The second row corresponds to maps created using model 1 and to a normal sample: (d) saliency map; (e) grad-CAM map; (f) GB grad-CAM map. The third row corresponds to maps created using model 2 (the heartbeat of interest is the first) and to an abnormal sample: (g) saliency map; (h) grad-CAM map; (i) GB grad-CAM map. The fourth row corresponds to maps created using model 2 and a normal sample: (j) saliency map; (k) grad-CAM map; (l) GB grad-CAM map. In all the presented cases the models gave the correct predictions. The scale presented at the top implies that the brighter pixels correspond to higher focus.

Table 4 Attribution Metric - Abnormal label vs Normal label. Mean value of the metric of the test samples that are abnormal and those which are normal. Each line for a different dataset: 1 - last heartbeat labelled; 2 - first heartbeat labelled. All presented values are percentages.

Set	Gradients		Grad-CAM		GB grad-CAM	
	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal
1	34.6 ± 10.3	29.8 ± 8.9	18.4 ± 15.5	23.3 ± 10.8	24.0 ± 17.9	27.8 ± 9.7
2	39.7 ± 19.3	23.2 ± 17.3	10.7 ± 10.2	8.5 ± 4.2	31.9 ± 24.1	13.8 ± 5.2

model 2 at the testing set, consisting of patients that are not in the other sets (interpatient classification). We hypothesise that, similarly to how humans classify ECG, the beats prior to the heartbeat of interest help the model to produce the correct prediction.

Nevertheless, our classification results of 91-94% were slightly bellow the state of the art results presented in [10]. The best study achieved an accuracy of 98 % in a multiclass problem using the same dataset we used. Since we performed binary classification we can conclude that our classification models were not as robust and accurate as the state of the art models. This can be explained by multiple factors. The reported studies treat ECG signals as 1D signals while we treat the signals as 2D (images). This increases the classification complexity and therefore may worsen the results. Additionally, we use all the signals in the MIT BIH database without any preprocessing. This can be problematic because some ECG signals in that database have a low signal-to-noise ratio. Finally, we used the dataset division proposed in [10], but samples from the training set were used to build the validation set, resulting in a testing set that was larger than the training set. By doing this, we are certain to use the same testing set as the reported studies. Since our main objective was the transparency of DL models, we did not focus on the optimisation of the classification results.

After the classification task, the next step was the application of the already enumerated explainability methods to create attribution maps. From those maps we computed our metric to measure the amount of focus of the model on the heartbeat of interest. The maps created using the gradients method do not distinguish between positive or negative contributions to the model prediction [19]. Attribution maps created using grad-CAM and GB grad-CAM method only consider positive contributions. For this reason the values of our metric are generally higher for attributions maps created using the gradients method. Nevertheless, we expect the model to be focused in the region of the labelled heartbeat, even if certain pixels of that region give negative contributions to the prediction.

Supposing that the model is not focused in any particular zone of the input images, the average value of our metric would be $9.7 \pm 4.0\%$. We called this value the random focus value. This value was estimated by computing the average proportion between the area of the region of interest and the area of the image across all test samples. The values shown in Tables 2, 3 and 4 are higher than the random focus value for all attribution maps except for specific cases in the grad-CAM attribution maps. This fact implies that both our models considered the region of the labelled heartbeat an important region to the prediction process. For further analysis we created 3 different scenarios of comparison: the generic case; the correct vs incorrect classification case and the abnormal vs normal label case.

From the generic case we can conclude that the obtained values of the attribution metric corroborate the better classification performance of model 1. Model 1 has higher values for all maps. From Table 2 we can also highlight the fact that the attribution metric using grad-CAM in model 2 was the only one bellow the random focus value. In fact, generally, grad-CAM has the lowest values for all cases. Grad-CAM maps are the most coarse attribution maps of the three different maps that were created [20]. Therefore the higher pixel values are more scattered across the map, when comparing with the other maps, resulting in lower values of our custom attribution metric. The histograms presented in Figure 1 show that despite having the higher average values, model 1 also have the higher deviation values for all maps, except for the ones created using gradients method - (a) and (d). The frequency of null values can also be seen as not negligible for the grad-CAM and GB grad-CAM methods - (b), (e) and (c), (f) . Those null values represent the cases where there is not any pixel attribution of positive contributions inside the region of interest. Gradients maps have a negligible frequency of null values because they consider positive and negative contributions to the prediction process.

With the second scenario, Table 3, we extrapolate a relation between the value of our custom attribution metric and the coherence between classification result and the actual label of the samples. We excepted that correct classifications were related with a more focused model (higher values of the metric) however this is not the case when applying the gradients method. This is again related with the nature of the method that does not distinguish between positive and negative contributions.

Finally, our third scenario, Table 4, focus on the relation between the actual label of the test sample and its attribution map. Here we cannot find a general tendency for both models. For model 1, both GB grad-CAM and grad-CAM have higher values when classifying normal samples. On the contrary for model 2, higher values are obtained when classifying abnormal samples. Gradients maps have higher values for abnormal samples in both models. These tendencies can be seen in Figure 2.

Notwithstanding that we are assessing the focus of our classification model in the region that contains the labelled heartbeat, we also hypothesise that high attribution values outside that region in our input images might be relevant for classification. There, the focus can be on the blank parts of the figures or in the remaining heartbeats that do not contribute to the label. In the first case, the model might be searching for heartbeats that should happen (e.g. if the signal is shorter than expected) or for higher amplitude signals than the ones that are present. In the second case, the model might be looking for inter-beat features, such as, the distance between R peaks, to classify the most important beat (important in some arrhythmia cases, such as tachycardia or bradycardia). In fact, the low values of our metric support this hypothesis, but further research is required to validate this.

Conclusions

In this study, we built two computer vision classification models and apply three different backpropagation based explainability methods to each, in order to create attribution maps. From the attribution maps, we then created a custom metric that measures the degree of importance of each pixel of the input image given the classification result. Then, we compared the obtained values of our metric across

different cases: the generic case, accurate vs inaccurate classification and abnormal vs normal samples.

The classification results were bellow the state of the art results. Both models achieve testing accuracy scores between 91-94 %. The values of our metric ranged between 8 and 38 % with high standard deviation values. Those values show that the focus of the classification models is sparse across the ECG image, even though there is a concentration in the heartbeat of interest.

To improve our project, we would try to improve the classification results by pre-processing the ECG signals that we used. Moreover, we would improve the computation of the heartbeat region of interest in order for it to become more precise. Besides, we would like to extend the knowledge on this subject by analysing the importance of other regions of the ECG images. We should explore specific waves inside of the heartbeat of interest to assess their importance to classification. Furthermore, it would be interesting to explore if DL models capture pseudo-time dependencies by computing the distant between R peaks and other commonly extracted features, that might explain the importance of regions of interest besides the labelled heartbeat.

Methods

Dataset Description

MIT BIH arrhythmia database was used for the exploration of explainability algorithms applied to ECG images. This dataset is comprised of 48 half-hour ECG recordings of 47 different subjects, where 23 were chosen randomly from a set of 4000 recording and 25 were chosen to include unusual arrhythmia events [21]. The sampling frequency of the recording is 360 Hz, with 11 bit resolution over a range of 10 mV. The labelling of each heartbeat and arrhythmia event was made by two different cardiologists. The total number of labels and, thus, of heartbeats is approximately 110.000, including noisy and virtually intractable parts of signals, that were discarded in this work.

Since the data corresponds to 1-D signals, the first step for the application of computer vision techniques is to convert it to images. To mimic real-world applications, these images will comprise sets of 6 heartbeats and the samples will be constructed based on a sliding-window approach. Using those same images, 2 different datasets were created, generating 2 different models with different training data. Those datasets can be described as follows:

1. Dataset 1 - **binary** label of each image is the label of the **last heartbeat**
2. Dataset 2 - **binary** label of each image is the label of the **first heartbeat**

Although there are 15 types of arrhythmia, we considered only the normal label and the remaining are comprised in the abnormal label.

To feed the created sample images into our DL model some transformations were performed. The images were cropped, to minimize the amount of image without relevant information. They were also resized (224x224) and normalized. Finally, following the recommendations in [22], the dataset was divided in two. Each subset with a different group of patients. Then, we created a validation set from the train set. By doing this we obtained: 37867, 11204 and 49617 images, in the training, validation and test sets, respectively.

Model Description

We used a ResNet50 to perform our binary classification task. This model was created to surpass the difficulty of training very large neural networks [23]. ResNet creators introduced residual blocks that ease the training process. To import, train, and evaluate the model and to perform all data transformations we used *Pytorch*[1].

The imported model was already trained for a natural image classification problem, on the ImageNet dataset [24]. Due to the significant difference between the pretraining task and the actual task of this project, the model was trained from scratch with a small learning rate. This was only possible due to the large number of samples in the datasets created for this project. In order to adapt the imported model to the task at hand, the last fully connected layer of the model was replaced by another fully connected layer but with an output size equal to the number of classes, i.e., output size of two.

To optimize the model we chose the Adam algorithm with weight decay. Additionally, we applied learning rate decay with a decay step of 4 epochs and gamma of 0.1. The initial learning rate was defined as $1e-5$ and $1e-4$, for datasets 1 and 2, respectively. Reducing the learning rate of the model as the training progresses allows the model to become more stable in advanced epochs.

The loss algorithm used was the weighted cross entropy algorithm. The usage of a weighted loss was important because of the class imbalance that is present in the dataset (very common in any medical dataset). For that reason, to the class with less samples was given a higher weight in the loss computation.

The evaluation metrics used were accuracy, precision and F_1 score. Using the evolution of the F_1 score metric in the validation dataset we developed an early stop mechanism. If the F_1 score of a certain epoch is less or equal than the mean of the same parameter at the last 4 epochs, then the training stops. This mechanism reduces the training time without compromising the performance of the model. Equations 1, 2, 3 show the mathematical formulas of the evaluation metrics.

$$Accuracy = \frac{Number\ correct\ predictions}{Total\ number\ of\ predictions} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$F_1\ score = \frac{TP}{\frac{1}{2}TP(FP + FN)} \tag{3}$$

Wherein TP, FP, FN are the true positives, false positives and false negatives, respectively.

[1]https://pytorch.org/

Explainability methods

In our project we used three different explainability methods: *gradients* (generates saliency maps), *grad-CAM* and *guided backpropagation grad-CAM*. Besides being local and post hoc methods, they are also backpropagation based methods. Accordingly with [19], backpropagation based methods compute attribution for all input features with a single forward and backward pass through the network. Some methods inside this category can only provide, in the attribution map, positive contributions to the final prediction result, while others shows the positive and negative contributions, which can degrade the results as it increases the noise in the map.

Gradients Method

Saliency Map is the oldest and probably one of the most used methods to explain the prediction of CNNs. The saliency map of the input of a CNN highlights the parts of the input that most contributes to the outcome and, so, the method attributes importance to the various pixels of an input pixels regarding the prediction of the network.

Based on the work [25] that introduced this method, the pixel importance is obtained by applying the somewhat inverse operation relative to the training of Neural Networks (NN). Neural Networks are usually trained by the application of backpropagation regarding the expected labels in order to optimise the loss function and is applied from the input to the output of the networks. However, to obtain the saliency maps, the same backpropagation algorithm is applied, but in this case the derivative is applied regarding the input image:

$$w = \frac{\partial y^c}{\partial I}\Big|_{I_0} \tag{4}$$

where y^c is the class score, I is the image and I_0 is the input image, specific for the task at hand, and w is the attribution map - analogous to the weights of NN.

Gradient-Weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping (grad-CAM) was first introduced in [26] as a generalisation of Class Activation Mapping (CAM). Unlike CAM, grad-CAM can be used to visualize any type of CNNs. Grad-CAM uses gradient information that flows to the last convolution layer to compute the importance, for the prediction, of each neuron. The last convolutional layers of a CNN retain spatial information and its neurons are focused on semantic class-specific information in the input image. For this reason grad-CAM is a class discriminative method.

Equation 5 shows how to compute the weight α_k^c. This weight captures the importance of a feature map k for a target class c. This value is the global average pooling of the gradient of the score (before softmax) for class, c, w.r.t the feature maps, A^k: $\frac{\partial y^c}{\partial A^k_{ij}}$.

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{5}$$

After this step, to compute the final map it is necessary to perform a weighted combination of forward activation maps and a ReLU, as shown in Equation 6. The application of the ReLU guarantees that the attribution map only depicts the positive contributions to the classification result.

$$L^c_{Grad-CAM} = ReLU\left[\sum_k \alpha^c_k A^k\right] \tag{6}$$

These operations create a coarse heat map of the same size as the convolutional feature maps. Although we can apply grad-CAM to any convolutional layer as we are trying to explain the decisions of our classifier we applied the method to the last convolutional layer of our ResNet.

Guided Backpropagation Gradient-Weighted Class Activation Mapping

The guided backpropagation grad-CAM (GB grad-CAM) was also introduced in [26]. This method was developed to tackle the lack of finer details in the attribution maps created using a grad-CAM. It consists in a pixel wise multiplication between a grad-CAM map and guided backpropagation (GB) map.

GB maps were first described in [27]. These maps are an improved version of the saliency maps. Instead of using a normal backpropagation approach, they use a guided backpropagation. GB combines two methods: 'deconvnet' [28] and backpropagation. These methods differ only in the way they handle backpropagation through the ReLU nonlinearity. The 'deconvnet' method, considers only the top gradient signal to compute the gradient in the nonlinearity and ignores the bottom input. GB combines this with backpropagation and masks out the values for which the top gradient or bottom data are negative. This prevents the backward flow of negative gradients.

Quantitative Analysis of Pixel Attribution Maps

In order to conclude about the magnitude of focus in the different regions of the ECG images we computed a proportion of attribution between the heartbeat of interest and the rest of the image. Firstly we computed a rectangular region of interest (ROI) that contains only the heartbeat of interest. This heartbeat is the labelled heartbeat, that varies accordingly with the dataset. Figure 3 illustrates an example of a computed ROI. In that case the last beat was the heartbeat of interest. Then we determine the proportion between the total sum of the pixel attribution map and sum of the map inside that region of interest. Using this proportion we can measure the percentage of focus inside the ROI.

Acknowledgements

We would like to acknowledge Dr. André Martins from Instituto Superior Técnico, Universidade de Lisboa, for his inputs in the early stages of this work.

Funding

This work was funded by FCT - Portuguese Foundation for Science and Technology and Bee2Fire under the PhD grant with reference PD/BDE/150624/2020. Moreover, it was funded by FCT - Portuguese Foundation for Science and Technology and PLUX Wireless Biosignals S.A. under the PhD grant with reference PD/BDE/150304/2019.

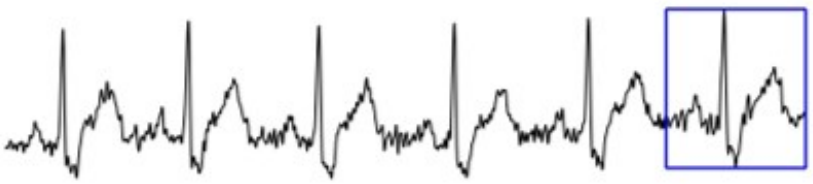

Figure 3 Normal electrocardiographic signal with a rectangular ROI that contains the last heartbeat.

Ethics approval and consent to participate
Not applicable

Availability of data and materials
All data is publicly available.

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

Authors' contributions
RV and BG both contributed in equal form in the development of the work reported here as well as in the writing of the paper. HG and PV participed in the scientific discussions inherent to this work. All authors read and approved the final manuscript.

Consent for publication
Not applicable.

Author details
[1]LIBPhys-UNL, Departamento de Física, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal. [3]Bee2Fire, S.A., Edi. Inov. Point, Sala 2.16, TagusValley-Tecnopolo do Vale do Tejo, R. José Dias Simão, Alferrarede, 2200-062 Abrantes, Portugal. [3]Bee2Fire, S.A., Edi. Inov. Point, Sala 2.16, TagusValley-Tecnopolo do Vale do Tejo, R. José Dias Simão, Alferrarede, 2200-062 Abrantes, Portugal. [4]Physics Department, Faculty of Science and Technology, NOVA University of Lisbon, Caparica Campus, 2829-516 Caparica, Portugal.

References

1. Vellido A. Societal Issues Concerning the Application of Artificial Intelligence in Medicine. Kidney Diseases. 2019;5(1):11–17.
2. Liang Y, Li S, Yan C, Li M, Jiang C. Explaining the black-box model: A survey of local interpretation methods for deep neural networks. Neurocomputing. 2021 jan;419:168–182. Available from: https://doi.org/10.1016/j.neucom.2020.08.011https://linkinghub.elsevier.com/retrieve/pii/S0925231220312716.
3. Hamon R, Junklewitz H, Sanchez I. Robustness and Explainability of Artificial Intelligence; 2020. Available from: https://publications.jrc.ec.europa.eu/repository/handle/JRC119336.
4. 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. Available from: https://doi.org/10.1038/s42256-019-0048-xhttp://www.nature.com/articles/s42256-019-0048-x.
5. Molnar C. Interpretable Machine Learning; 2019. https://christophm.github.io/interpretable-ml-book/.
6. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: A review of machine learning interpretability methods. Entropy. 2021;23(1):1–45.
7. Ploug T, Holm S. The four dimensions of contestable AI diagnostics- A patient-centric approach to explainable AI. Artificial Intelligence in Medicine. 2020;107(June):101901. Available from: https://doi.org/10.1016/j.artmed.2020.101901.
8. Chapter 3 - Rights of the data subject — General Data Protection Regulation (GDPR). General Data Protection Regulation (GDPR); 2018. (Accessed: 14-06-2021). Available from: https://gdpr-info.eu/chapter-3/.
9. Clarke N, Vale G, Reeves EP, Kirwan M, Smith D, Farrell M, et al. GDPR: an impediment to research? Irish Journal of Medical Science. 2019;188(4):1129–1135.
10. Luz EJdS, Schwartz WR, Cámara-Chávez G, Menotti D. ECG-based heartbeat classification for arrhythmia detection: A survey. Computer Methods and Programs in Biomedicine. 2016;127:144–164.
11. Luz EJDS, Nunes TM, De Albuquerque VHC, Papa JP, Menotti D. ECG arrhythmia classification based on optimum-path forest. Expert Systems with Applications. 2013;40(9):3561–3573.
12. Dong X, Wang C, Si W. ECG beat classification via deterministic learning. Neurocomputing. 2017;240:1–12.
13. Rahhal MMA, Bazi Y, Alhichri H, Alajlan N, Melgani F, Yager RR. Deep learning approach for active classification of electrocardiogram signals. Information Sciences. 2016;345:340–354.
14. Pyakillya B, Kazachenko N, Mikhailovsky N. Deep Learning for ECG Classification. In: Journal of Physics: Conference Series. vol. 913; 2017. .

15. Rim B, Sung NJ, Min S, Hong M. Deep learning in physiological signal data: A survey. Sensors (Switzerland). 2020;20(4).
16. Pyakillya B, Kazachenko N, Mikhailovsky N. Deep Learning for ECG Classification. In: Journal of Physics: Conference Series. vol. 913; 2017. Available from: http://iopscience.iop.org/article/10.1088/1742-6596/913/1/012004/pdf.
17. Mousavi S, Afghah F, Acharya UR. HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks. Computers in Biology and Medicine. 2020;127(June):104057. Available from: https://doi.org/10.1016/j.compbiomed.2020.104057.
18. Maweu BM, Dakshit S, Shamsuddin R, Prabhakaran B. CEFEs: A CNN Explainable Framework for ECG Signals. Artificial Intelligence in Medicine. 2021;115:102059. Available from: https://www.sciencedirect.com/science/article/pii/S093336572100052X.
19. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. Journal of Imaging. 2020;6(6):1–19.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. Revista do Hospital das Cl??nicas. 2016;17:331–336. Available from: http://arxiv.org/abs/1610.02391.
21. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Engineering in Medicine and Biology Magazine. 2001;20(3):45–50.
22. Luz EJ, Schwartz WR, Cámara-Chávez G, Menotti D. ECG-based heartbeat classification for arrhythmia detection: A survey [Journal Article]. Comput Methods Programs Biomed. 2016;127:144–64.
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016;2016-December:770–778.
24. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–255.
25. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps; 2014.
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017. p. 618–626.
27. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. 3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings. 2015;p. 1–14.
28. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Analytical Chemistry Research. vol. 12; 2014. p. 818–833. Available from: https://linkinghub.elsevier.com/retrieve/pii/S221418121630074Xhttp://link.springer.com/10.1007/978-3-319-10590-1{_}53.