Domain Knowledge Guided Deep Atrial Fibrillation Classification and Its Visual Interpretation

Xiaoyu Li
Xi'an Jiaotong University
Xi'an, Shaanxi, China
xiaoyuli@stu.xjtu.edu.cn

Buyue Qian
Xi'an Jiaotong University
Xi'an, Shaanxi, China
qianbuyue@xjtu.edu.cn

Jishang Wei
HP Lab
Silicon Valley, USA
weijishang@gmail.com

Xianli Zhang
Xi'an Jiaotong University
Xi'an, Shaanxi, China
xlbryant@stu.xjtu.edu.cn

Sirui Chen
Xi'an Jiaotong University
Xi'an, Shaanxi, China
chensirui@stu.xjtu.edu.cn

Qinghua Zheng
Xi'an Jiaotong University
Xi'an, Shaanxi, China
qhzheng@xjtu.edu.cn

ABSTRACT

Hand-crafted features have been proven useful in solving the electrocardiograph (ECG) classification problem. The features rely on domain knowledge and carry clinical meanings. However, the construction of the features requires tedious fine tuning in practice. Lately, a set of end-to-end deep neural network models have been proposed and show promising results in ECG classification. Though effective, such models learn patterns which usually mismatch human's concept, and thereby it is hard to get a convincing explanation with interpretation methods. This limitation significantly narrows the applicability of deep models, considering it is difficult for cardiologists to accept the unexplainable results from deep learning. To alleviate such limitation, we are bringing the best from the two worlds and propose a domain knowledge guided deep neural network. Specifically, we utilize a deep residual network as a classification framework, within which key feature (P-wave and R-peak position) reconstruction tasks are adopted to incorporate domain knowledge in the learning process. The reconstruction tasks make the model pay more attention to key feature points within ECG. Furthermore, we utilize occlusion method to get visual interpretation and design a visualization at both heartbeat level and feature point level. Our experiments show the superior performance of the proposed ECG classification methods compared to the model without P-wave and R-peak tasks, and the patterns learnt by our model is more explainable.

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357998

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Applied computing** → **Health informatics**; • **Human-centered computing** → *Visual analytics*.

KEYWORDS

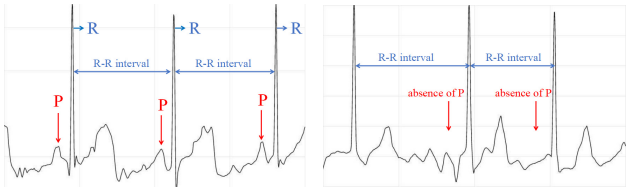
ECG Classification; Deep Residual Network; Visual Interpretation; Occlusion Method

ACM Reference Format:
Xiaoyu Li, Buyue Qian, Jishang Wei, Xianli Zhang, Sirui Chen, and Qinghua Zheng. 2019. Domain Knowledge Guided Deep Atrial Fibrillation Classification and Its Visual Interpretation. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357998

1 INTRODUCTION

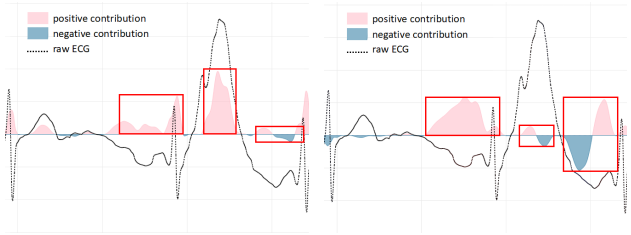
Electrocardiogram monitors hearts' electronic activities, which has been proved as an indispensable tool for diagnosis of cardiovascular diseases. In practice, cardiologists review ECG signals at various time intervals to identify abnormal heart activities. This labor-intensive process is prone to mistakes and usually suffers from inter- and intra-physician variability. Automatic detection systems have been developed to alleviate the aforementioned difficulties and assist doctors in revealing arrhythmia.

A variety of machine learning algorithms have been designed to advance automatic arrhythmia detection[5, 14, 23, 30, 32]. Conventional approaches train classifiers, such as random forest[2, 11] and support vector machine (SVM)[26], with hand-crafted features. Feature design and selection is nontrivial, which requires in-depth domain knowledge and is often involving a try-and-error procedure. Most recently, deep neural networks has emerged as an end-to-end solution that train classifiers directly on raw ECG data[17, 18, 31]. Although deep neural network has made remarkable progress in improving detection accuracy and convolutional neural network (CNN) is reported to have achieved cardiologist level performance[19], it is not clear why CNN makes a specific decision. Outcome prediction without a clear explanation is difficult for acceptance in a practical clinical scenario. Though



(a) Typical normal ECG (b) Typical AF ECG

Figure 1: Illustration for P-wave and irregular heart rhythm. There are three heart beats for each example, with red notation for P-wave and blue notation for R-peak. In the AF ECG, there is the absence of P-wave, and R-R intervals are not in equal length which means the irregular heart rhythm.



(a) Interpretation from original ResNet (b) Interpretation from our model

Figure 2: Example of interpretation resulted from the original ResNet and our model. The black line denotes raw ECG signal, and area map denotes support score from model for current decision of AF, with positive value in red area to support and negative value in blue area to blame. We draw additional red bounding box to stress key areas. We can conclude that the original ResNet pays much more attention to the raised region with high gradients purely driven by data, and our model pays more attention to the absence of P-wave, which is a significant feature for AF diagnosis.

we can get contribution map of input through various interpretation methods[28, 29, 33], patterns learnt by purely data-driven deep learning model often mismatch human's conception.

Our work aims at empowering deep neural networks with both performance improvement and the ability to capture more explainable pattern by incorporating feature reconstruction task to a classification neural network. During training, the classifier learning is guided by the domain reconstruction process. After training, we utilize a heartbeat-aware occlusion method[33] to the classification network to find decision support on the input raw data for decision explanation. There are a few challenges associated with such an approach. First, the reconstruction of which features is domain dependent. Combining the feature reconstruction with the classification task requires special treatments such as regarding R-peak and P-wave reconstruction as object detection problem. Second, it is important to maintain or boost the performance of the main classification task while introducing the feature construction tasks, because improperly setted structure for construction task often makes the classification task perform worse. Thirdly, original occlusion methods are used to cover up different portions of the scene with a gray square and compare how the output changes. The square is usually set as certain length and slides with certain step to get feature-point-level contribution. But we also need proper high-level contribution to guide users' attention for interactive visualization, because the raw ECG is ususally too long.

In our work, we illustrate the idea of model design in the context of detecting Atrial Fibrillation (AF), a specific type of arrhythmia. In AF detection, cardiologists specifically seek a few patterns, such as the absence of P-wave and the irregular heart rhythm, to determine AF. Illustration of the absence of P-wave and irregular heart rhythm is shown as Fig.1.There are three heart beats for each example, with red notation for P-wave and blue notation for R-peak. In the AF ECG, we can see the absence of P-wave and R-R intervals are not in equal length which means the irregular heart rhythm. We mimic the procedure for cardiologists to judge AF — first force model to learn what and where the key feautres are(the P-waves and R-peaks), and then learn to classify AF with the features in mind. Specifically, we incorporate P-wave and R-peak detection, which are regarded as object detection problems, in the AF detection problem. A 34 layer residual network (ResNet)[6, 19] is used for AF detection and feature maps from certain layers of ResNet are extracted to predict peak position and confidence. To choose appropriate layers for P-wave and R-peak detection, we consider different abstraction level for P-wave and R-peak respectively, and consider the evidence that it is easier to find P-waves based on R-peak in the meantime. We first pretrain our model with only the P-wave and R-peak detection problem, and then add the main classification problem to finetune. After training and evaluation, for visual interpretation, we also use two kinds of occlusion methods, the original one, and the heartbeat-aware one, to get interpretation for visualization. We use the former to provide detailed contribution of each feature point. Then we modify the occlusion method to be able to mask R-R intervals, in which mode the region of interest can match cardiologists' mental concept. We apply both occlusion methods to produce visual interpretation and compare the difference between the interpretation result from the best of our models and the original ResNet. A typical example is shown as Fig.2 — an abnormal wave is surrounded by right two R-peaks, and the region before R-peak, where there is the absence of P-wave, matters most. From the Fig.2, we can conclude that our model concentrates more on the guided P-waves, which is a significant feature cardiologists appreciate.

In summary, the major contributions of this paper are as follows:

- We introduce a domain knowledge guided approach to empower deep neural network to capture patterns which are more explainable, and improve performance in the meantime for AF classification problem. Since the additional labels, detected by object detection method, are only some key feature points, such approach can be easily applied in other domains.
- We present a new type of visual representation for outcome interpretation on ECG data, which is based on the original occlusion and the heartbeat-aware occlusion method. With such representation, we visualize and analyze the patterns captured by different models.

2 RELATED WORK

We propose a domain knowledge based AF classification methods, regarding waves reconstruction as object detection task and providing interpretation with visualization of region of interest. Thus, in this section, we mainly introduce related ECG classification methods, object detection methods and interpretation methods.

The most related is AF classification methods. Conventionally, hand-crafted feature based methods are widely developed and still keep the state-of-the-art performance in the situations with limited data[4, 5, 16, 27, 30, 32]. The five top teams in CinC/Challenge 2017, respectively extract 79, 62, 150, more than 150 and more than 600 features and get final champion together with blind test F1 measure of 0.83. Such kind of methods work beyond the data limitation with domain knowledge, but it is time consuming to engineer so much features and the process strong relies on knowledge and experience. With deep neural network getting popular with its data-driven automatic abstract ability, end-to-end deep model is getting more promising. [19] proposes a 34 layers deep convolutional network in cardiologist level with their collected data. [22, 31, 34] also use similar convolutional neural network or recurrent neural network based methods to solve the AF classification problem from time domain or frequency domain. These methods show the potential of the strong representation ability but limited with data amount, lacking interpretation. It is worth mentioning that [**?**] uses recurrent network with attention mechanism to provide beat level explanation, showing the weight of every beat successfully. It's impressive and we step further to provide beat level and pointwise interpretation in the meantime.

It does not help much to solely present the result from deep neural network, because doctors just do not accept that[12]. Some methods among the deep interpretation methods have shown the ability to indicate which part of the input strongly influence the result. [25] proposes to use the gradient of the output with respect to pixels of an input image to compute a "saliency map" of the image in the context of image classification, which is similar to deconvolutional network[33]. Here, we could treat an ECG signal of certain length as a 1*n image. Then, [28] combines Simonyan's approach and deconvolutional network into guided backpropagation, which zero's out the importance signal at a ReLU if either the input to the ReLU during the forward pass is negative or the importance signal during the backward pass is negative. [29] proposes integrated gradients where they integrate the gradients as the input are scaled up from some starting value (e.g. all zeros) to their current value. [24] proposes DeepLIFT, a method for decomposing the output prediction of a neural network on a specific input by backpropagating the contribution of all neural in the network to every feature of the input, using a "difference-from-reference". Occlusion method is first proposed by [33] to visualize the response of deep neural network to a specific input as a perturbation based method. [35] optimizes the occlusion method with a more rigorous approach at both removing information from the input and effect evaluation of such information removal, providing better additional insight. It is reported that occlusion method better identifies the few most important feature than many gradient-based methods and the method in [35] is much slower than the original occlusion method[1]. Thus, we use and modify occlusion method from [33], adapting the method for AF diagnosis, masking information dynamically with R-R intervals, providing beat level and pointwise level contribution of input in the meantime.

To get interpretation more intuitive, we introduce auxiliary task of peak reconstruction where we treat a peak point as an object and reformulate the task as object detection problem. With limitation of conventional algorithm for feature extraction, here we only detect the highest point of peaks. For object detection, many typical methods[7, 13, 21] treat the input as many grids with map relation with the length of the last feature map. We take the similar manner to detect P-wave and R-peak respectively.

3 METHODOLOGY

3.1 Classification Problem Formulation

To diagnose AF, cardiologists usually observe two waveforms: R-peak and P-wave. We take AF diagnosis as a typical classification task and the detection of P-wave and R-peak as two object detection tasks. We define R-peak and P-wave detection as locating the highest feature points of R-peak and P-wave in ECG data. Our model takes raw ECG signal $S = [s_1, ..., s_n]$ as input and outputs one-hot class label $Y = [y_1, ..., y_N]$ with two-tuples (x, c) of certain number. We treat ECG signal as multiple grids and detect P-wave and R-peak in every grid with a two-tuple (x, c), x for position index of the R-peak or P-wave to perform regression and c for confidence to perform prediction, referring to YOLOv1[20]. We use categorical cross entropy loss for classification and mean square error loss for detection respectively. Then, we combine classification loss and detection loss to define the final objective function as below:

$$L_{classification}(S, Y) = \frac{1}{N} \sum_{i=0}^{N} logp(Y = y_i|S) \qquad (1)$$

$$
\begin{aligned}
L_{detection}(S,(X,C)) = \lambda_{coord} \sum_{i=0}^{M} 1_i^{obj}(x_i - \hat{x}_i)^2 \\
+ \sum_{i=0}^{M} 1_i^{obj}(c_i - \hat{c}_i)^2 \\
+ \lambda_{noobj} \sum_{i=0}^{M} 1_i^{noobj}(c_i - \hat{c}_i)^2
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
L_{final}(S,Y,(X,C)) = & L_{classification}(S,Y) \\
& + \frac{1}{2} L_{detection-P}(S,(X_P,C_P)) \\
& + \frac{1}{2} L_{detection-R}(S,(X_R,C_R))
\end{aligned} \tag{3}
$$

3.2 Model Architecture and Framework

We aim to incorporate domain knowledge and guide learning process by addition key feature reconstruction tasks which are solved with object detection method. The intuition behind such idea is as follows: since experts judge instances by certain key features, it shall help model to classify instances by learning and keeping in mind what and where these key features are, and it is exactly what object detection method does. Thus, we first pretrain our model with only key feature reconstruction tasks and then finetune the whole model with both feature reconstruction tasks and classification task.

Our network model is shown as Fig.3. We use convolutional network with shortcut connections which is similar to Residual Network[8]. We set filter length of 64 and filters number of 64 * k for each convolutional layer, where k starts as 1 and increments every 4 block. We set kernel size of convolutional layer as 16, and max pooling size as 2. We apply Batch Normalization, Dropout and the pre-activation block design, which is similar to [19] and [6].

In the middle of the network, there are ConvBlocks shown as in yellow box. We only add the Maxpooling layers to ConvBlocks at odd orders. After fifth ConvBlocks and sixth ConvBlocks, we respectively add a branch with convolutional layer and Sigmoid layer to detect R-peaks and P-waves. We set the P-wave branch behind the R-peak because we usually easily notice R-peaks and find P-waves based on R-peak locations which means R-peak detection should help P-wave detection. The filters number of the last convolutional layer in the branch is 2 and kernel size is 1, mapping the feature map into only 2 channels. Thus, we get a tensor with the shape of (b, f, 2), where b denotes batch size, f denotes length of the feature map, 2 means length of the tuple (x, c). In this way, we treat the raw input as f grids according to the linear transformation for the convolutional network. For each grid, we predict whether there is a R-peak in the grid or not with value c as confidence, and perform regression for R-peak feature point position x, shown as Fig.4. With the grids, we can only predict the relative position in the grid instead of absolute position, making the transformation relatively easier for the network. We take the official data in

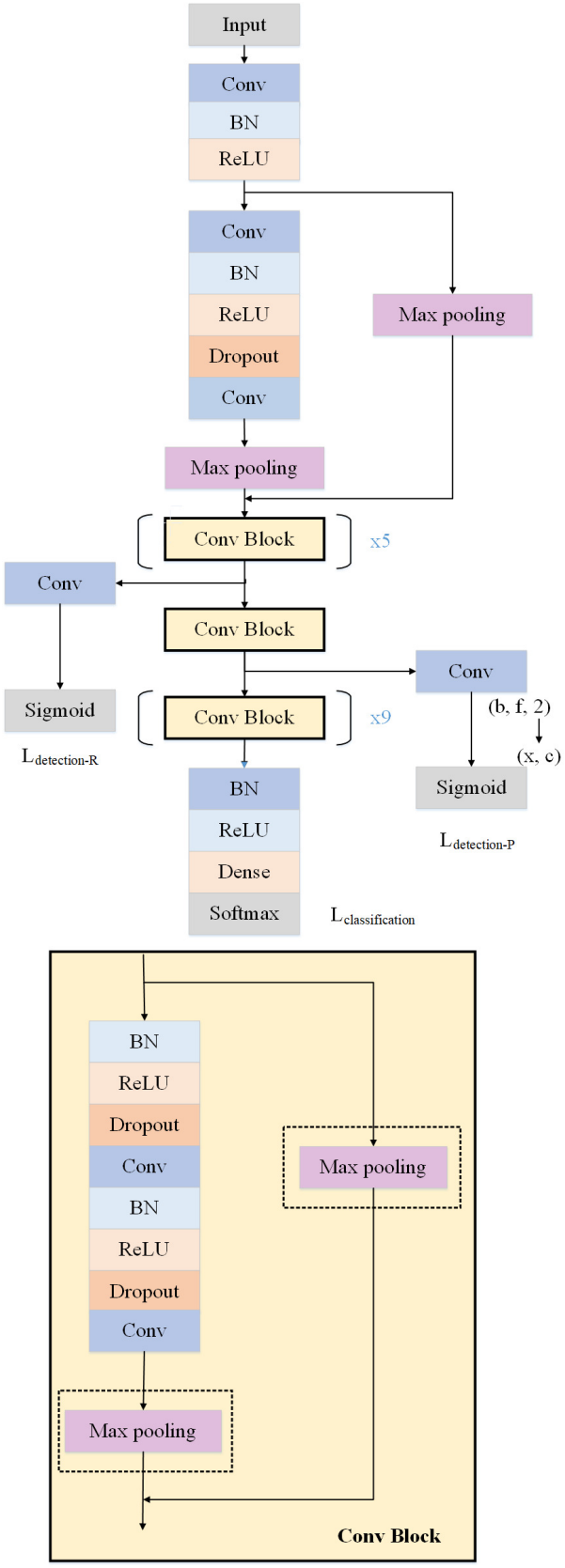


Figure 3: Overview of our network

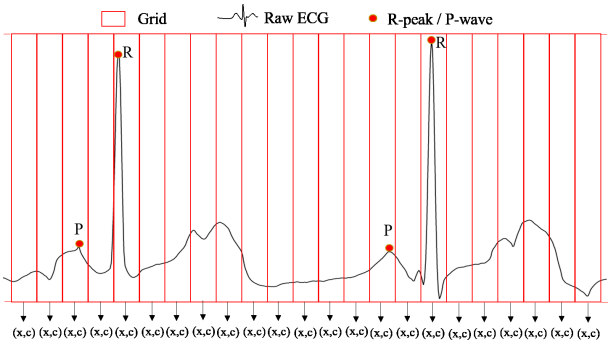


Figure 4: Regarding raw ECG as multiple grids, and detecting P-waves and R-peaks in each grid.

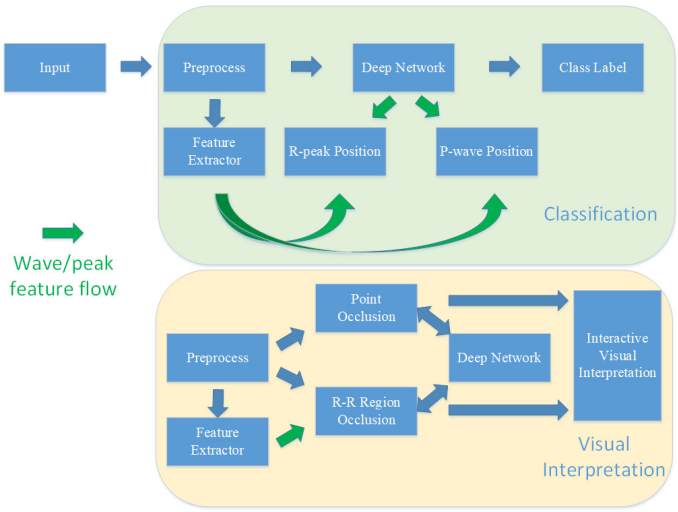


Figure 5: Overview of our framework

CinC/Challeng 2017[3] as an example. The ECG data ranges from 9 seconds to 60 seconds with sampling frequency of 300hz and we pad each ECG into a vector with length of 71 * 2^8 (18176) with zeros. The number of heart beats is roughly less than 180 and l here is 1136, which means negative instances are much more than positive ones. Thus, we set $\lambda_{noobj} = 0.5$ to penalize the confidence of negative instances and $\lambda_{coord} = 5$ to encourage R-peak feature point position regression. In such procedure, it is a key step to decide where to set the two branches. We choose certain positions for the two branches based on the following consideration. Firstly, considering that ResNet builds up a hierarchy of features from the first layer until the last layer, we believe the R-peak and P-wave reconstruction tasks need to levarage some intermadiate layers. We do not choose first layers because that the features captured by first layers are too concret, with only some basic shapes usually. We also do not choose last layers because the features captured by last layers are enough for classification which means these features have already contained the high-level information based on the "key features". Thus, we choose intermediate layers as a start point. Secondly, we consider the relation between R-peak and P-wave. Actually, R-peak is easier to recognize than P-wave. Based on R-peak, it also will be easier to find P-wave because P-wave always show up a bit ahead of R-peaks. Thus, we set R-peak branch before P-wave branch. Then, after only several trials, we get a proper branch positions setting.

As for the main classification task, there are still 9 ConvBlocks left, followed by Batch Normaliztion, Relu, dense layer and a softmax layer. We use the output from the softmax to perform the final classification. To get better initialization, we pretrain the whole model without the main classification task first. Then we add the classification task back and take the same manner to finetune.

At the training stage of the above model, R-peak and P-wave positions are required as ground truth. But these tags are not available in the most dataset. Alternatively, we utilize conventional methods from [10, 15] to extract R-peak and P-wave positions as shown in the "classification" part of our framework, Fig. 5. The same R-peak position tags will also be sent to our heartbeat-aware occlusion method.

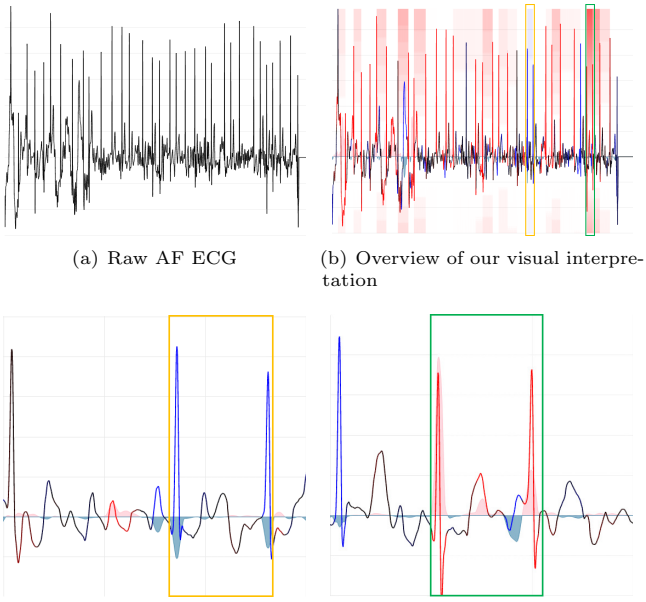
3.3 Visual Interpretation

After the classification procedure, we also provide visual interpretation to explain that how each part of ECG influences the final classification result. In this part, we utilize the occlusion method, a perturbation-based interpretation approach, to calculate the contribution of each feature points to the final outcome. The original occlusion method directly masks input with a sliding window with certain length and step. The occluded data runs through the model to generate a result, which is compared against the result from the original input to determine the contribution of the occluded area, which means we can easily get the contribution of each feature point of raw ECG. It is reported that the contribution usually focus on the main subject only when using bigger patches[1], so we set the patch length of 15 and the step size of 1 to get feature-point-level contribution. Then, there is a problem that the whole ECG is quite long and the feature-point-level contribution is usually too dense for cardiologists to concentrate on important regions. Thus, we further modify the original occlusion method. Since we are able to get R-peaks position tags, which means we can get each R-R intervals, we dynamically assign the length of R-R intervals to the patches and steps to get the contribution of each heart beat. Then, with some visual mapping, we can guide cardiologists' attention with beat-level contribution and provide detail with feature-point-level contribution through visualization techniques.

We design a new visual representation for interpretation and build a tool for cardiologists to check result from deep model. We'll illustrate our visual encoding and the effectiveness of our visualization with an example shown as in Fig.6. The ECG is labeled with "AF" and the contribution is from baseline (34 layers ResNet without additional branch).

Though the outcome is also "AF", cardiologists can not directly take the result from a black box, without checking important regions. When a cardiologist start checking, the overview of raw ECG just looks like a mess, shown as Fig.6(a). It is hard to get valuable information with the whole raw ECG itself quickly. But in our visualization, there is a background as in Fig.6(b), consisting of rectangles of red and blue in different depth, with opacity of 0 in the middle with gradient to avoid chromatic aberration for ECG signal visualization. We encode the contributions of each R-R intervals to the color of the rectangles. The "red" and "blue" respectively indicate whether this region supports or blames the model's decision, and the depth of color indicates how important the model makes decision depending on the region. It's easy for a cardiologist to select their region of interest. Then, the cardiologist can pan and zoom to concentrate on the regions he may be interested in. In this example, the two typical regions are shown in the green (the most positive) and orange (the most negative) bounding boxes respectively. Pan and zoom to focus on any of them, then it comes to the next two figures.

As shown in Fig.6(c) and Fig.6(d), we also visually encode the feature-point-level contribution into the gradient color of raw input, red for "positive", blue for "negative" and black for zeros. Although it is convenient for cardiologists to check directly, comparison between feature points seems intractable. It's usually hard to tell whether a feature point are colored with more or less intensity. Thus, we further encode the contribution spatially. We draw an area map with the feature-point-level contribution, corresponding to each feature point in the line of raw input. Above the zero axis, the area map is red, and below the zero axis is blue. The height of one feature point in area map means how much the contribution is. Beside, the area size of a certain region in the area map indicates how important the region is to the classification result. Thus, cardiologists only need to concentrate on the region with larger areaes. The Fig.6(c) shows the region of the orange bounding box in the overview and the Fig.6(d) shows the region of the green one. Unfortunately, since the outcome is from beseline, we see that the most influential part of both figures are R-peaks, without any relationship with the absence of P-waves. If we replace the baseline with our model, we can see the remarkable evidence for "AF" of the absence of P-wave with a large red area map, as shown in Fig.9(c).



(a) Raw AF ECG

(b) Overview of our visual interpretation

(c) Region that mostly blames current decision

(d) Region that mostly supports current decision

Figure 6: The Fig.6(c) and Fig.6(d) are the corresponding regions for the bounding box with highest contribution to support and blame current decision respectly, showing the confusing pattern learnt by original ResNet

Table 1: Data Summary

Type	#recording	**Table Column Head**				
		Mean	***SD***	***Max***	***Median***	***Min***
Normal	5154	31.9	10.0	61.0	30	9.0
AF	771	31.6	12.5	60	30	10.0
Other	2557	34.1	11.8	60.9	30	9.1
Total	8482	32.5	11.4	61.0	30	9.0

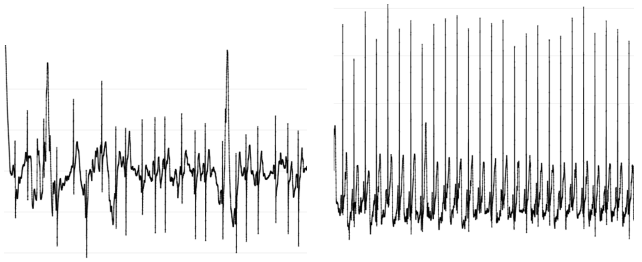
4 EVALUATION

4.1 Experiment Setting

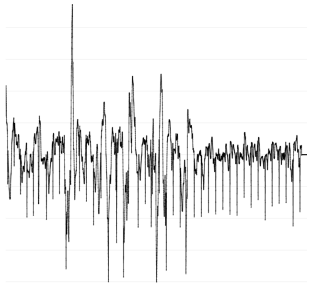
Large ECG dataset with label is hard to get. The data collected by [19] is large enough but has not been released yet. To get dataset as large as possible, we use the data shared by physionet/Challenge 2017[3], including recordings lasting from 9s to 61s donated by AliveCor. The ECG data is with frequency of 300 Hz, bandwidth of 0.5-40 Hz and a $\pm$ 5 mV dynamic range. Each ECG record corresponds to an individual. We remove 46 ECG data of "Noisy" class, and the left includes 5154 "Normal" data, 771 "AF" data, and 2557 "Other rhythm" data, shown as Table.1.

We utilize conventional algorithms to get P-wave[15] and R-peak[10] point positions for each recording as ground truth for detection. We use raw ECG data as direct input and split the whole data as 8:1:1 into training set, validation set and test set. As a trick to alleviate the training problem caused by the imbalance, we duplicate the data of "AF" and "Other rhythm" to balance the training set.

We set baseline as 34 layers residual network, the same model as [6]. We directly train the baseline with Adam optimizer with initial learning rate of 0.001 and decay learning rate by a ratio of 0.1 when validation loss does not get lower continually for 3 epoches. We initialize the weights of convolutional layers as in [9]. As for our model, we take the same manner to train. The only difference is that we firstly



(a) Typical instance for "AF" (b) Typical instance for "Normal"



(c) Typical instance for "Other"

Figure 7: Typical examples of raw ECG respectively from four classes. The instances of "AF" are of irregular heart rhythm with absence of P-wave in heart beats. The instances of "Other" include any other arrhythmia.

pretrain our model without the main classificatin branch, and then add the main branch back to finetune. Since our additional branch can be optionally added, we try 3 settings for our model: *Model-R* with branch for R-peak, *Model-P* with branch for P-wave, and *Model-RP* with both branches. We evaluate each model with 5 fold cross validation for 5 times and get the average metrics, including accuracy, average precision, recall and F1.

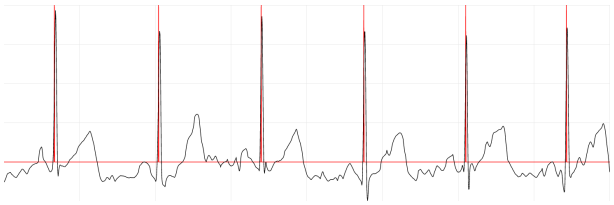
After the training process, we pick the best model of baseline and our models respectively trained with the same training set. Besides above metrics, we also do two kinds of evaluation to prove that the pattern captured by our model is more convincing. Firstly, we demonstrate how the pattern our model learned is more explainable with the occlusion methods and visual representation clarified. Secondly, to supplementarily prove the example is not only typical one, we also provide objective analysis at the end.

4.2 Comparison of Performance

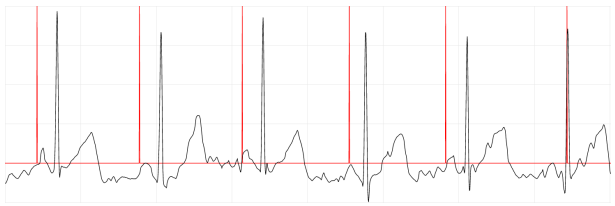
The experiment result shows the our knowledge guidance works well, shown as Tabel.2. As we mentioned, the absence of P-wave and the irregular heart rhythm are signifi t patterns for AF classification according to cardiologists. R-peak is helpful, but not as much as the absence P-wave. Thus, all of our models outperform baseline and *Model-P* performs better than *Model-R*. Besides, the performance of our model

Table 2: Classification result for each model

Model	Accuracy	Precision	Recall	F1
Baseline	0.8143	0.7836	0.7703	0.7707
Model-R	0.8229	0.7951	0.7813	0.7820
Model-P	**0.8314**	**0.8042**	**0.7960**	**0.7959**
Model-RP	0.8296	0.8019	0.7922	0.7916



(a) Raw input and extracted R-peak



(b) Raw input and extracted P-wave

Figure 8: Example for feature extractor. The red line indicates result from feature extractor for R-peak and P-wave, with black line as raw ECG input. The Fig.8(a) shows R-peak positions are extracted quite well but the Fig.8(b) shows the P-wave positions are just roughly right.

Model-RP is between *Model-R* and *Model-P*. This should be due to the distraction caused by the guidance from both R-peak and P-wave detection task. The experiment result perfectly matches medical conception and the performance is improved by the additional domain knowledge.

However, the performance of our models is still limited by the robustness of the feature extractor in the experiment, because the ground truth of our R-peak and P-wave tags may be not true. Visualizing the result from the feature extractor, we find that the R-peak positions are extracted quite well but P-wave positions are just roughly right, shown as Fig.8 for example, ranging about 5 to 50 points behind the R-peak, where the right is the front. Conventional methods are much better at R-peak extraction than P-wave extraction, because R-peaks mostly will not disappear and show up with high gradient but P-wave can be absent usually in abnormal ECG. This backward could be alleviated if more P-wave tagged data collected and deep neural network applied to purely detect P-wave. Consequently, if we get accurate ground truth, our model would perform better. Besides, our original intention is to maintain or boost the performance of original ResNet, and

what's more important is that our model can capture patterns which are more explainable, shown as the next subsection.

4.3 Comparison of Visual Interpretation

We compare the visual interpretation of our model and baseline. Above all, we take an "AF" instance with right outcome and compare the corresponding visual interpretation from *Model-P* with the interpretation from baseline. For comparison, we respectively choose the regions that support current outcome mostly, from both visual interpretations, and pan and zoom into the two regions in both interpretations.

The visualization result is shown as Fig.9. The Fig.9(a) and Fig.9(d) are the overview respectively from the *Model-P* and baseline. The region in the orange bounding box is the R-R interval that *Model-P* thinks most important. The region in the green bounding box is the R-R interval that *Baseline* thinks most important. We select certain regions and get the corresponding magnified details to compare. We also use red dash bounding box to stress important area for comparison. As we clarified, the area map and the color indicate the contribution to the model's current outcome. The blue stands for "negative" and the red stands for "positive". In the Fig.9(e), it shows the highest contribution on the left red R-peak. But in the Fig.9(b), there is only a little contribution on the other red R-peak by contrast, which means the pattern *Model-P* learned is different from baseline. In the Fig.9(c), there is a high red peak area on the left of the right R-peak. The red area is exactly corresponding to the region of input where the P-wave is absent. In contrast, the red area in the same region from baseline is much smaller, as shown in the Fig.9(f). That is to say, although the baseline has noticed the absence of P-wave, such information is drown by other pattern and our *Model-P* pays much more attention to the absence of P-wave by contrast. Compared to the baseline, the pattern our *Model-P* learned is more explainable intuitive.

We also provide supplementary objective evaluation for qualitative analysis. We calculate Kullback-Leibler divergence between the distribution of feature-point contribution and the distribution of extracted P-wave positions, for each test set and model, with average result shown as Table.3. The lower KL divergence means current model pays more attention to P-waves. From this Table.3, the model with best performance in Table.2, the Model-P(the model with only P-wave reconstruction task), get the lowest KL divergence as expected, which means that our model can show better visual interpretation than baseline statistically.

Further more, we also cooperate closely with cardiologists from the hospital affiliated to our university. We ask them to use the AF classification tool and elecit their feedback. From their feedback, we can conclude two points. First point is that, although the model fails sometimes, the tool based on *Model-P* still works better than the tool based on the *original model* generally. The other point is that it is quite acceptable and attractive to provide visual interpretation for cardiologists rather than just providing outcome from a black-box model. They are more confident with the result with our visual interpretation.

Table 3: KL divergence between the distribution of feature-point contribution and the distribution of P-wave positions for each model

Model	Original	Model-R	Model-P	Model-RP
KL	6.2028	6.2348	6.080	6.094

5 DISCUSSION

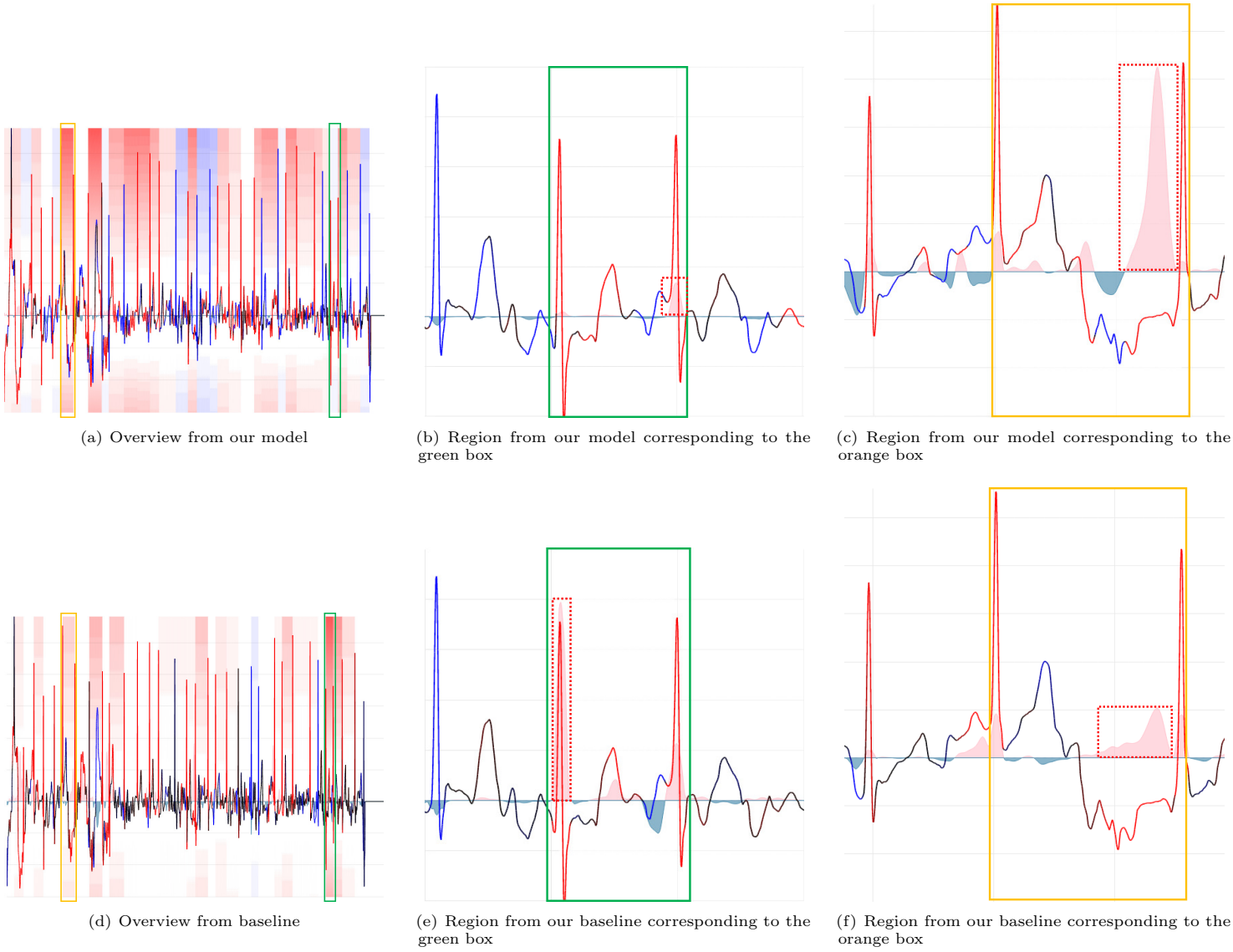
During the comparison of visual interpretation, we find that the original model mostly makes decision mainly with the R-peak. This may be due to the importance of R-peak for all of the other arrhythmia and that the training data is not enough for the model to notice enough significance of the absence of P-wave. By contrast, our model usually decides with P-wave for instances of AF class.

In addition, the object detection method can be utilized to detect other key features in other problem, which means that the incorpation of object detection method makes the model more generalizable to encoding other experts' domain knowledge. Thus, our approach to guide learning process with domain knowledge, is not limited in AF classification, and can be adopted in many areas. For example, for EEG(Electroencephalogram) disease classification, the reconstruction of spike wave position should help detect Lennox-Gastaut syndrome if spike wave labeling is available, and for human activity recognition task, reconstruction of hand waving position should also help model classify whether the subject is sitting or running.

There is a problem to further clarify — can we be sure boost the performance of the main classification task while introducing feature reconstruction task? Regret to tell, our answer is "No". The correlation between accuracy and comprehension of visual interpretation should be empirically positive in most situation where the features are fundamental evidence for the classification task, but such correlation has not been proven theoretically. An important way to alleviate such problem is the participation from experts who exam visual interpretation and make final decisions. That's why we combine object detection method and visualization techniques to improve model performance and provide better visual interpretation in the meantime.

6 CONCLUSION

We utilize domain knowledge to guide learning process of deep model to capture patterns which are more explainable and boost performance in the meantime for AF classification problem. Specifically, we add key feature points (P-wave and R-peak) reconstruction task, which is solved as an object detection problem, to incorporate domain knowledge in deep residual network with additional branches. The guidance of P-wave and R-peak forces the network pay more attention to certain key feature points which helps both classification and interpretation. In addition, we propose a new visual



(a) Overview from our model

(b) Region from our model corresponding to the green box

(c) Region from our model corresponding to the orange box

(d) Overview from baseline

(e) Region from our baseline corresponding to the green box

(f) Region from our baseline corresponding to the orange box

Figure 9: Comparison of visual interpretation between Model-P and Baseline. The upper two figures are the overviews and others are magnified region corresponding to the orange and green bounding box. From the area map in red dash bounding box in the Fig.9(e) and Fig.9(c), we can know that, *Baseline* pays more attention to R-peak and our *Model-P* pays much more attention to the absence of P-wave by contrast.

representation based on original occlusion and heartbeat-aware occlusion method to interpret the outcome from deep models for ECG data. Our evaluation shows the effectiveness of such domain knowledge guidance in deep residual network for AF classification. Future work includes trying our method in other applications with similar nature.

7 ACKNOWLEDGMENTS

ACKNOWLEDGMENTS

This work is sponsored by the National Key Research and Development Program of China with grant number with grant number 2018YFC130078; the "Reliability and Persistence of Big Data Hybrid Storage" project under the National Key Research and Development Program of China with grant number 2016YFB1000303; the "Multi-model Based Patient Similarity Learning for Medical Data Modelling and Learning" project under National Natural Science Foundation of

China General Program with grant number 61672420; the Project of China Knowledge Center for Engineering Science and Technology; the National Natural Science Foundation of China InnovationResearch Team No. 61721002; Ministry of Education Innovation Research Team No. IRT13035; the Key Project of Natural Science Foundation of China under grant No. 61532015.

REFERENCES

[1] Marco Ancona, Enea Ceolini, Cengiz ztireli, and Markus Gross. 2017. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. (2017).

[2] Christoph Hoog Antink, Steffen Leonhardt, and Marian Walter. 2017. Fusing QRS Detection, Waveform Features, and Robust Interval Estimation with a Random Forest to Classify Atrial Fibrillation. In *Computing in Cardiology Conference*.

[3] Gari Clifford, Chengyu Liu, Benjamin Moody, Li Wei Lehman, Ikaro Silva, Qiao Li, Alistair Johnson, and Roger Mark. 2017. AF Classification from a Short Single Lead ECG Recording: the Physionet Computing in Cardiology Challenge 2017. In *Computing in Cardiology Conference*.

[4] Erin E Coppola, Prashnna K Gyawali, Nihar Vanjara, Daniel Giaime, and Linwei Wang. 2017. Atrial Fibrillation Classification from a Short Single Lead ECG Recording Using Hierarchical Classifier. In *Computing in Cardiology*.

[5] Shreyasi Datta, Chetanya Puri, Ayan Mukherjee, Rohan Banerjee, Anirban Dutta Choudhury, Rituraj Singh, Arijit Ukil, Soma Bandyopadhyay, Arpan Pal, and Sundeep Khandelwal. 2017. Identifying Normal, AF and other Abnormal ECG Rhythms using a Cascaded Binary Classifier. In *Computing in Cardiology Conference*.

[6] Marco A F Pimentel Adam Mahdi Maarten De Vos Fernando Andreotti, Oliver Carr. 2017. Comparing Feature Based Classifiers and Convolutional Neural Networks to Detect Arrhythmia from Short Segments of ECG. In *Computing in Cardiology Conference*.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. 2017. Mask R-CNN. *IEEE Transactions on Pattern Analysis & Machine Intelligence* PP, 99 (2017), 1–1.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. (2015), 770–778.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. (2015), 1026–1034.

[10] P. Kathirvel, M. Sabarimalai Manikandan, and K. P. Soman. 2011. An Efficient R-peak Detection Based on New Nonlinear Transformation and First-Order Gaussian Differentiator. *Cardiovascular Engineering & Technology* 2, 4 (2011), 408–425.

[11] Martin Kropf, Dieter Hayn, and Gunter Schreier. 2017. ECG Classification Based on Time and Frequency Domain Features Using Random Forrests. In *Computing in Cardiology Conference*.

[12] Zachary C. Lipton. 2017. The Doctor Just Won't Accept That! (2017).

[13] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. 2015. SSD: Single Shot MultiBox Detector. (2015), 21–37.

[14] Ruhi Mahajan, Rishikesan Kamaleswaran, John Andrew Howe, and Oguz Akbilgic. 2018. Cardiac Rhythm Classification from a Short Single Lead ECG Recording via Random Forest. In *Computers in Cardiology*.

[15] D. Makowski. 2016. NeuroKit: A Python Toolbox for Statistics and Neurophysiological Signal Processing (EEG, EDA, ECG, EMG...). (2016).

[16] Filip Plesinger, Petr Nejedly, Ivo Viscor, Josef Halamek, and Pavel Jurak. 2017. Automatic Detection of Atrial Fibrillation and Other Arrhythmias in Holter ECG Recordings using PQRS Morphology and Rhythm Features. In *Computing in Cardiology Conference*.

[17] Bahareh Pourbabaee, Mehrsan Javan Roshtkhari, and Khashayar Khorasani. 2017. Deep Convolution Neural Networks and Learning ECG Features for Screening Paroxysmal Atrial Fibrillatio Patients. *IEEE Transactions on Systems Man & Cybernetics Systems* PP, 99 (2017), 1–10.

[18] B Pyakillya, N Kazachenko, and N Mikhailovsky. 2017. Deep Learning for ECG Classification. In *Journal of Physics Conference Series*. 012004.

[19] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. *CoRR* abs/1707.01836 (2017). arXiv:1707.01836 http://arxiv.org/abs/1707.01836

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Computer Vision and Pattern Recognition*. 779–788.

[21] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6517–6525.

[22] Jonathan Rubin, Saman Parvaneh, Asif Rahman, Bryan Conroy, and Saeed Babaeizadeh. 2017. Densely Connected Convolutional Networks and Signal Quality Analysis to Detect Atrial Fibrillation Using Short Single-Lead ECG Recordings. (2017).

[23] Yuxi Zhou Qingyun Wang Junyuan Shang Hongyan Li Junqing Xie Shenda Hong, Meng Wu. 2018. ENCASE: an ENsemble ClASsifiEr for ECG Classification Using Expert Features and Deep Neural Networks. In *Computers in Cardiology*.

[24] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. (2017).

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Computer Science* (2013).

[26] Radovan Smisek, Jakub Hejc, Marina Ronzhina, Andrea Nemcov, Lucie Nemcova, Jiri Chmelik, Jana Kolarova, Ivo ProvazniK, Lukas Smital, and Martin Vitek. 2017. SVM Based ECG Classification Using Rhythm and Morphology Features, Cluster Analysis and Multilevel Noise Estimation. In *Computing in Cardiology Conference*.

[27] Dionisije Sopic, Elisabetta De Giovanni, Amir Aminifar, and David Atienza. 2017. A Hierarchical Cardiac Rhythm Classification Methodology Based on Electrocardiogram Fiducial Points. In *Computing in Cardiology Conference*.

[28] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for Simplicity: The All Convolutional Net. *Eprint Arxiv* (2014).

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. (2017).

[30] Toms Teijeiro, Constantino A Garca, Daniel Castro, and Paulo Flix. 2017. Arrhythmia Classification from the Abductive Interpretation of Short Single-Lead ECG Records. (2017).

[31] Zhaohan Xiong, Martin Stiles, and Jichao Zhao. 2017. Robust ECG Signal Classification for the Detection of Atrial Fibrillation Using Novel Neural Networks. In *Computing in Cardiology Conference*.

[32] Morteza Zabihi, Ali Bahrami Rad, Aggelos K. Katsaggelos, Serkan Kiranyaz, Susanna Narkilahti, Moncef Gabbouj, Morteza Zabihi, Ali Bahrami Rad, Aggelos K. Katsaggelos, and Serkan Kiranyaz. 2017. Detection of Atrial Fibrillation in ECG Hand-held Devices using a Random Forest Classifier. In *Computing in Cardiology*.

[33] Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. 8689 (2013), 818–833.

[34] Martin Zihlmann, Dmytro Perekrestenko, and Michael Tschannen. 2017. Convolutional Recurrent Neural Networks for Electrocardiogram Classification. (2017).

[35] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. (2017).