

Research and Applications

Evaluating gradient-based explanation methods for neural network ECG analysis using heatmaps

Andrea Marheim Storås , PhD^{1,2}, Steffen Mæland, PhD³, Jonas L. Isaksen, PhD⁴, Steven Alexander Hicks , PhD^{*1}, Vajira Thambawita, PhD¹, Claus Graff, PhD⁵, Hugo Lewi Hammer, PhD^{1,2}, Pål Halvorsen, PhD^{1,2}, Michael Alexander Riegler , PhD^{1,2}, Jørgen K. Kanters, MD^{4,6}

¹Department of Holistic Systems, SimulaMet, 0170 Oslo, Norway, ²Department of Computer Science, Oslo Metropolitan University, 0167 Oslo, Norway, ³Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway, ⁴Laboratory of Experimental Cardiology, University of Copenhagen, 2200 Copenhagen, Denmark, ⁵Department of Health Science and Technology, Aalborg University, 9260 Aalborg, Denmark, ⁶Center for Biological Research, University of California, San Francisco, San Francisco, CA 94143, United States

*Corresponding author: Department of Holistic Systems, SimulaMet, Stensberggata 27, 0170 Oslo, Norway (steven@simula.no)

Drs M.A. Riegler and J.K. Kanters contributed equally to the work.

Abstract

Objective: Evaluate popular explanation methods using heatmap visualizations to explain the predictions of deep neural networks for electrocardiogram (ECG) analysis and provide recommendations for selection of explanations methods.

Materials and Methods: A residual deep neural network was trained on ECGs to predict intervals and amplitudes. Nine commonly used explanation methods (Saliency, Deconvolution, Guided backpropagation, Gradient SHAP, SmoothGrad, Input \times gradient, DeepLIFT, Integrated gradients, GradCAM) were qualitatively evaluated by medical experts and objectively evaluated using a perturbation-based method.

Results: No single explanation method consistently outperformed the other methods, but some methods were clearly inferior. We found considerable disagreement between the human expert evaluation and the objective evaluation by perturbation.

Discussion: The best explanation method depended on the ECG measure. To ensure that future explanations of deep neural networks for medical data analyses are useful to medical experts, data scientists developing new explanation methods should collaborate tightly with domain experts. Because there is no explanation method that performs best in all use cases, several methods should be applied.

Conclusion: Several explanation methods should be used to determine the most suitable approach.

Key words: explainable artificial intelligence; machine learning.

Introduction

Electrocardiograms (ECGs) are time series of the electrical activities of the heart and are widely used clinically. Machine learning (ML) models, including complex deep neural networks, have shown impressive results on analyzing ECGs and can even detect cardiac findings that medical experts are typically unable to infer from the ECG alone.¹

Humans have difficulties understanding model decisions and lack of model explainability has been identified as a barrier for implementation of ML systems in medicine.² Gradient-based explanation methods can visualize feature importance using heatmaps.

Gradient-based explanations provide the regions of the data that are important to the model's prediction. These are often visualized as heatmaps to convey explanation to the end user. Many model explanation methods with different properties exist. Unfortunately, there are no clear guidelines about how to pick the most appropriate method, and currently, studies apply different explanation methods. Moreover, there is a tendency to

generate explanations without critically considering which methods, if any, are optimal for the specific medical use case. Our work aims to highlight these issues by thoroughly evaluating 9 of the most frequently applied explanation methods for explaining deep neural network models. We apply our evaluation on ECG data, but this is important for all medical image analyses.

Many of the proposed explanation methods are based on similar logic. In this work, we focus on gradient-based methods, which use the *gradient* of the prediction with respect to the input. A gradient is a derivative of a function with more than one variable. The gradient captures the local slope of the function, predicting the effect of taking a small step from a point in any direction. In that way, the gradient helps quantifying how a small change in the input changes the output. The differences between gradient-based methods lie in additional operations applied after or during the gradient computation that aim to improve and sharpen the explanations. The highlighted areas in the heatmaps based on the explanations

generated by the neural network, were compared with the correct segmentation masks annotated by medical experts.

The goal of the present study was to compare various explainability methods with objective and qualitative evaluation to determine which explanation method is preferred among experts. To our knowledge, this is the first time a comprehensive evaluation of explanation methods, including both qualitative and objective evaluation metrics, is described in the field of cardiology.

Materials and methods

Population

The study was performed on realistic digital ECGs, produced by a generative adversarial network.³ The generative model, named *Pulse2pulse*,⁴ was trained on real ECG data combined from the Danish General Suburban Population Study (GESUS)^{5,6} and the Inter99 study.⁷ Afterwards, the model was used to synthesize 150 000 10-s 12-lead ECGs, “sampled” at 500 Hz. The synthetic data not only covers the same distribution as the real data but has also shown to possess the same properties, effectively mirroring the characteristics of the actual ECG datasets. Using synthetic data rather than real patient data enhances privacy, increases the amount of training data available.

ECG analysis

The Marquette 12 SL algorithm (GE Healthcare, Chicago, IL, USA) was used for automated measurements and extraction of median beats⁸ from the generated synthetic ECGs. For the present study, we used 6 clinically relevant ECG features: 3 amplitudes (R-wave, J-point, and T-wave amplitudes—ie, to evaluate cardiac hypertrophy and ischemia) and 3 intervals (PR, QRS, and QT intervals—ie, to evaluate the risk of cardiac arrhythmias and eventually need for a cardiac pacemaker) as seen in Figure 1.

Machine learning models

ix convolutional neural networks were trained separately on the QT, QRS, and PR intervals and the T_{V5} , R_{V5} , and $J\text{-point}_{V5}$ amplitudes on the median ECG and 10-s ECG, respectively.

In all experiments, we used a *ResNet*-inspired⁹ convolutional network implemented in PyTorch version 1.9.1,¹⁰ structurally identical to the one described in reference,¹¹ but retrained for this study. Using the Adam optimizer and mean squared error as loss function, the models were trained for 100 epochs. The initial learning rate was set to 0.008, and the rate was reduced by a factor of 0.5 if the performance metric did not improve for 10 epochs. Sixty percent of the ECGs were used for training and 20% were used for validation. The remaining 20% of the data were reserved for the hold-out test set. This test set was used to evaluate the final models, using mean absolute error (MAE) and root mean squared error (RMSE) as evaluation metrics. The code is publicly available online at <https://github.com/smaeland/ecg-heatmaping-review>.

Heatmaps

Nine different explanation methods were applied (Table 1), implemented using the Captum library¹² based on the PyTorch framework.¹⁰ The heatmaps were created with 12 ECG leads (Full heatmap), but to reduce noise we also

generated averaged heatmaps. The average heatmaps were created by averaging over all 12 leads, and then presented on an overlay of ECG lead V5 to give a clean simple presentation. To avoid canceling out positive and negative attributions, we took the absolute values of the importance before averaging. The experts were instructed to evaluate which heatmap method was most understandable for a certain ECG measurement.

Below follows a brief description of each of the explanation methods.

*Saliency*¹³ computes the gradient of the output with respect to the input to create maps that highlight which pixels most affect the model's predictions.

*Deconvolution*¹⁴ uses transposed convolution operations to trace activations from output back to input, revealing which input features activate specific network layers.

*Guided backpropagation*¹⁵ modifies standard backpropagation by only allowing positive gradients to flow back to the input layer, highlighting the features that most activate the output.

*Smoothgrad*¹⁶ reduces noise in saliency maps by averaging the gradients of the input image perturbed with noise.

*Input \times gradient*¹⁷ aims to improve the visual sharpness of the heatmap by element-wise multiplying the gradient with the input data. In other words, each input feature is multiplied with its corresponding gradient. This way, the sign and strength of the input can be included.

*DeepLIFT*¹⁷ compares the activation of each neuron to a reference activation and assigns contributions based on the difference. For this method, it is required to choose the reference, which should be a baseline representing the absence of signal. We default to using a zero vector.

*Integrated gradients*¹⁸ attributes a model's prediction to its input features by computing gradients between a baseline and the actual input, integrating these to show each feature's contribution. We use a zero vector as the baseline in this study.

*Gradient SHAP*¹⁹ has similarities to Integrated gradients, but adds random noise to the input. The baseline for Gradient SHAP is a zero vector, as for the other methods.

*Grad-CAM*²⁰ is only applicable to convolutional networks. For a specific convolution layer, the corresponding gradient of the prediction is computed and multiplied with layer activations. Consequently, the resulting heatmap has the same dimensions as the feature map of the convolution layer, which is often smaller than the input dimensions. In order to obtain fine-grained heatmaps of correct dimensionality, the authors propose merging the method with Guided backpropagation to form *Guided GradCAM*.²⁰ This is done by upsampling the heatmap obtained by Grad-CAM to the size of the input and multiplying it element-wise with the gradients from Guided backpropagation.²⁰

Expert evaluation of heatmaps

The ECG features predicted by the deep neural network were evaluated by 3 experts (C.G., J.K.K., J.L.I.) with more than 5 years experience in ECG analysis and Deep Learning (Table 1). Each ECG expert examined representative heatmaps generated by the 9 explanation methods for each of the different ECG measures. The ECG experts were presented each explanation method (9) for each ECG measure (6) both for full heatmaps including all 12 leads and for averaged heatmaps (2). This process is visualized in Figure 2, which shows the process from extracting the ECGs from the source

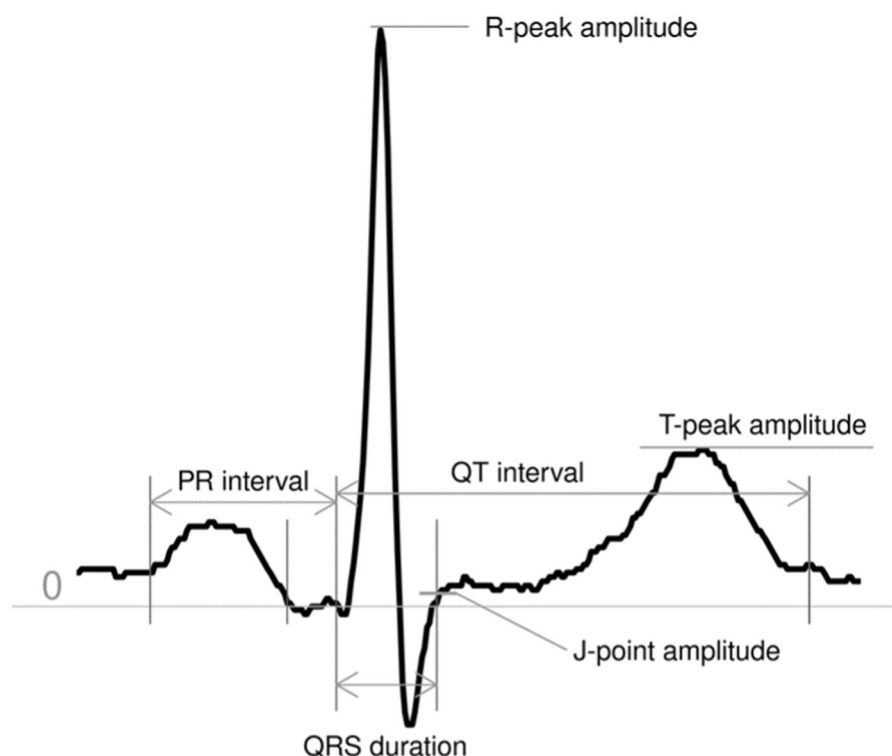


Figure 1. Illustration of an ECG, showing the amplitudes and intervals investigated in this study. The x-axis is the time axis, and the y-axis is the voltage axis.

Table 1. Overview of explanation methods evaluated in this study.

Method	Model requirements	Baseline	Noise sampling
Saliency ¹³		No	No
SmoothGrad ¹⁶		No	Yes
Input*Gradient ¹⁷		No	No
Deconvolution ¹⁴	ReLUActivations	No	No
Guided Backpropagation ¹⁵	ReLUActivations	No	No
DeepLIFT ¹⁷		Yes	No
Integrated Gradients ¹⁸		Yes	No
Guided Grad-Cam ²⁰	Convolutional Network	No	No
Gradient SHAP ¹⁹		Yes	Yes

dataset to expert evaluation. All heatmaps were presented twice (2) to estimate intraobserver variation, so all experts reviewed 216 heatmaps for each of 51 ECGs (1224 decisions between 9 heatmaps). The heatmaps were randomly shuffled and the experts were blinded for the ECG ID and explanation method, but not for the ECG feature of interest. All experts were familiar with heatmaps. The experts were tasked with selecting the explanation method which in the expert's opinion best explained the measurement of the ECG marker at hand. The voting guide provided to the ECG experts prior to the heatmap evaluation is included in [Appendix A.1](#).

Objective heatmap evaluation

For objective evaluation, we applied a perturbation procedure consisting of iteratively randomizing data at the location of highest heatmap importance in the full heatmaps and run a new prediction on the perturbed data.²¹ Since the model always is run on the full 12-lead ECG, it is not possible to do objective evaluation on the averaged ECGs. For each

iteration, the location to inject randomized data is selected in decreasing importance based on the initial generated heatmap. If the heatmap successfully highlights the areas that are important to the model's predictions, the prediction error will increase quickly within few iterations of perturbation, meaning that the most important regions of information have been perturbed. For a poorly created heatmap that misses important regions or highlights regions that are not relevant to the model's prediction, the prediction error will increase late or not at all. Since our prediction targets are defined with very high resolution (both in amplitude and time), we followed the example of reference²² and perturbed only a single data point per iteration.

To perturb the data point, we added a random value which was drawn from a normal distribution with a mean of 0 and standard deviation equal to 20% of the difference between the largest and the smallest value in the ECG data sample under study. By performing region perturbation 200 times, a little <3% of the data was randomized. The procedure was

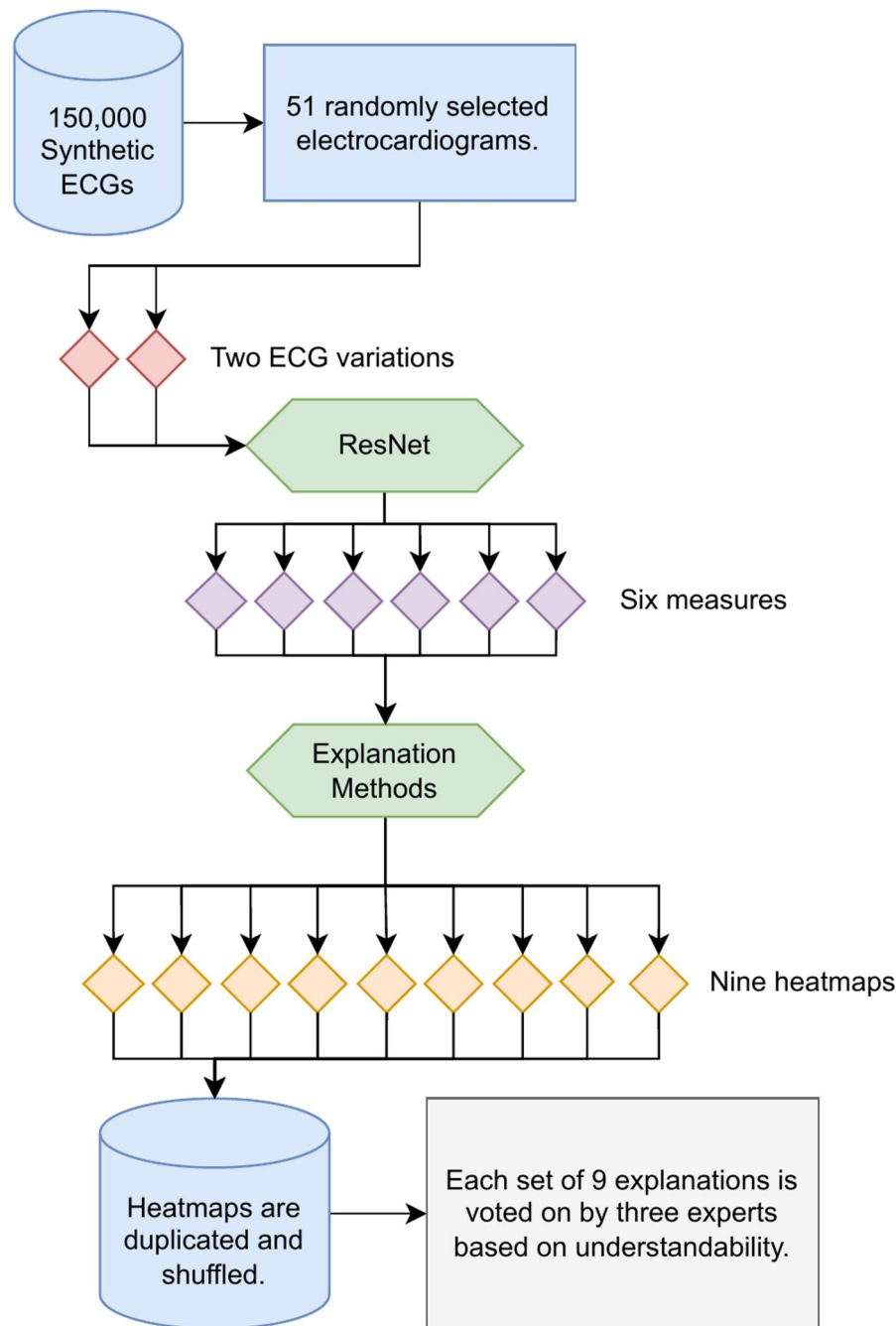


Figure 2. A flow diagram illustrating the study. It details the extraction of 51 ECG samples, analysis using a ResNet model to predict 6 measures, application of 9 explanation methods, and expert evaluation of shuffled heatmaps.

done for a set of 5000 randomly selected ECG samples, and the results were averaged across all samples. Prediction error curves over the 200 iterations of perturbation were plotted for all combinations of heatmaps and models. To have a baseline representing a method that is completely independent of the models' weights and performance, we also included prediction error curves for randomly created heatmaps. As a metric to quantify the results of the perturbation, the areas under the error curves up to 15 iterations of perturbation were computed. We chose 15 iterations to capture only the rise of the curve (not the plateau), and verified that the ranking of the different methods stayed the same for any choice

between 10 and 25 iterations. A large area under the curve indicates large effects of the perturbations and therefore accurate heatmaps.

Statistical calculations

The experts' votes were used to rank the explanation methods, both within each target prediction group and across all predictions. Inter- and intra-annotator agreement was calculated in terms of Cohen's κ .²³ Values in the intervals $[-1.0, 0.0]$, $(0.01, 0.20]$, $(0.21, 0.40]$, $(0.41, 0.60]$, $(0.61, 0.80]$, $(0.81, 0.99]$ refer to disagreement and slight, fair, moderate, substantial, and near perfect agreement, respectively, while 1

Table 2. Number of votes for different explanation methods for different observation types.

	PR	QRS	QT	R-peak	J-point	T-peak
Full heatmaps (12-lead)						
Deconvolution	0	0	0	0	0	0
DeepLIFT	33	0	2	0	117	0
Gradient SHAP	2	41	2	0	7	1
Guided backpropagation	33	19	52	0	1	24
Guided Grad-CAM	1	6	2	0	2	7
Input \times gradient	1	0	2	306	8	272
Integrated gradients	15	68	0	0	35	0
Saliency	105	42	65	0	34	0
SmoothGrad	116	130	181	0	102	2
Averaged heatmaps (one lead)						
Deconvolution	0	0	0	0	0	0
DeepLIFT	81	9	24	0	137	0
Gradient SHAP	7	37	3	0	26	0
Guided backpropagation	125	122	98	0	5	3
Guided Grad-CAM	1	34	0	0	0	0
Input \times gradient	3	1	8	306	14	302
Integrated gradients	22	23	5	0	61	1
Saliency	22	35	71	0	62	0
SmoothGrad	45	45	97	0	1	0

The upper part of the table shows the results for full heatmaps, while the lower part represents averaged heatmaps. The highest number of votes are highlighted in bold.

indicates perfect agreement.^{23,24} 95% confidence intervals for the κ values were calculated using bootstrapping.

Results

The present study revealed that the performance of attention map algorithms on medical data varied considerably. There was no single method that was superior for all ECG measurements, meaning that an individual explanation method had to be chosen for each ECG measure. The networks' predictive performances were similar to previously published results¹¹ on the 6 presented ECG measures with respect to MAE and RMSE. Appendix A.2 provides the performance metrics for the models developed in this study.

Expert heatmap evaluation

Table 2 displays the number of votes for the explanation methods and observation categories for complete and averaged heatmaps. Across all markers, DeepLIFT, Guided backpropagation, Input \times gradient, Saliency, and SmoothGrad received many votes whereas Deconvolution, Gradient SHAP, guided Grad-CAM, and Integrated Gradients received few to no votes. However, different explanation methods were preferred for various observation categories. See examples for averages heatmaps on Figures 3 and 4. All explanation methods for each ECG measure can be found in Appendix A.4.

For amplitude predictions, DeepLIFT (small amplitude, J-point_{V5} amplitude) and Input \times gradient (large amplitudes, R_{V5}- and T_{V5}-amplitude) were preferred for both full and averaged heatmaps (Figures 5 and 6). Notably, Input \times Gradients was unanimously selected for the tallest R_{V5}-amplitude. For the interval measures, SmoothGrad was preferred for full heatmaps, and Guided backpropagation was preferred for average heatmaps. It is noteworthy that most heatmaps occasionally were selected, except for Deconvolution that never was chosen by the experts.

The intra- and inter-annotator agreements were explored for the qualitative heatmap evaluation survey. Intra-annotator agreement was substantial, ranging between 0.6 and 0.9 (Appendix A.3). The inter-annotator agreement ranged from perfect for R_{V5}-amplitude heatmaps to poor for J-point_{V5} amplitude (Table 3). Generally, the inter-annotator agreement was higher for average heatmaps than for full heatmaps, except for the QT interval where agreement was similar.

Objective evaluation

Figure 7 shows the effect of adding noise to the ECG guided by heatmaps. The absolute increase in error was smaller for interval predictions (Figure 7A-C) than for amplitude predictions (Figure 7D-F). The areas under the curve Figure 7 were calculated up to iteration number 15 and listed in Table 4. DeepLIFT, Gradient SHAP, Saliency, and SmoothGrad performed consistently well with steep error curve slopes. Guided backpropagation and Input \times gradient did well except for the R_{V5}-amplitude. Integrated gradients were a stable heatmap. Guided GradCAM performed under par, and Deconvolution was at the level of random perturbation.

Discussion

The study demonstrates that commonly used explanation methods in ECG analysis ranges from highly effective to completely ineffective. In addition, there is no universally favored heatmap for ECG analysis, and a suitable heatmap must be carefully chosen for each ECG parameter.

If heatmaps fail to satisfy healthcare staff, they may refrain from utilizing the neural network models. This would hinder the implementation of technology that could otherwise facilitate expedited and precise examination of medical data. It is imperative that data scientists who are creating new explanation methods closely work with these experts.

The heatmaps were reproducible for each individual expert; however, there were significant differences in repeatability amongst different experts, ranging from modest (J-point amplitude) to nearly flawless (R-wave amplitude). In general, experts exhibited greater reproducibility when analyzing averaged heatmaps as opposed to full heatmaps encompassing all 12 leads. This is likely due to the dispersion of information over all leads, making it more challenging for the expert to assess and comprehend. In contrast, the averaged heatmap provides a clearer and more comprehensible representation. Simple heatmaps may be more favorable for the utilization of explainable AI.

For amplitude measurements, Table 2 and Figure 1 demonstrated that the Input \times Gradient method accurately identified the peak of the R-wave, which is where R-wave amplitude should be measured. The R-peak, the highest amplitude wave in the human ECG, is characterized by its sharpness and high frequency content. Consequently, it is the most easily identifiable wave for humans as well as computers, explaining the perfect agreement between the experts.

Additionally, the Input \times Gradient method provided comparable results for measuring the T-wave amplitude. Interestingly, most other approaches completely failed to detect the R-wave peak, whereas all methods except deconvolution successfully recognized the T-wave peak, as anticipated for measuring T-wave amplitude. The T-wave peak lacks the same level of distinctness as the R-wave, and its rounded

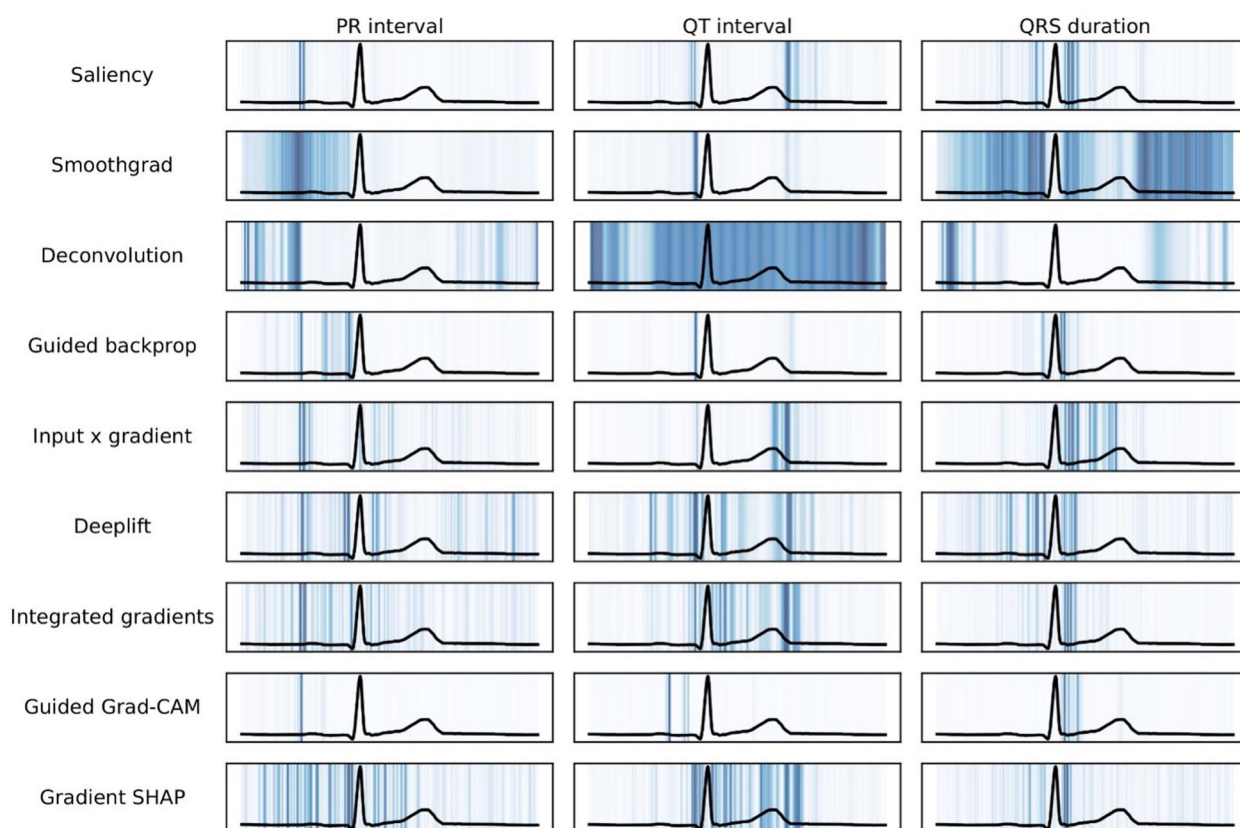


Figure 3. Visual comparison of explanation methods for interval predictions. Regions with a darker blue intensity contribute more significantly to the prediction.

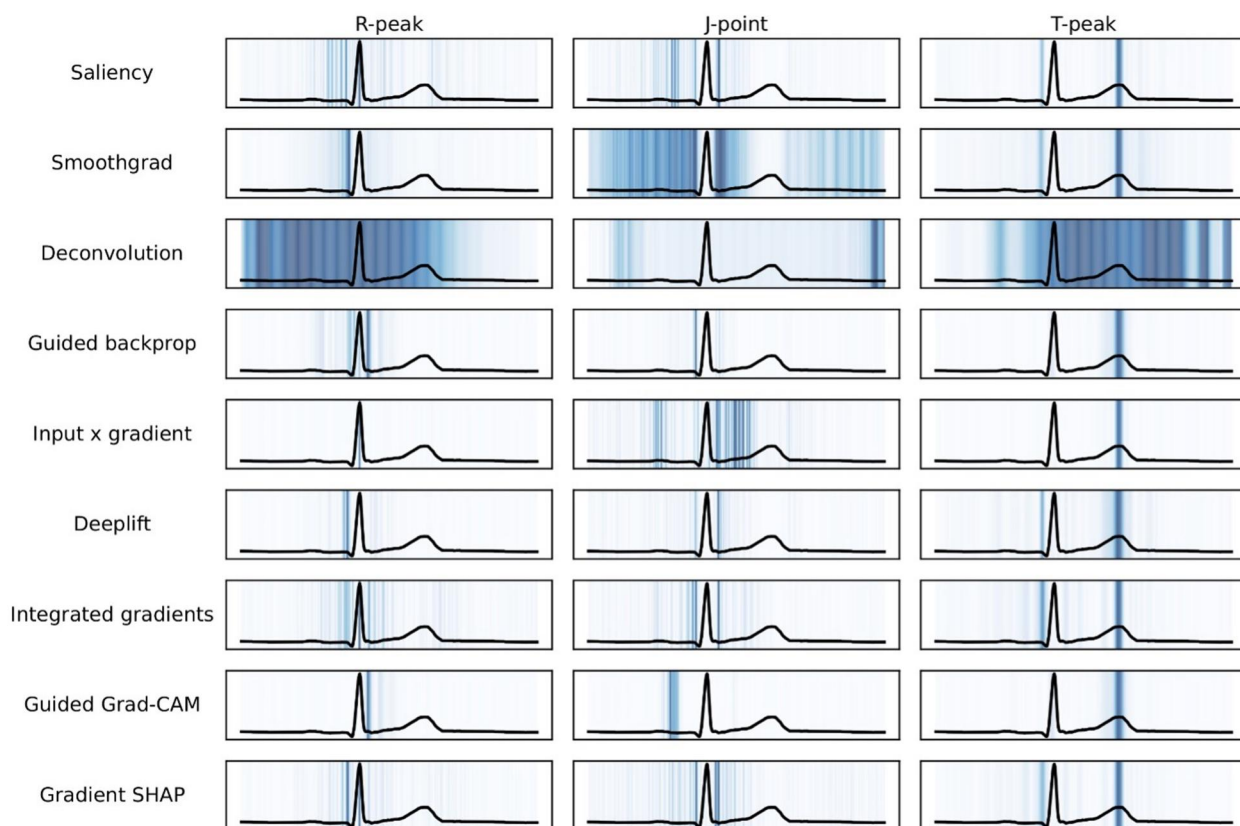


Figure 4. Visual comparison of explanation methods for amplitude predictions. Regions with a darker blue intensity contribute more significantly to the prediction.

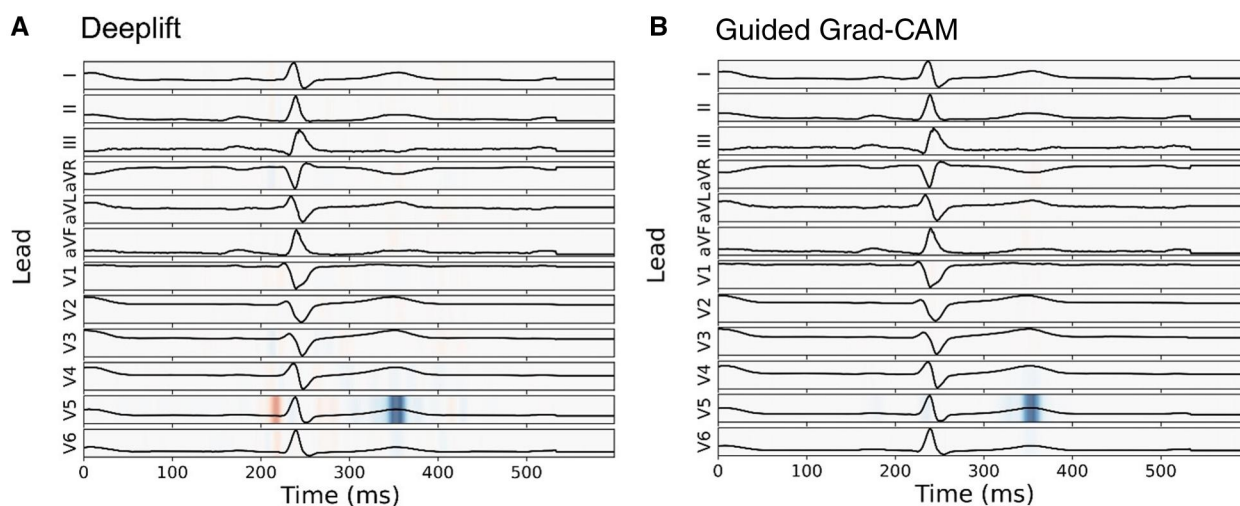


Figure 5. A 12-lead ECG overlaid with heatmaps for predicting T-wave amplitude. Panel (A) shows a heatmap generated by DeepLIFT, identified as the most effective method in objective heatmap evaluations. Panel (B) displays a heatmap from Guided Grad-CAM. In both panels, blue areas indicate regions that increase prediction values, while red areas indicate regions that decrease prediction values.

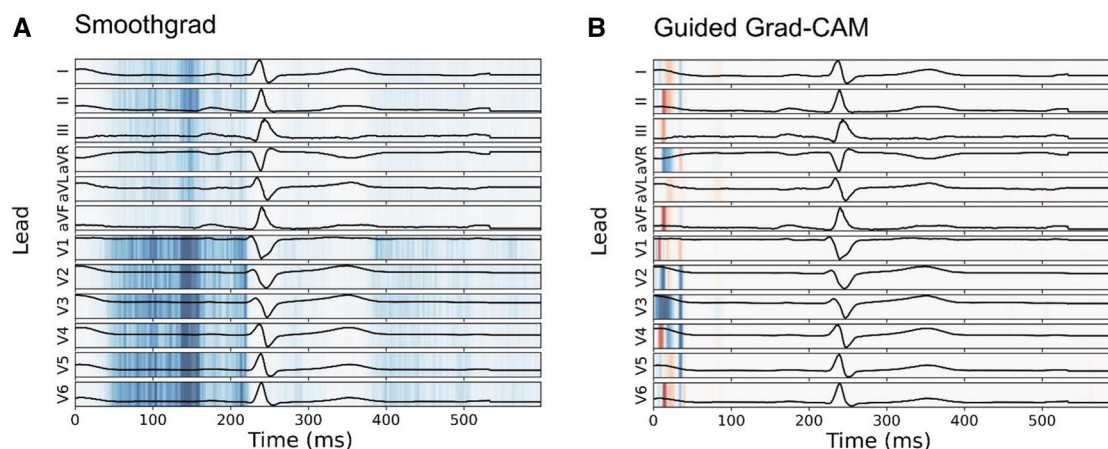


Figure 6. A 12-lead ECG overlaid with heatmaps for PR interval prediction. Panel (A) features a heatmap generated by SmoothGrad, rated by experts as the best method for PR interval analysis, versus Panel (B) which shows a heatmap from Guided Grad-CAM. In both panels, blue areas signify regions that increase prediction values, and red areas signify regions that decrease prediction values.

Table 3. Inter-annotator agreement (Cohen's κ) between the 3 annotators for the different ECG features, and for all ECG features in total.

Observation type	Full heatmaps	Averaged heatmaps
PR interval	0.15 (0.12, 0.18)	0.49 (0.39, 0.60)
QRS duration	0.10 (0.06, 0.14)	0.59 (0.49, 0.67)
QT interval	0.25 (0.20, 0.32)	0.23 (0.15, 0.32)
R-amplitude in V5	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
J-point elevation in V5	0.07 (0.03, 0.11)	0.25 (0.19, 0.32)
T-amplitude in V5	0.78 (0.70, 0.86)	0.97 (0.94, 0.99)
Total	0.39 (0.37, 0.41)	0.59 (0.56, 0.62)

The values in parentheses are 95% confidence intervals.

shape makes it more prone to noise. The T-wave may exhibit humps or notches,²⁵ resulting in several peaks that may possess comparable magnitude, hence confounding detection by both neural networks and humans.

The J-point is the amplitude at the termination of the QRS complex, just prior to the initiation of the ST segment. The Input \times Gradient heatmap showed poor performance in

determining the J-point_{V5} amplitude, where the experts favored DeepLIFT for both averaged and full heatmaps.

The experts favored SmoothGrad for generating full heatmaps and Guided Backpropagation for average heatmaps with interval measurements. Guided Backpropagation produced average heatmaps that effectively highlighted the well-defined start of the PR/QT/QRS interval, as determined by experts. In contrast, other methods either generated incomprehensible attention patterns or failed to accurately identify the beginning or end of the interval.

The preferred explanation method varies between models predicting different tasks, which is in line with earlier research.²⁶ Jin et al.²⁶ grading glioma in brain magnetic resonance imaging (MRI) scans and detecting lesions in knee MRI scans, suggest that the performance of explanation methods is task dependent. Even though an explanation method performs well for explaining one deep neural network model, this method can be less optimal when switching to another neural network architecture and/or medical task. These results, therefore, warrant exploration of several

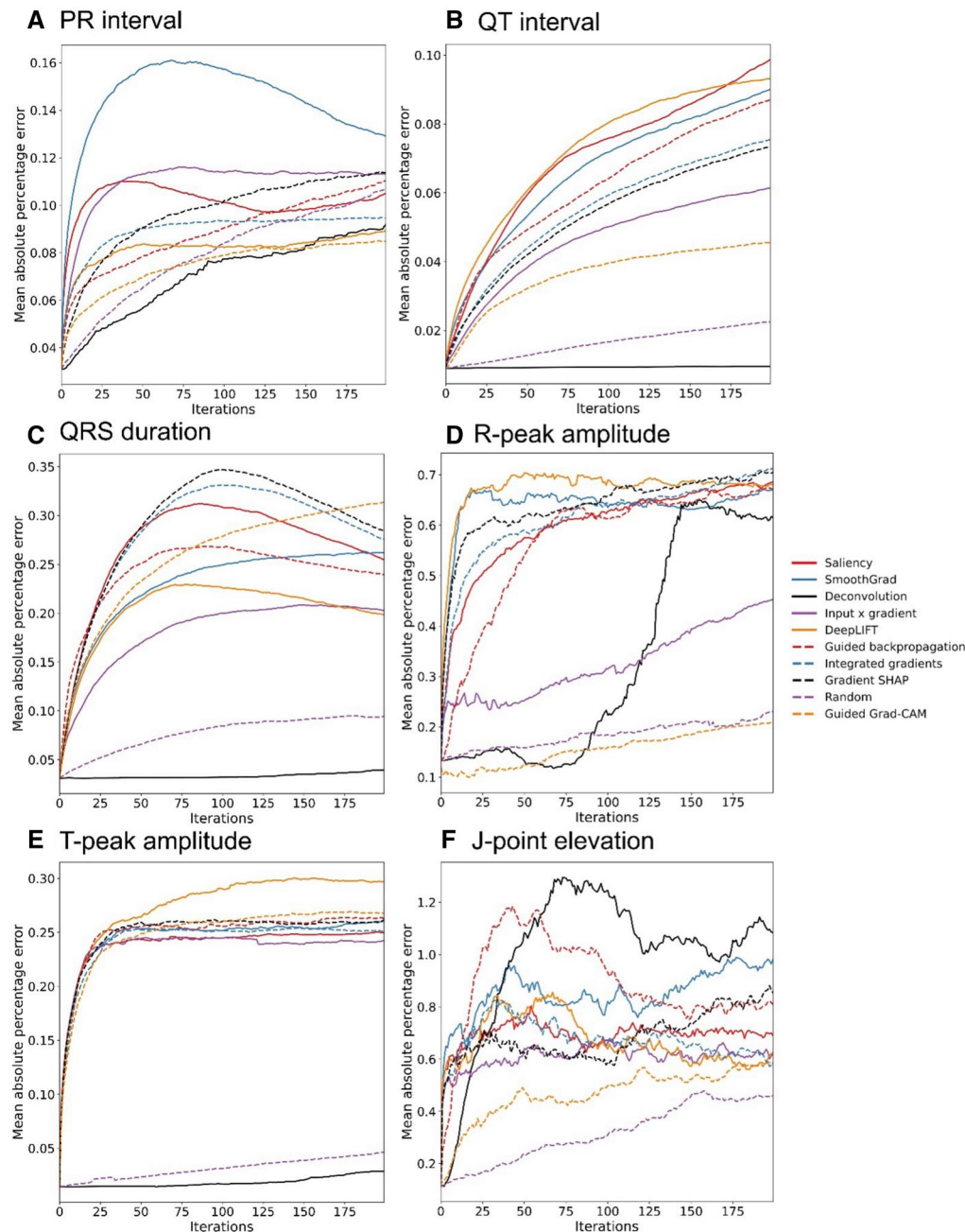


Figure 7. Comparison of explanation methods by region perturbation (A–F). Steeper curves indicate stronger effects of explanation-guided perturbations.

different methods when explaining ML models. The findings that multiple explanation methods are necessary for a comprehensive analysis are of particular importance when ML-models in combination with heatmaps are used to extract novel biological information. In that analysis, it is important to know what kind of information (interval, amplitude) a method can identify and how that information would be mapped onto the ECG. Our findings help make that mapping possible for future ECG-based studies.

The experts' heatmap evaluation determines if the areas of attention have biological significance for the ECG experts, while the objective evaluation determines if interfering with the very important areas according to the heatmaps with noise affects the prediction. Consequently, places of high

importance on the heatmap should be more impacted than areas of lower importance. Differences between objective measurements and experts' assessments may in part owe to differences in how experts would analyze the ECG and how ML models analyze the ECG.

The objective heatmap evaluation indicated that SmoothGrad was the preferable approach for evaluating the PR interval. It was the second-best method for evaluating the QT interval and acceptable for evaluating the QRS interval, close to the evaluation of the experts. There is a noticeable difference in the evaluations of the amplitude heatmaps. The objective assessment favored SmoothGrad for both J-point and T-wave amplitude, whereas Deep Lift was selected for R amplitude. The experts exhibit a preference for SmoothGrad when

Table 4. Area under the mean absolute percentage error curve for the first 15 iterations.

Method	PR	QRS	QT	R _{V5} - amplitude	J-point	T _{V5} - amplitude
Deconvolution	0.53	0.47	0.14	2.06	3.58	0.22
Deep Lift	0.94	1.44	0.39	8.04	8.47	2.49
Gradient SHAP	0.79	1.62	0.28	6.90	8.25	2.45
Guided backpropagation	0.86	1.91	0.33	3.34	7.18	2.49
Guided Grad-CAM	0.74	1.59	0.21	1.61	3.28	2.22
Input×gradient	1.07	1.21	0.24	3.60	7.54	2.50
Integrated gradients	0.94	1.64	0.27	5.71	8.32	2.34
Random	0.57	0.58	0.14	2.09	2.01	0.24
Saliency	1.27	1.65	0.32	5.32	8.87	2.49
SmoothGrad	1.50	1.54	0.35	7.48	9.81	2.50

The highest value for each ECG parameter is highlighted in bold. Marker-specific rank indicated in parenthesis as best (1), medium (2), poor (3), based on the range from winner to random control.

it comes to the J-point amplitude, while distinctly favoring the Input × gradient for R- and T-amplitude.

The discrepancy in the Input × gradient between experts and objective evaluation is unexpected, considering that Input × gradient effectively identified the R-peak in the ECGs. The objective evaluation evaluated the Input × gradient, sixth out of 9, and it is expected that disturbing the narrow R-wave will result in a significant rise in model prediction error. Given that the amplitude of the R-wave in lead V5 is the highest, the impact of adding a random value to the ECG will be relatively little compared to adding the same value to a low amplitude section of the ECG. In the context of explaining the R_{V5}-amplitude model, the model's sensitivity to perturbations will be reduced if the heatmap identifies regions with strong signals, as demonstrated by the Input × gradient approach. Conversely, when an explanation emphasizes sections of the ECG that do not correspond to the peak of the R-wave, the impact of disturbance will be greater and more likely to result in a more pronounced rise in the error curve. Consequently, the objective evaluation approach shows a preference for explanation methods that do not emphasize the R-peak. This likely accounts for the unexpected findings on the R-peak in Table 4.

The interval forecasts exhibited a more gradual rise in error as the disturbance increased, in contrast to the amplitudes. This outcome is anticipated, as the relevant information within intervals is distributed across numerous ECG channels, resulting in redundancy. Put simply, interval models include multiple leads to make predictions and changing a single data point is unable to eliminate all the necessary information for accurate predictions.

Our results on ECG time series confirm findings from heatmaps for chest X-ray analysis. They concluded that none of the examined explanation methods were able to generate consistent explanations that highlighted the regions of interest.²⁷ Another study performed on fundus images for diabetic retinopathy detection systematically evaluated 10 different heatmaps methods.²⁸ The results showed that the best explanation method differed between different neural networks. Similar to the conclusion of the X-ray study,⁵ that heatmaps were unable to highlight clinically relevant regions in the fundus images.²⁸ Finally, a systematic evaluation of 16 explanation methods for analyzing MRI scans of the brain

and knee, respectively, concluded that the explanation methods were insufficient for clinical use.²⁶

It is possible that the unusual heatmap focus is a result of the neural network detecting important features that the ECG expert is unaware of. Nevertheless, it is peculiar that the various explanation methods exhibited significant variations when applied to the identical ECG approach. Measuring an amplitude involves determining the highest point on a curve. This may appear to be a simple issue, but it is surprising that many heatmap approaches did not prioritize accurately identifying the maximum point.

Limitations

The scope of our investigation was limited to ECG analysis, and it remains uncertain whether the heatmap evaluation outcomes would be consistent for different datasets and models. Due to the substantial work involved, the analysis was limited to a small number of annotators and ECGs. Furthermore, the ECGs that were used to conduct this study were fully synthetic. Despite these representing the distribution and characteristics of actual ECGs, this can still be recognized as a limitation of the study.

Recommendations

When using heatmaps to explain deep neural networks on ECG analysis, we recommend:

- Several explanation methods should be tested and compared to find the most appropriate method combined with qualitative evaluation by medical experts.
- Use heatmaps based on DeepLIFT, Guided backpropagation, Input × gradient, Saliency, or SmoothGrad methods.
- Do not use the Deconvolution heatmap.
- Use a collection of representative observations from the dataset.

Conclusion

The study was unable to establish a singular explanation method that could account for all models of ECG measurements. We recommend investigating several explanation methods in collaboration with domain experts.

Author contributions

Andrea Marheim Storås, Steffen Mæland, Jonas L. Isaksen, Steven Alexander Hicks, Vajira Thambawita, Michael Alexander Riegler, and Jørgen K. Kanthers conceived and designed the study. Andrea Marheim Storås, Steffen Mæland, and Steven Alexander Hicks developed the software used to perform the study. Jonas L. Isaksen, Claus Graff, and Jørgen K. Kanthers participated in the study. Andrea Marheim Storås, Michael Alexander Riegler, Steven Alexander Hicks, Jørgen K. Kanthers, Jonas L. Isaksen, Vajira Thambawita, and Hugo Lewi Hammer were involved in analyzing the study data. All authors participated in reviewing and revising the manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interest

The authors declare that they have no competing interests.

Data availability

The ECGs are fully annotated and publicly available at <https://osf.io/6hved/>.

References

1. Attia ZI, Harmon DM, Behr ER, et al. Application of artificial intelligence to the electrocardiogram. *Eur Heart J*. 2021;42:4717-4730.
2. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput Appl*. 2020;32:18069-18083.
3. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. *Adv Neural Inf Process Syst*. 27:9.
4. Thambawita V, Isaksen JL, Hicks SA, et al. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Sci Rep*. 2021;11:21896.
5. Bergholdt HK, Bathum L, Kvetny J, et al. Study design, participation and characteristics of the Danish General Suburban Population Study. *Dan Med J*. 2013;60:A4693.
6. Henriksen LF, Petri AS, Hasselbalch HC, et al. Increased iron stores prolong the QT interval—a general population study including 20 261 individuals and meta-analysis of thalassaemia major. *Br J Haematol*. 2016;174:776-785.
7. Ghouse J, Have CT, Weeke P, et al. Rare genetic variants previously associated with congenital forms of long QT syndrome have little or no effect on the QT interval. *Eur Heart J*. 2015;36:2523-2529.
8. GE Healthcare. *Marquette™ 12SL™ ECG Analysis Program Physician's Guide 2056246-002 Revision C*. 2015. Accessed October 28, 2024. <https://landing1.gehealthcare.com/rs/005-SHS-767/images/45351-MUSE-17Nov2022-6-1-Quick-Reference-Guide-LP-Diagnostic-Cardiology.pdf>
9. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE; 2016:770-8. <https://doi.org/10.1109/CVPR.2016.90>
10. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc.; 2019:Article 721, 8026-8037.
11. Hicks SA, Isaksen JL, Thambawita V, et al. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Sci Rep*. 2021;11:10949.
12. Kokhlikyan N, Miglani V, Martin M, et al. Captum: a unified and generic model interpretability library for PyTorch. arXiv:2009.07896. 2020. <https://doi.org/10.48550/arXiv.2009.07896>
13. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034. 2014. <https://arxiv.org/abs/1312.6034>
14. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, et al., eds. *Computer Vision—ECCV 2014*. Cham: Springer International Publishing 2014:818-33. https://doi.org/10.1007/978-3-319-10590-1_53
15. Springenberg JT, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net. arXiv:1412.6806. 2015. <https://doi.org/10.48550/arXiv.1412.6806>
16. Smilkov D, Thorat N, Kim B, et al. SmoothGrad: removing noise by adding noise. arXiv:1706.03825. 2017. <https://doi.org/10.48550/arXiv.1706.03825>
17. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *Proceedings of the 34th International Conference on Machine Learning*. Sydney, NSW, Australia: JMLR.org; 2017:3145-53.
18. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning*. Sydney, NSW, Australia: JMLR.org; 2017:3319-28.
19. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.; 2017:4768-77.
20. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128:336-359.
21. Samek W, Binder A, Montavon G, et al. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Netw Learn Syst*. 2017;28:2660-2673.
22. Ancona M, Ceolini E, Öztireli C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks. 2018. <https://doi.org/10.48550/arXiv.1711.06104>
23. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
25. Graff C, Matz J, Christensen EB, et al. Quantitative analysis of T-wave morphology increases confidence in drug-induced cardiac repolarization abnormalities: evidence from the investigational IKr inhibitor Lu 35-138. *J Clin Pharmacol*. 2009;49:1331-1342.
26. Jin W, Li X, Fatehi M, et al. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Med Image Anal*. 2023;84:102684.
27. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol Artif Intell*. 2021;3:e200267.
28. Van Craenendonck T, Elen B, Gerrits N, et al. Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection. *Transl Vis Sci Technol*. 2020;9:64.