Theresa Bender, Philip Hempel, and Nicolai Spicher*

Concordance analysis between deep learning model predictions and electrocardiography thresholds from cardiology guidelines

https://doi.org/10.1515/cdbme-2024-2018

Abstract: Deep learning models for the classification of electrocardiograms (ECGs) are able to learn disease-specific patterns, but they are rarely implemented in medical practice due to their "black box" nature. Post-hoc explainable artificial intelligence (XAI) methods compute regions of interest (ROI) which are of importance for a model's decision making. However, it needs to be further analyzed whether a model focuses on the morphological or rhythmical information within the ROIs. We evaluate a pre-trained ResNet for sinus bradycardia (SB) and sinus tachycardia (ST) classification on the PTB-XL dataset using the XAI method Integrated Gradients. We compare the confidence of the model predictions to ECG features used by clinicians using correlation analysis. Correlation is highest for RR intervals (SB: 0.44) and atrial as well as ventricular heart rates (ST: 0.51), with the majority exceeding clinical thresholds for both disorders, indicating that the model learned rhythmical features. Except for QT intervals in ST classification, morphological features such as duration and amplitudes of P-/T-waves do not show any correlation.

Keywords: Deep Learning, Electrocardiogram, Explainable Artificial Intelligence, Integrated Gradients

1 Introduction

Artificial neural networks take raw electrocardiograms (ECGs) as input and output probabilities for the presence of the diseases they were trained on. Due to the increasing availability of large datasets, these networks are able to learn disease-specific patterns based on millions of ECGs and reach high sensitivity and specificity [1]. Despite their broad application in other fields [2, 3], these "black boxes" are hardly implemented in medical practice, since they do not provide insights in their decision making. Explainable artificial intelligence (XAI) methods address this shortcoming by computing regions of interest (ROI) pointing to the most relevant parts of an input signal for the model's decision.

Recently, we proposed an open-source framework based on the post-hoc XAI method Integrated Gradients (IG) [4] which we customized to the analysis of networks for ECG classification [5] and age prediction [6]. Results revealed several insights; for example that the ROI of a model for diagnosis of atrial fibrillation (AF) was centered on the P-wave and the model learned to use its existence for ruling out AF. While the ROI can be shown to a cardiologist as a visual landmark and might increase trust in the network, the informative value is rather limited and not comparable to evidence-based gold standard, since features recognized by these models may not match those mentioned in clinical guidelines [7]. Instead, an analysis of the network's decision w.r.t. standard ECG parameters based on exact heartbeat interval features [8, 9], such as the RR interval, would have higher informative value, especially being able to distinguish between morphological and rhythmical features recognized in each ROI.

Hence, in this work we extend the open-source XAI framework proposed in [5] by integrating evidence-based ECG features used by cardiologists and analyze their correlation with a ResNet's decisions.

2 Medical Background

We focus on two rhythmical ECG abnormalities, sinus bradycardia (SB) and sinus tachycardia (ST), which are solely diagnosed based on evaluation of rest ECG without further symptoms or blood results necessary. In a normal heart cycle, the sinus node initiates the contraction of the heart, starting with the atria. This is shown in an ECG as P-wave, followed by the QRS-complex representing the contraction of the ventricles, and a T-wave indicating ventricular depolarization. A normal frequency of these heart cycles is $50 - 100$ bpm at rest, with a QT interval of $350 - 550$ ms, a P duration of $50 - 100$ ms and QRS duration of $60 - 100$ ms [11]. Moreover, the normal amplitude measured in lead II should not exceed $2/3$ of the R-peak for T-waves [11].

Theresa Bender, Philip Hempel, Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany
***Corresponding author: Nicolai Spicher,** Department of Medical Informatics, University Medical Center Göttingen, Robert-Koch-Str. 40, Göttingen, Germany, e-mail: nicolai.spicher@med.uni-goettingen.de

Tab. 1: Features from PTB-XL+ [10] sorted by number of recordings exceeding the clinical threshold. Percentages are given with regard to SB and ST recordings classified with high confidence (> 0.5), respectively. Pearson correlation coefficient (PCC) is calculated for each feature compared to model confidence. bpm: beats-per-minute, ms: milliseconds.

Sinus bradycardia			**PCC**	**Sinus tachycardia**			**PCC**
RR_Mean_Global	> $1,200$ ms	99.50 %	0.44	HR_Ventr_Global	> 100 bpm	99.75 %	0.51
HR_Ventr_Global	< 50 bpm	95.99 %	-0.28	RR_Mean_Global	< 600 ms	99.75 %	-0.40
HR_Atrial_Global	< 50 bpm	94.49 %	-0.23	HR_Atrial_Global	> 100 bpm	96.71 %	0.26
P_Dur_II	> 100 ms	83.21 %	0.05	QT_Int_Global	< 350 ms	70.38 %	-0.34
QRS_Dur_Global	> 100 ms	41.35 %	0.03	T_Amp_II	> $2/3*$R_Amp_II	12.15 %	-0.06
T_Amp_II	> $2/3*$R_Amp_II	9.77 %	0.04	P_Dur_II	< 50 ms	8.35 %	-0.03
QT_Int_Global	> 550 ms	0.50 %	0.27	QRS_Dur_Global	> 120 ms	6.84 %	-0.07

An SB means the heart rate is regular and below 50 bpm, while an ST means its regular and above 100 bpm. Both can be detected manually via mean PP-intervals, i.e. the distance between two heartbeats measured from the initial sinus node activity [11], which usually corresponds to the mean RR-interval.

3 Methods

We used PTB-XL [12] containing $21,414$ 12-lead ECGs of patients older than 16 years with annotation for the associated diseases as dataset. We exclude 363 patients under 16 since the network applied was trained on adults only. The mean age of the patients is 59.74 years (± 16.51) with 48% woman. Within PTB-XL there are 506 and 685 recordings annotated as SB and ST, respectively.

We analyze a state-of-the-art pre-trained ResNet [13] trained on more than two million ECGs to predict the presence of both SB and ST. First, we predict both disorders and calculate the F1-score to measure model performance. To gain insight into the decision process, we then make use of the XAI framework proposed in [5] which is build on Python and the iNNvestigate library for the computation of IG relevances.

IG attribute the prediction of a model f on unseen data to its input features x, using a baseline input $\tilde{x}$ for attribution calculation. The IG are defined as the path integral of the gradients along the straight-line path from $\tilde{x}$ and x, defined as $\tilde{x} + \alpha(x - \tilde{x})$ for $\alpha \in [0, 1]$. The integrated gradient for the i-th input dimension is then defined as

$$\mathrm{IG}_i(x) := (x_i - \tilde{x_i}) \cdot \int_0^1 \frac{\partial f(\tilde{x} + \alpha(x - \tilde{x}))}{\partial x_i} d\alpha, \qquad (1)$$

where $\frac{\partial f(x)}{\partial x_i}$ is the gradient of output $f(x)$ along the i-th dimension [14].

We analyze whether features located at ROIs identified from IG relevances correlate with the confidence of the ResNet. For each feature, we calculate for all recordings classified with high confidence (> 50 %) a) the percentage of recordings where the feature is abnormal with thresholds from Section 2, as well as b) the Pearson correlation coefficient (PCC) with the ResNet confidence.

We include the rhythmical features RR interval (*RR_Mean_Global*), ventricular heart rate (*HR_Ventr_Global*), and atrial heart rate (*HR_Atrial_Global*), as well as the morphological features QT interval (*QT_Int_Global*) and QRS duration (*QRS_Dur_Global*), averaged over all leads. Additionally, we extract the morphological features R amplitude (*R_Amp_II*), T amplitude (*T_Amp_II*), and P duration (*P_Dur_II*), averaged over lead II, as there are no global equivalents. Features were selected based on cardiology literature (cf. Section 2) and extracted with a commercial ECG delineation algorithm, GE Healthcare's Marquette™ 12SL™, from the recently published PTB-XL+ dataset [10].

4 Results

Analyzing the PTB-XL dataset with the Ribeiro model produces F1-scores of 0.67 for SB and 0.83 for ST, respectively.

Figure 1 displays the results of XAI analysis for a true positive SB result. As can be seen in this example, the ROI is centered on the QRS complex as well as on the T-wave where moderate positive values are accumulated. In some leads, such as V1 and II, P-waves appear in the ROI as well. In Figure 2 similar ROIs can be seen for a true positive ST recording, although ROIs centered on P- and T-waves are scarce. In both classifications, most ROIs can be found in I, II, aVR and V1.

In Table 1 features related to these ROIs are compared to clinical thresholds for all recordings classified with high confidence (> 0.5). For both SB and ST, heart rate as well as RR mean are abnormal in more than 94% of cases, respectively. Furthermore, a long P-wave duration in SB (83.21 %) and a short QT interval in ST (70.38 %) can be observed. On the

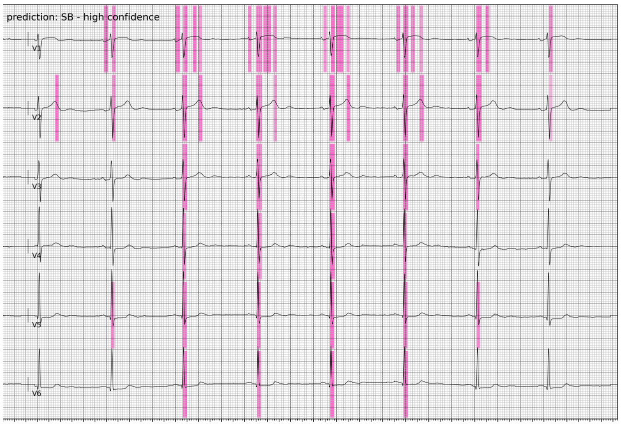

Fig. 1: XAI results for recording 4334 from PTB-XL [12], classified correctly with SB (0.69). Positive IG attributions (pink) with a threshold of 0.4 can be seen on QRS-complexes, sometimes P- and T-waves, mainly in V1.

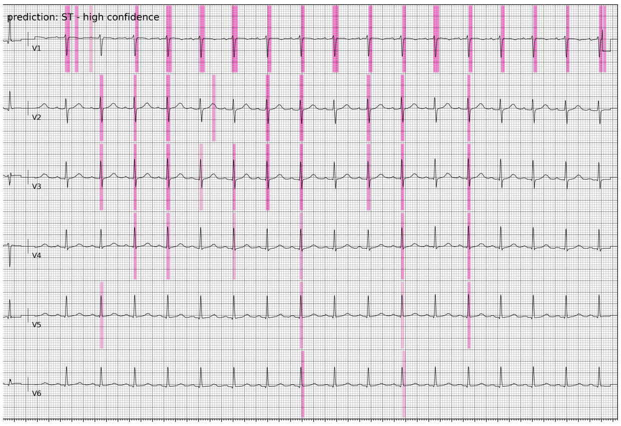

Fig. 2: XAI results for recording 16903 from PTB-XL [12], classified correctly with ST (0.69). Positive IG attributions (pink) with a threshold of 0.4 can be seen on QRS-complexes, mainly in V1.

contrary, QRS duration shows no correlation ($|\text{PCC}| < 0.07$) in both cases.

Figure 3 shows the agreement between the mean RR interval and network confidence. ECGs are represented by gray (healthy controls), blue (SB classification with low confidence (model output ≤ 0.5)), or red (SB classification with high confidence (> 0.5)) dots. A clear trend w.r.t. model confidence can be observed showing that the higher the confidence of the network, the higher the RR interval with a PCC of 0.44 showing moderate correlation. Predictions with high confidence were in 99.5% above the clinical threshold of $1,200$ ms.

Similarly, Figure 4 shows the agreement between the mean ventricular heart rate and network confidence. A high PCC of 0.51 can be observed. Predictions with high confidence were in 99.75% above the clinical threshold of 100 bpm.

5 Discussion

In concordance with clinical guidelines for both SB and ST [8, 9], the ResNet analyzed in this work seems to base it's decisions on rhythmical features only. Correlation is highest for ventricular and atrial heart rates as well as the RR interval, which is similar to the ventricular heart rate. Although the QT interval as a morphological feature correlates to the models confidence as well, this could be explained due to this interval changing in response to the heart rate.

However, the lower concordance of network decisions and atrial heart rates compared to other rhythmical features suggests that the ResNet concentrated on RR intervals (i.e. ventricular heart rate), rather than PP intervals mentioned in cardiology guidelines. This is underlined by ROIs which are concentrated mostly on QRS complexes for both, SB and ST. However, this might probably be an effect of the PTB-XL+ ground truth which was determined using an ECG delineation algorithm, which are known to perform better in detecting R-peak features than P-wave features [15]. This was also underlined by the feature *HR_Atrial_Global* containing 25 ”N/A” values for both disorders.

In summary, the analyzed ResNet presumably learned rhythmical features for SB and ST, however, not the same features as defined by cardiology guidelines. In future work we will analyze these concordances further by examining more features and including new methods for feature importance such as proposed in [16]. Additionally, due to the low correlation of model confidence and QRS duration, it could be assumed that the model is able to distinguish between atrial and ventricular tachycardia, with the latter requiring broad QRS complexes for diagnosis, which will be investigated by the inclusion of further labels.

Author Statement

Research funding: This research was funded by the Lower Saxony “Vorab” of the Volkswagen Foundation and the Ministry for Science and Culture of Lower Saxony (grant no. 76211-12-1/21). Conflict of interest: Authors state no conflict of interest. Informed consent: Not applicable. Ethical approval: This work only makes use of public data and does not contain any additional information involving human participants obtained by the authors.

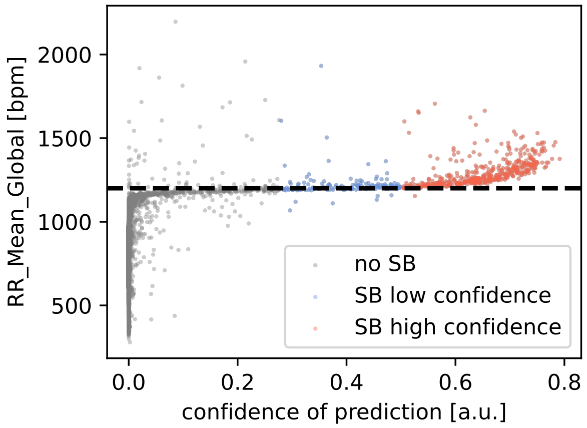

Fig. 3: Each dot in the scatter plots represents a single ECG classified by the network, with gray dots representing healthy controls and colored dots SB patients. The y-axes show the global mean RR interval from PTB-XL+ [10] in relation to the confidence of the ResNet for SB. The dashed line indicates the evidence-based threshold for bradycardic heart rates of < 50 bpm or $1,200$ ms.

References

[1] Zhang Y, Li J, Wei S, Zhou F, Li D. Heartbeats classification using hybrid time-frequency analysis and transfer learning based on ResNet. IEEE Journal of Biomedical and Health Informatics. 2021;25(11):4175-84.

[2] Haque MF, Lim HY, Kang DS. Object detection based on VGG with ResNet network. In: 2019 International Conference on Electronics, Information, and Communication (ICEIC). IEEE; 2019. p. 1-3.

[3] Kiliç Ş, Askerzade I, Kaya Y. Using ResNet transfer deep learning methods in person identification according to physical actions. IEEE Access. 2020;8:220364-73.

[4] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17. JMLR.org; 2017. p. 3319-28.

[5] Bender T, Beinecke JM, Krefting D, Muller C, Dathe H, Seidler T, et al. Analysis of a Deep Learning Model for 12-Lead ECG Classification Reveals Learned Features Similar to Diagnostic Criteria. IEEE journal of biomedical and health informatics. 2023.

[6] Hempel P, Bender T, Gandhi K, Spicher N. Towards explaining deep neural network-based heart age estimation. In: 2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology. Malta: IEEE; 2023. p. 41-2.

[7] Poon AI, Sung JJ. Opening the black box of AI-Medicine. Journal of Gastroenterology and Hepatology. 2021;36(3):581-4.

[8] Kusumoto FM, Schoenfeld MH, Barrett C, Edgerton JR, Ellenbogen KA, Gold MR, et al. 2018 ACC/AHA/HRS Guideline on the Evaluation and Management of Patients With Bradycardia and Cardiac Conduction Delay: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines, and the Heart Rhythm Society. Circulation. 2019;140(8):e333-81.

[9] Blomström-Lundqvist C, Scheinman MM, Aliot EM, Alpert JS, Calkins H, Camm AJ, et al. ACC/AHA/ESC guidelines for the management of patients with supraventricular arrhythmias—executive summary A Report of the American College of Cardiology/American HeartAssociation Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines(Writing Committee to Develop Guidelines for the Management of Patients With Supraventricular Arrhythmias)Developed in collaboration with NASPE–Heart Rhythm Society. European Heart Journal. 2003;24(20):1857-97.

[10] Strodthoff N, Mehari T, Nagel C, Aston PJ, Sundar A, Graff C, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset. Scientific Data. 2023;10(1):279.

[11] Trappe HJ, Schuster HP. EKG-Kurs für Isabel. Stuttgart: Georg Thieme Verlag; 2017.

[12] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a large publicly available electrocardiography dataset. Scientific data. 2020;7(1):154.

[13] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature communications. 2020;11(1):1760.

[14] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. vol. 70 of Proceedings of Machine Learning Research. PMLR; 2017. p. 3319-28. Available from: https://proceedings.mlr.press/v70/sundararajan17a.html.

[15] Spicher N, Kukuk M. Delineation of electrocardiograms using multiscale parameter estimation. IEEE journal of biomedical and health informatics. 2020;24(8):2216-29.

[16] Mehari T, Sundar A, Bosnjakovic A, Harris P, Williams SE, Loewe A, et al.. ECG Feature Importance Rankings: Cardiologists vs. Algorithms; 2023.

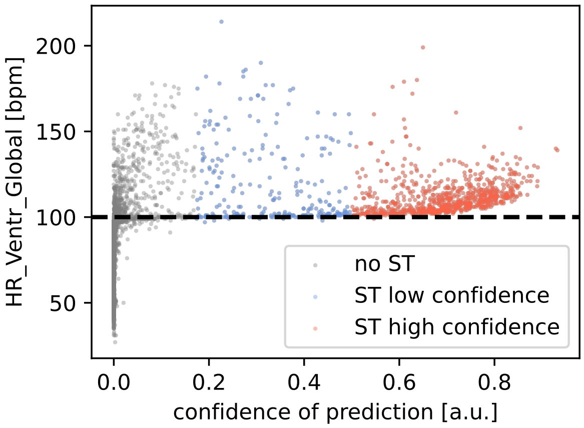

Fig. 4: Each dot in the scatter plots represents a single ECG classified by the network, with gray dots representing healthy controls and colored dots ST patients. The y-axes show the global ventricular heart rate from PTB-XL+ [10] in relation to the confidence of the ResNet for ST. The dashed line indicates the evidence-based threshold for tachycardic heart rates of > 100 bpm.