Lennart Graf*, Philip Hempel, Tabea F. A. Steinbrinker, Stephan von Haehling, Dagmar Krefting, and Nicolai Spicher

Influence of Pacemaker Artifacts on Deep Neural Networks for ECG Classification

https://doi.org/10.1515/cdbme-2025-0181

Abstract: Patients with pacemakers represent a high-risk group for cardiovascular events. While deep learning has recently shown promising results in detecting many abnormalities in electrocardiograms (ECGs), little attention has been paid to its performance in the presence of pacemaker activity. In this study, we evaluate the impact of pacemakers on the performance of a state-of-the-art deep neural network (DNN). Using the MIMIC-IV ECG dataset for validation, we compared model performance between ECGs from patients with and without implanted pacemakers. We observed a notable decrease in performance in pacemaker patients compared to non-paced patients (area under curve (AUC) 0.850 vs 0.770, sensitivity 0.630 vs. 0.423). To understand this discrepancy, we applied an explainable artificial intelligence (XAI) method to analyze the relevance of the model's predictions. Lead V1 was identified as the most relevant lead for the prediction of the model, even in the presence of a pacemaker. In addition, the false negative (fn) predictions of the model for pacemaker ECGs were most influenced by the P wave segment in lead V1. Our findings highlight the need for careful adjustment and training of DNNs in healthcare to achieve fair models that generalize well across diverse patient populations.

Keywords: Electrocardiogram, atrial fibrillation, deep learning, explainable artificial intelligence, integrated gradients.

***Corresponding author: Lennart Graf,** Department of Medical Informatics, University Medical Center Göttingen and Campus Institute Data Science, Robert-Koch-Str. 40, Göttingen, Germany, e-mail: frederiklennart.graf@med.uni-goettingen.de
Philip Hempel, Tabea F. A. Steinbrinker, Department of Medical Informatics, University Medical Center Göttingen and Campus Institute Data Science, Göttingen, Germany
Stephan von Haehling, Department of Cardiology and Pneumology, University Medical Center Göttingen and German Centre for Cardiovascular Research (DZHK), Partner Site Lower Saxony, Göttingen, Germany
Dagmar Krefting, Department of Medical Informatics, University Medical Center Göttingen and Campus Institute Data Science and German Centre for Cardiovascular Research (DZHK), Partner Site Lower Saxony, Göttingen, Germany
Nicolai Spicher, Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany and Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark

1 Introduction

A pacemaker is an implanted device generating electrical impulses to the heart, which provoke the heart to beat. Pacemakers are indicated in various groups at cardiovascular risk, e.g. dysfunctions of the sinus node or high-grade atrioventricular blocks resulting in bradycardia. Hence, patients with a pacemaker are a risk group, requiring careful monitoring of cardiovascular status. The electrocardiogram (ECG) is the standard tool for cardiac assessment, offering a non-invasive and multidimensional perspective on the heart's physiological activity. Atrial fibrillation (AF) is a serious disease as it is associated with an increased risk for other cardiovascular diseases such as heart failure or stroke. In particular, paroxysmal AF often occurs asymptomatically and intermittently, making it especially difficult to detect in single screening events.

While AF detection from ECG is well studied in the overall population [1], its detection is challenging in pacemaker patients, where pacing artefacts can mask irregular rhythms, but even with a pacemaker the risk of stroke remains elevated [2]. Studies have shown that the annual incidence of AF after pacemaker implantation is at least 5% and lifetime risk is estimated to be 30-50%. This is primarily due to advanced age and a higher prevalence of cardiovascular comorbidities in this patient population [3–5].

Moreover, ECGs from paced patients are challenging to interpret. Modern pacemakers deliver short, low-voltage electrical stimuli. Detection of these pacing artifacts can be challenging, especially at low sampling rates or with excessive filtering. Hence, algorithms might fail to detect pacemaker activity and interpret the resulting waveforms as pathological findings; e.g. pacing-induced QRS complexes are misinterpreted as left bundle branch block or myocardial infarction [6].

Recently, DNNs have been suggested for the automatic assessment of ECGs [7]. These might pave the way towards novel screening programs to prevent cardiovascular events before their manifestation [8]. However, although studies have developed multiple DNNs for automatic ECG processing, there has been limited focus on analysing the performance of pre-trained models when predicting ECGs from pacemaker patients. Thereby, in this work, we analyse the influence of pacemaker activity on the performance of a state-of-the-art DNN in detecting AF.

Open Access. © 2025 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

2 Material and Methods

2.1 DNN model

We used a pretrained DNN for classification which was trained on 2,322,513 Brazilian ECGs. It accepts an array representing a 10-second 12-lead ECG as input and outputs a classification of six different diagnoses (atrioventricular block, right bundle branch block, left bundle branch block, sinus tachycardia, sinus bradycardia, and AF) [7]. However, in this work we limit the analysis to the prediction of AF only.

2.2 Dataset

As a test dataset, we used the MIMIC-IV ECG dataset annotated with ICD-10 diagnostic codes [9–12] to analyze four subgroups: (i) AF patients with pacemaker, (ii) AF patients without pacemaker, (iii) Non-AF control patients with pacemaker, and (iv) Non-AF control patients without pacemaker.

AF-ECGs were included if patients had a diagnosis of chronic AF (I48.2). Non-AF ECGs were included if ICD codes were available for the corresponding hospital stay and no diagnosis of AF (I48) was documented. A Non-AF control ECG was selected and matched for sex, age (± 1 year) and pacemaker status (Z95.0/Z45.01). ECGs were excluded if patients had ICD codes indicating an implantable cardioverter-defibrillator (Z45.02, Z45.09, Z95.810), if signals were incomplete, or if the ECG was recorded before the first documented pacemaker code. If a pacemaker code was documented during a stay, all subsequent ECGs of that patient were considered to have been recorded with a pacemaker. All prior ECGs of these patients were excluded.

Using these criteria, we identified AF and Non-AF control ECGs. In a final step, we performed 1:1 matching to obtain balanced cohorts. This resulted in the subgroups i)-iv) with each containing 541 ECG recordings in total.

2.3 Analysis

Using the DNN introduced in sec. 2.1, we predicted AF on the MIMIC-IV subset and compared the AUC and confusion matrices from patients with and without implanted pacemakers. For the confusion matrices, we used the threshold (0.390), which was provided alongside the DNN [7].

Furthermore, in order to examine how the model's classification is influenced by the different ECG segments (e.g. P-/T-wave, QRS complex), we used an established XAI pipeline [13]. It is based on the integrated gradients method [14] and assigns "relevance" values to each sample of an input ECG.

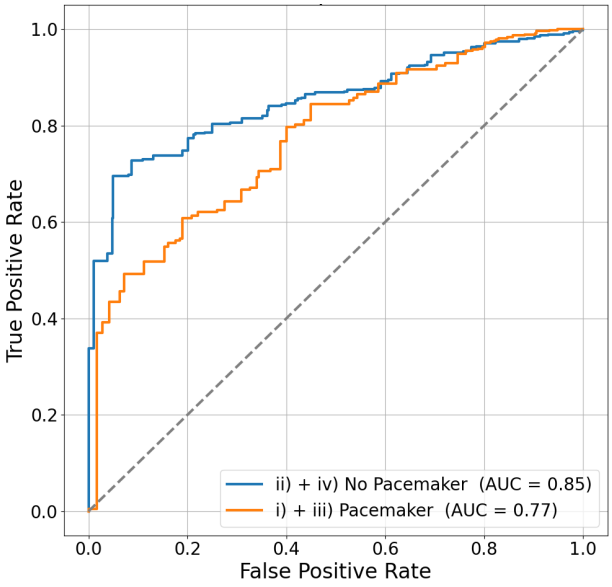

Fig. 1: ROC curve for AF detection

These can be positive, if a sample is indicating towards AF classification, or negative, if a sample is indicating towards a non-AF classification. Previous research [13] has shown that these effects are rather weak and require averaging over multiple ECG recordings to show clear effects. Hence, we customized this to our data by averaging relevance values for each subgroup i)-iv) by segmenting the generated relevances into physiologically defined intervals (P-wave, QRS, S-T, T-wave), and averaged absolute values across leads and segments.

3 Results

As shown in Fig. 1, the model achieved an AUC of 0.850 for ECGs from patients without an implanted pacemaker, compared to 0.770 for ECGs from paced patients. Tab. 1 shows that specificity and precision were comparable between groups, but the performance was better in ECGs without pacemakers, especially in terms of sensitivity (0.63 vs. 0.42). This resulted in a lower number of fn in the non-pacemaker group (200 vs. 312), and a correspondingly higher number of true positives (tp) (341 vs. 229), thereby showing a decreased performance in the presence of a pacemaker.

Tab. 1: Classification metrics for AF detection

Group	Acc.	Prec.	Sens.	Spec.	F1
ii) + iv) No Pacemaker	0.790	0.927	0.630	0.950	0.750
i) + iii) Pacemaker	0.690	0.909	0.423	0.957	0.578

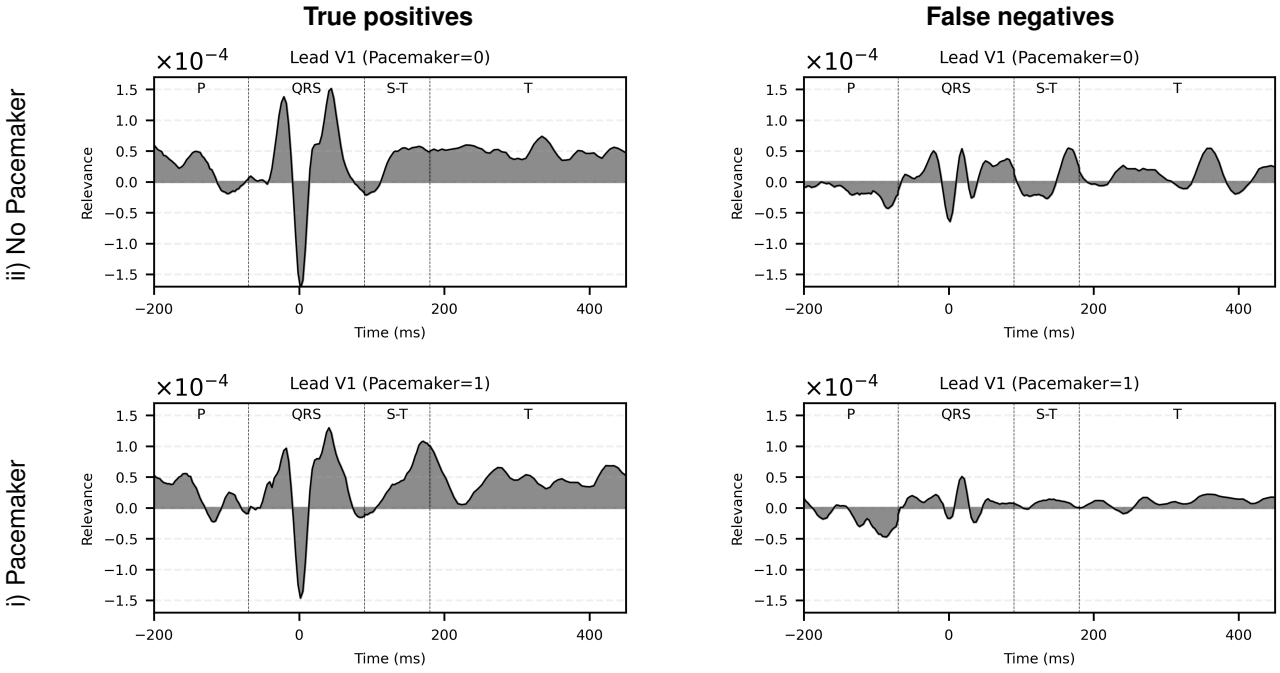

Fig. 2: Mean relevance scores for tp (left column) and fn(right column) in AF classification, separately for groups i) and ii) (n=341, 200, 229, 312). The gray area under the curve reflects the summed positive and negative relevance.

Tab. 2: Most important ECG leads ranked by pos./neg relevances for AF detection in ECGs.

Rank	ii) + iv) No Pacemaker	i) + iii) Pacemaker
1st pos.	V1 (0.00332)	V1 (0.00337)
2nd pos.	V5 (0.00282)	V3 (0.00237)
3rd pos.	V3 (0.00261)	V2 (0.00200)
1st neg.	V1 (-0.00137)	V1 (-0.00106)
2nd neg.	II (-0.00096)	II (-0.00060)
3rd neg.	I (-0.00086)	I (-0.00053)

Tab. 2 depicts the summed positive and negative relevance for AF detection in ECGs without and with implanted pacemaker. The most important lead for the model decision is lead V1, in the presence or absence of a pacemaker.

Fig. 2 depicts the mean relevance that the model assigns to the individual ECG segments in lead V1. A comparison between the tp and fn for the patient cohorts with i) and without implanted pacemakers ii) shown. The further the curve deviates from zero at a certain point in time, the more the model has used this segment to decide in or against AF. Values close to zero indicate that the corresponding segment was only of minor importance for the model. Overall, we observed a decrease in the summed positive and negative relevance from tp to fn across all ECG segments in lead V1. The overall relevance decreased by 52.7% in patients without a pacemaker and by 70.6% in patients with an implanted pacemaker. The largest relative decrease was found in the T segment in patients without pacemakers (-67.0%) and in the ST segment in patients with implanted pacemakers (-84.5%). In contrast, the smallest difference was observed in the ST segment (-35.4%) for patients without a pacemaker and in the P wave (-31.1%) for patients with a pacemaker.

4 Discussion

A fundamental challenge of DNNs in healthcare is the fairness of the models, with their generalization capabilities being an important part of ensuring fairness [15]. Models should not only perform well on the training data, but also work correctly on new, unknown data [16]. Our results demonstrate this issue in the context of AF detection in ECGs: the model performance decreased from an AUC of 0.850 to 0.770 in paced patients. In particular, sensitivity dropped from 0.63 to 0.42 (Tab. 1). This could be a result of a distribution shift in the dataset, as the model was trained on Brazilian data but evaluated on a U.S. test set, which could be due to the lower prevalence of pacemakers in Brazil [17]. This underlines the need for model robustness to population-specific characteristics.

In fn pacemaker ECGs i), the absolute relevance in the P wave segments of lead V1 decreased proportionally the least compared to tp cases, while the negative relevance increased, suggesting that these segments were particularly important for

the model's negative classification. From a physiological perspective, pacemaker can suppress or modulate atrial activity, which leads to different characteristics in the ECG. While AF without a pacemaker is typically characterised by absent P waves, pacemaker ECGs may exhibit different features suggestive of AF, such as prolonged P-wave duration (e.g., >130 ms in right atrial appendage pacing) [18]. This results in less diagnostically useful information for the model, which may have contributed to the observed higher number of fn in pacemaker ECGs.

On the other hand, the high relevance of V1 in both groups suggests that this lead is sensitive to atrial activity. This finding is consistent with V1 being widely used for the detection of atrial disorders [19].

This work has some limitations. Although a pacemaker diagnosis code may be present, it does not guarantee that active pacing was present during the ECG recording. Similarly, even though we focussed on patients with diagnosis of chronic AF, rhythm changes in the recorded ECGs cannot be completely excluded. However, these aspects reflect the challenges of working with real clinical data and large retrospective cohorts. These potentially confounding factors may have contributed to the moderate metrics observed. Moreover, 84% of AF cases in the MIMIC-IV dataset were labeled as unspecified and were not included, highlighting the potential need for larger datasets with more granular diagnostic labels.

5 Conclusions

Overall, our findings highlight the importance of training and evaluating DNNs in healthcare with a focus on fairness. To promote broad applicability and robustness, the diversity of patient groups must be considered.

Author Statement

This project was funded by the German Federal Ministry of Education and Research as part of the ACRIBiS project of the Medical Informatics Initiative (FK: 01ZZ2317B).Conflict of interest: Authors state no conflict of interest.

References

[1] Ma C, Xiao Z, Zhao L, et al. A review on atrial fibrillation detection from ambulatory ECG. IEEE Trans Biomed Eng 2024;71:876–892.

[2] Healey JS, Connolly SJ, Gold MR, Israel CW, Van Gelder IC, Capucci A, et al. Subclinical atrial fibrillation and the risk of stroke. N Engl J Med 2012;366:120–129.

[3] Nielsen JC. Mortality and incidence of atrial fibrillation in paced patients. J Cardiovasc Electrophysiol 2002;13:S17–22.

[4] Glikson M, Nielsen JC, Kronborg MB, Michowitz Y, Auricchio A, Barbash IM, et al. 2021 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy. Eur Heart J 2021;42:3427–520.

[5] Kreimer F, Gotzmann M. Pacemaker-induced atrial fibrillation reconsidered—associations with different pacing sites and prevention approaches. Front Cardiovasc Med 2024;11.

[6] Smulyan H. The computerized ECG: friend and foe. Am J Med 2019;132(2):153–160.

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

[8] Hempel T, Bender T, Gandhi K, Spicher N. Towards explaining deep neural network-based heart age estimation. In: Proc IEEE EMBS Special Topic Conf Data Sci Eng Healthc Med Biol. Malta: IEEE; 2023. p. 41–42.

[9] Gow B, Pollard T, Nathanson LA, Johnson A, Moody B, Fernandes C, et al. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset (version 1.0). PhysioNet 2023.

[10] Strodthoff N, Lopez Alcaraz JM, Haverkamp W. MIMIC-IV-ECG-Ext-ICD: Diagnostic labels for MIMIC-IV-ECG (version 1.0.0). PhysioNet 2024.

[11] Strodthoff N, Lopez Alcaraz JM, Haverkamp W. Prospects for Artificial Intelligence-Enhanced ECG as a Unified Screening Tool for Cardiac and Non-Cardiac Conditions – An Explorative Study in Emergency Care. Eur Heart J Digit Health 2024;ztae039.

[12] Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000;101:e215–e220.

[13] Bender T, et al. Analysis of a deep learning model for 12-lead ECG classification reveals learned features similar to diagnostic criteria. IEEE J Biomed Health Inform 2024;28(4):1848–1859.

[14] Kokhlikyan N, Miglani V, Martin M, Wang E, Reynolds J, Melnikov A, et al. Captum: A unified and generic model interpretability library for PyTorch. arXiv 2020;2009.07896.

[15] Liu M, Ning Y, Teixayavong S, et al. A translational perspective towards clinical AI fairness. NPJ Digit Med 2023;6:172.

[16] Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy AI: From principles to practices. ACM Comput Surv 2023;55(9):177.

[17] Rojel U, Diaz JC, Figueiredo MJO, Di Biase L, Saad E, Aguinaga-Arrascue L, et al. Current state of arrhythmia care in Latin America: A statement from the Latin American Heart Rhythm Society. Heart Rhythm O2 2025;6(1):112–126.

[18] Endoh Y, Miyazawa T, Kakuta T, Yashiro S, Hori Y, Katoh T, et al. Identification of pacing rhythm using deep learning with twelve-lead electrocardiogram. J Cardiovasc Electrophysiol 2003;14:1019–24.

[19] Lebek S, Wester M, Pec J, Poschenrieder F, Tafelmeier M, Fisser C, et al. Abnormal P-wave terminal force in lead V1 is a marker for atrial electrical dysfunction but not structural remodelling. ESC Heart Fail 2021;8:4055–4066.