Transfer Learning to Focus Self-Learning AI on Rhythm Improves Interpretability in Atrial Fibrillation Detection

Alexander Hammer, Matteo Zannini, Hagen Malberg, Martin Schmidt

Institute of Biomedical Engineering, TU Dresden, Germany

Abstract

Explainable AI (xAI) can identify samples that are relevant for deep neural networks (DNNs) to detect cardiovascular diseases in ECGs. However, interpretability is limited as these explanations are not always related to common diagnostic criteria. xECGArch comprises two convolutional neural networks (CNNs) with different temporal focus. As shown previously, the long-term CNN (LT-CNN) emphasizes QRS complexes, whose relevance changes correlate with rhythm. To test whether QRS morphology is decisive, we applied transfer learning to make the LT-CNN focus on rhythm for atrial fibrillation (AF) detection, using 10 s single-lead ECGs from public databases. The LT-CNN was trained on 9,675 ECGs to detect R peaks and tested on 1,320 unseen ECGs, reaching a sample-accurate F1 score of 98.1%. The pre-trained model was then fine-tuned on AF detection using 8,868 ECGs, with the weights of none or the first 3 to 8 of 9 layers frozen, reaching F1 scores of 87.6% to 93.3% on 986 unseen ECGs, decreasing with increasing number of frozen layers. A systematic validation of explanations extracted with deep Taylor decomposition shows increasing ($p < 0.001$) model focus on R peaks for AF detection with more frozen layers, peaking at 7. Our results indicate that transfer learning can guide DNNs to use specific features, enhancing interpretability and moving toward trustworthy AI for clinical applications.

1. Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia, with a lifetime risk of 22%–36% [1]. Untreated, it increases morbidity, especially stroke, by up to 5 times and mortality by up to 2 times [2]. Early detection allows interventions that can prevent severe outcomes [3]. Because AF is often paroxysmal in early stages, detection requires long-term electrocardiographic monitoring [3], which is labor-intensive to analyze. Deep neural networks (DNNs) from the field of deep learning (DL) achieve high performance in AF detection from electrocardiograms (ECGs) [4]. However, due to their complexity, their decision-making process lacks explainability and their self-learned features lack interpretability [5, 6]. Nevertheless, both is required for their integration into clinical routine as trustworthy diagnostic support [5, 7].

Explainable AI (xAI) methods provide *post-hoc* explanations that approximate each input value's relevance for the DNN decision. This helps to identify relevant ECG segments [6, 7]. However, it is unclear which information is used by a DNN from these segments and how it relates to diagnostic criteria, which is why the interpretability of the explanations is limited [6].

xECGArch [8] is a DL architecture, comprising two convolutional neural networks (CNNs) with different temporal focus, making it interpretable by design. A systematic validation of explanations extracted with deep Taylor decomposition (DTD) demonstrated that the short-term CNN self-learns morphological features from the ECG while the long-term CNN (LT-CNN) focuses on QRS complexes [6, 8–11], with relevance changes in the QRS complexes correlating with rhythm [6]. However, it is unclear whether the QRS morphology is a decisive factor to the LT-CNN. Therefore, in this study, we investigate the potential of transfer learning (TL) to make the LT-CNN focus particularly on rhythm for AF detection in single-lead ECGs.

TL commonly involves pre-training a model on a large dataset and fine-tuning it on a smaller one, with the classification task being identical (e.g., [12]) or more specific (e.g., [13]) during fine-tuning. Weimann and Conrad [14] tested various pre-training tasks, including heart rate categorization, for fine-tuning on detecting AF, normal sinus rhythm, or other pathologies, using optimal pre-training weights as the starting point. In contrast, our approach pre-trains on sample-precise R peak detection to avoid learning morphological features, and freezes early-layer weights during fine-tuning to preserve information. We then examine how the number of frozen layers affects classification performance and model explanations.

2. Methods

In this study, the untrained LT-CNN [8] was pre-trained on R peak detection and fine-tuned on AF detection.

Table 1. Dataset description for pre-training and fine-tuning tasks. *AF*, atrial fibrillation; *bpm*, beats per minute; *f*, female; *HR*, heart rate; *m*, male; *N*, normal sinus rhythm; *O*, other pathology; *SD*, standard deviation.

Set	Class	*n*	Sex [%]	Age	HR	Label [%]
			f / m	*mean ± SD*	*mean ± SD*	*AF / N / O*
Pre-training (R peak detection)						
Train	-	16,189	45.7 / 54.3	64.3 ± 17.4	81.7 ± 25.2	7.1 / 31.9 / 61.0
Test	-	2,858	45.5 / 54.4	64.4 ± 17.6	82.7 ± 26.3	7.1 / 31.4 / 61.5
Total	-	19,047	45.7 / 54.3	64.3 ± 17.4	81.8 ± 25.4	7.1 / 31.8 / 61.1
Fine-tuning (AF detection)						
Train	AF	4,420	42.3 / 57.7	71.9 ± 11.8		
	non-AF	4,448	45.9 / 54.1	60.7 ± 16.7		
	Total	8,868	44.1 / 55.9	66.3 ± 15.5		49.8 / 5.1 / 45.1
Test	AF	507	42.2 / 57.8	71.6 ± 12.3		
	non-AF	479	44.1 / 55.9	60.6 ± 17.5		
	Total	986	43.1 / 56.9	66.3 ± 16.0		51.4 / 4.1 / 44.5
Total	AF	4,927	42.3 / 57.7	71.9 ± 11.9		
	non-AF	4,927	45.7 / 54.3	60.7 ± 16.8		
	Total	9,854	44.0 / 56.0	66.3 ± 15.6		50.0 / 5.0 / 45.0

2.1. Dataset Description

Originally, xECGArch has been trained and tested on 9,854 ECGs from 4 publicly available databases that have been used in the George B. Moody PhysioNet Challenge 2021 [15, 16]. In this study we used the same data with the original training and test splits for the fine-tuning task. For pre-training, we used 19,047 ECGs from the same databases, ensuring that no recordings that appeared in the test set of the fine-tuning task were present in the pre-training set. From each recording, we used Einthoven II and the middle 10 s. The composition of the data by age, gender and heart rate (pre-training) is described in Table 1. The data was pre-processed in line with [8].

R peak annotations were generated automatically using 6 open source QRS detectors [17–22]. Detected R peaks were accepted for annotation if at least two QRS detectors found an R peak within a 0.1 s window. The first of the found peaks within the window was further corrected at the peak using the PhysioNet [16] waveform database function `correct_peaks` to make sure that the annotation is on top of a local maximum.

2.2. Model Training and Validation

The LT-CNN consists of 9 convolutional blocks, containing 1D convolution, batch normalization and a rectified linear unit each, global average pooling and a softmax activation layer. Here, we added a dropout layer with 25% drop out rate after the last convolutional layer for extended robustness. For the pre-training task, the global average pooling layer has been removed to match input and output size for one-hot-encoded R peak detection.

For the fine-tuning task the weights from the best model from the pre-training task were loaded and frozen in none or the first 3 to 8 layers in separate experiments, while training the remaining layers.

For each experiment, we conducted a grid search on the training set in a 5-fold cross validation, using the Adam optimizer to minimize the categorical cross-entropy loss. Due to the imbalance between data points with and without R peak, a 0.7 to 0.3 weighting was applied during pre-training. During grid search, we optimized the learning rate [1e-3, 1e-4, 1e-5] and the batch size (pre-training: [2, 3, 4], fine-tuning: [8, 16, 32]). The maximum training duration was set to 150 epochs with early stopping after 20 epochs on plateau. The best model according to F1 score was applied to an unseen test set, containing 15% of the data during pre-training or 10% during fine-tuning.

2.3. Evaluation of Model Explanations

We extracted model explanations using DTD as it provided the most trustworthy explanations for xECGArch in a systematic comparison of 13 xAI methods using perturbation [8]. We then determined relative relevance (rR) values by recording-wise scaling relevance values to [0, 1].

Subsequently, mean rR per interval and recording was calculated in line with [6], using iterative two-dimensional signal warping (i2DSW) [23, 24] for robust fiducial point detection. The intervals included the interval from Q peak to R peak (Q), the R peak ± 1 sample (R), the interval from R peak to S peak (S), and everything beside the QRS complex (notQRS). The mean rR per segment and recording was normalized on average rR per recording ($\mathrm{rR}_{\mathrm{norm}}$).

Using a 2-factor analysis of variance (ANOVA) followed by Tukey-Kramer *post-hoc* test, we investigated the effect of intervals and the number of frozen layers as independent factors on the mean rR as the dependent factor.

3. Results

After pre-training, the LT-CNN reached an F1 score of 98.1% in point-precise R peak detection on the unseen test set, containing a total of 39,379 annotated R peaks, which equals 13.8 ± 4.4 (mean ± standard deviation (SD)) R peaks per recording and an average heart rate of 82.7 ± 26.3 beats per minute (bpm), compared to 39,508 predicted R peaks, which equals 13.8 ± 4.5 R peaks per recording and an average heart rate of 82.9 ± 26.7 bpm. Additional model performance metrics are summarized in Table 2.

The fine-tuned LT-CNN reached an overall decreased F1 score in AF detection of 87.6% to 93.3% for 8 to 0 frozen layers compared to 95.1%, reached by the original LT-CNN [8]. Noticeably, the model performance keeps steady for 0 to 3 frozen layers and decreases steadily with increasing number of frozen layers, with a particularly steep drop in the F1 score of 3.3% between 7 and 8 frozen layers, compared to 0.1% to 1.3% for previous steps (see Table 2).

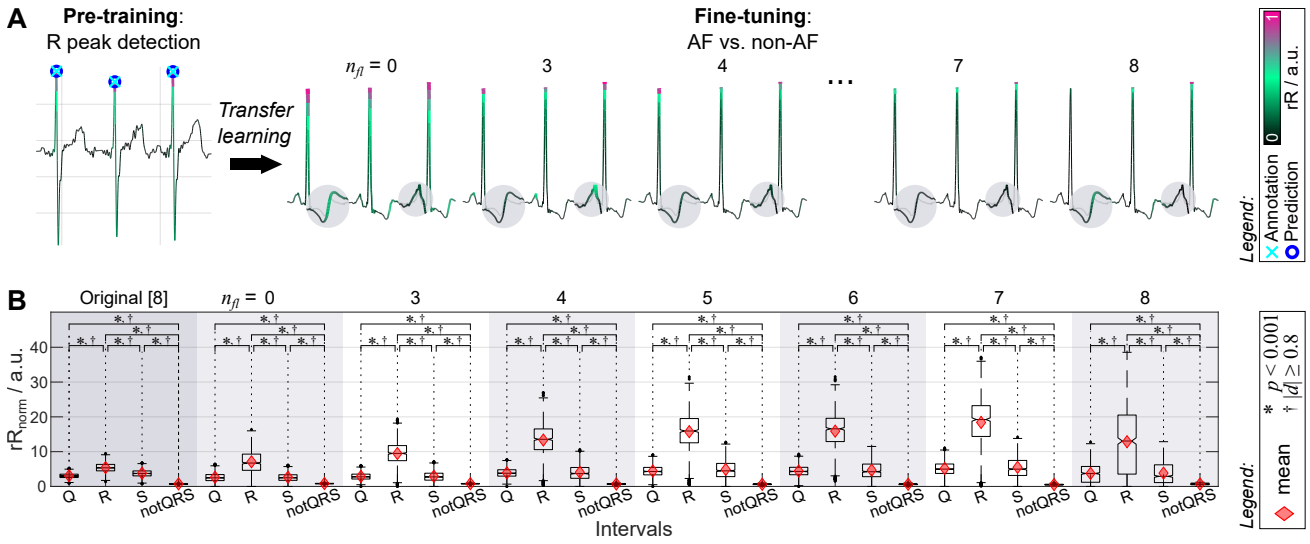

Figure 1. Model explanations in terms of sample-wise relative relevance (rR) during pre-training and fine-tuning with different numbers of frozen layers (n_{fl}) on atrial fibrillation (AF) detection; **A** as saliency maps for example ECGs and **B** systematically investigated for rR averaged per interval type and normalized on average rR per ECG recording ($\mathrm{rR}_{\mathrm{norm}}$), with significance values according to Tukey-Kramer *post-hoc* test for inter-interval differences for each n_{fl}-configuration.

Table 2. Model performance during 5-fold cross validation (for training and validation set) for optimal learning rate (lr) and batch size (bs) and during test on unseen test set with different numbers of frozen layers (n_{fl}). All performance metrics are given in %. *SD*, standard deviation.

Metric	**Pre-train**	**Orig. [8]**	**Fine-tune**						
n_{fl}	-	-	0	3	4	5	6	7	8
lr	1e-3	1e-4	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
bs	3	8	16	32	16	16	8	16	32
Training (5-fold cross validation)									
F1									
mean	98.1		95.8	95.7	94.2	94.0	93.4	92.2	88.4
SD	0.3		1.7	0.8	0.6	1.1	1.5	1.0	0.8
max	98.5		97.1	96.4	94.6	94.8	93.4	92.3	87.5
Validation (5-fold cross validation)									
F1									
mean	97.8		92.8	92.6	91.6	91.4	90.3	89.6	87.2
SD	0.2		0.4	0.4	0.6	0.8	1.2	0.9	0.8
max	98.1		93.2	93.1	92.3	91.9	91.7	90.6	88.3
Test									
Accuracy	100.0	95.3	93.2	93.2	91.9	91.5	90.4	90.1	86.4
Sensitivity	98.2	94.9	91.3	92.1	91.7	90.3	90.9	93.1	93.7
Precision	97.9	95.6	95.3	94.5	92.5	92.9	90.4	88.2	82.3
Specificity	100.0	95.6	95.2	94.4	92.1	92.7	89.8	86.8	78.7
F1	98.1	95.1	93.3	93.3	92.1	91.6	90.7	90.6	87.6

Exemplary model explanations in Figure 1 A show a clear focus on QRS complexes during R peak detection. During fine-tuning, an increasing focus on the R peak and a decreasing rR of surrounding areas, especially the P and T waves, is observed with increasing number of frozen layers. However, from 7 to 8 frozen layers, the focus of the R peak decreases again and the rR of the T wave increases.

Figure 1 B shows the systematic investigation on rR per interval and number of frozen layers. ANOVA revealed both of them and their interaction to be significant factors ($p < 0.001$) on rR. The *post-hoc* analysis was performed only for significant main factors. Results of the *post-hoc* analysis are shown in Figure 1 B for rR differences between intervals within the same model configuration and in Table 3 for rR differences between R intervals of different model configurations. For each number of frozen layers, the R peak was significantly ($p < 0.001$) the most relevant interval. It was observed that with an increasing number of frozen layers, the rR of the R peak increased significantly ($p < 0.001$) in relation to the mean rR of the entire signal until it dropped significantly ($p < 0.001$) in median from 7 to 8 frozen layers, with the interquartile range increasing.

4. Discussion & Conclusion

The LT-CNN achieved a sample-accurate F1 score of 98.1% on the unseen test set for R peak detection, clearly focusing the QRS complexes with the R peaks being most relevant. The number of frozen layers affected the models ability to detect AF in the xECGArch test dataset, with an F1 score improving from 87.6% to 93.3% for 8 to 0 layers frozen. Furthermore, with an increasing number of frozen layers of up to 7, there is a focus shift from the surrounding area to the R peaks, before it drops for 8 frozen layers. This indicates that the DNN requires sufficient free layers, which weights can be adjusted during fine-tuning, to solve the transfer task. An exact threshold for the number of layers to be frozen cannot be derived from this study.

Table 3. Results of the Tukey-Kramer *post-hoc* test on differences in R intervals of different model configurations regarding their relative relevance, including significance values (*: $p < 0.001$) and Cohen's d for effect sizes († $0.2 \leq |d| < 0.5$, †† $0.5 \leq |d| < 0.8$, †††: $0.8 \leq |d|$).

	Model configurations - number of frozen layers (n_{fl})						
	$n_{fl} = 0$	3	4	5	6	7	8
Orig. [8]	* ††	* †††	* †††	* †††	* †††	* †††	* †††
$n_{fl} = 0$		* ††	* †††	* †††	* †††	* †††	* †††
3			* †††	* †††	* †††	* †††	* †
4				* †	* †	* ††	- -
5					- -	* †	* †
6						* †	* †
7							* ††

Overall, our findings suggest that TL can guide a DNN to use specific characteristics for solving a task on costs of a small reduction in accuracy. This might in future applications enhance the models' interpretability, a prerequisite for trustworthiness and the use of AI in clinical practice.

Acknowledgments

This project is co-funded by the European Union and co-financed from tax revenues on the basis of the budget adopted by the Saxon State Parliament.

References

[1] Mou L, *et al*. Lifetime risk of atrial fibrillation by race and socioeconomic status: ARIC study (atherosclerosis risk in communities). Circ Arrhythm Electrophysiol 2018; 11(7):e006350.

[2] Odutayo A, *et al*. Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: Systematic review and meta-analysis. BMJ 2016;354:i4482.

[3] Brundel BJJM, *et al*. Atrial fibrillation. Nat Rev Dis Primers 2022;8(1):21.

[4] Stracina T, Ronzhina M, Redina R, Novakova M. Golden standard or obsolete method? Review of ECG applications in clinical and experimental context. Front Physiol 2022; 13:867033.

[5] Holzinger A, *et al*. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov 2019;9(4):e1312.

[6] Hammer A, *et al*. Fusion of automatically learned rhythm and morphology features matches diagnostic criteria and enhances AI explainability. NPJ Artif Intell 2025;1(19).

[7] Gershman SJ, Horvitz EJ, Tenenbaum JB. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science 2015; 349(6245):273–278.

[8] Goettling M, Hammer A, Malberg H, Schmidt M. xECGArch: A trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features. Sci Rep 2024;14:13122.

[9] Hammer A, *et al*. An explainable AI for trustworthy detection of atrial fibrillation on reduced lead ECGs in mobile applications. Eur Heart J 2024;45(Supplement_1):ehae666.3497.

[10] Hammer A, *et al*. Explainable and interpretable AI visualises self-learned clinically relevant ECG characteristics of rhythm and morphology paving the way for trustworthy diagnostic support. Eur Heart J 2025;46(Supplement_1).

[11] Hammer A, Malberg H, Schmidt M. Morphology features self-learned by explainable deep learning for atrial fibrillation detection correspond to fibrillatory waves. In CinC 2024, volume 51. Karlsruhe, Germany; 1–4.

[12] Avetisyan A, *et al*. Deep neural networks generalization and fine-tuning for 12-lead ECG classification. Biomed Signal Process Control 2024;93:106160.

[13] Wang Z, Stavrakis S, Yao B. Hierarchical deep learning with generative adversarial network for automatic cardiac diagnosis from ECG signals. Comput Biol Med 2023; 155:106641.

[14] Weimann K, Conrad TOF. Transfer learning for ECG classification. Sci Rep 2021;11(1):5251.

[15] Reyna MA, *et al*. Issues in the automated classification of multilead ECGs using heterogeneous labels and populations. Physiol Meas 2022;43(8):084001.

[16] Goldberger A, *et al*. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circ 2000;101(23):e215–e220.

[17] Johnson AE, *et al*. R-peak estimation using multimodal lead switching. In CinC 2014, volume 41. Cambridge, MA, USA; 281–284.

[18] Khamis H, *et al*. QRS detection algorithm for telehealth electrocardiogram recordings. IEEE Trans Biomed Eng 2016;63(7):1377–1388.

[19] Moeyersons J, *et al*. R-DECO: An open-source Matlab based graphical user interface for the detection and correction of R-peaks. PeerJ Comput Sci 2019;5:e226.

[20] Emrich J, Koka T, Wirth S, Muma M. Accelerated sample-accurate R-peak detectors based on visibility graphs. In European Signal Processing Conference (EUSIPCO), volume 31. IEEE, 2023; 1090–1094.

[21] Moody GB, Pollard T, Moody B. WFDB software package (version 10.6.2). PhysioNet, 2021.

[22] Makowski D, *et al*. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behav Res Methods 2021;53(4):1689–1696.

[23] Schmidt M, *et al*. Two-dimensional warping for one-dimensional signals—Conceptual framework and application to ECG processing. IEEE Trans Signal Process 2014; 62(21):5577–5588.

[24] Schmidt M, Baumert M, Malberg H, Zaunseder S. Iterative two-dimensional signal warping—Towards a generalized approach for adaption of one-dimensional signals. Biomed Signal Process Control 2018;43:311–319.

Address for correspondence:

Alexander Hammer
Institute of Biomedical Engineering, TU Dresden
Fetscherstr. 29, 01307 Dresden, Germany
alexander.hammer@tu-dresden.de