Detecting Atrial Fibrillation from Reduced-Lead Electrocardiograms of Mobile Patches Using Interpretable Features

Alexander Hammer[1], Boris Schmitz[2,3], Hagen Malberg[1], Martin Schmidt[1]

[1]Institute of Biomedical Engineering, TU Dresden, Dresden, Germany; [2]Department of Rehabilitation Sciences, Faculty of Health, University of Witten/Herdecke, Witten, Germany; [3]DRV Clinic Königsfeld, Center for Medical Rehabilitation, Ennepetal, Germany

Abstract

Long-term electrocardiograms (ECGS) recorded with mobile patches can help to detect paroxysmal diseases like atrial fibrillation (AF) when combined with automated ECG analysis. However, mobile patches provide a reduced number of leads that differ in signal morphology. We therefore investigated how reduced-lead ECGs affect AF detection, using 2,478 publicly available 12-lead ECGs of 30 s each. The feature set comprised 186 interpretable features per lead, including heart rate variability, morphology features, and signal quality indices. Binary decision tree ensembles were trained to detect AF, normal sinus rhythm (N), and other anomalies (O) in 8 different lead configurations. We also prospectively evaluated the applicability of the 3-lead model to 1,601 mobile long-term ECGs from the TIMELY eHealth project. Although the discriminability of AF, N, and O decreased with the number of leads, we achieved a minimum F1 score of 0.907 for single-lead III, compared with the highest F1 score of 0.957 when using 12-leads. P wave and PQ segment morphology features were consistently the most relevant. In mobile long-term ECGs, we were able to correctly identify AF in 5 patients, 2 of whom had no previous diagnosis. Overall, we achieved reliable classifications across different lead configurations and were able to demonstrate the potential of our approach for mobile applications.

1. Introduction

Atrial fibrillation (AF) is the most prevalent cardiac arrhythmia globally [1]. Early detection and treatment of AF can reduce the risk of morbidity and mortality[2], which are significantly increased without intervention [1]. AF is typically paroxysmal and asymptomatic in its early stages [3], which makes diagnosis challenging. The use of mobile electrocardiography patches, along with automated AF detection algorithms, can increase the likelihood of early detection of AF episodes. Combined with high-performance algorithms for automated ECG analysis, they represent a promising tool for diagnostic support. However, due to a reduced number of leads compared to the standard 12-lead ECG, they provide less information. The results of the PhysioNet/Computing in Cardiology (CinC) Challenge 2021 demonstrated inconsistencies between different approaches, yet collectively indicated a reduction of ECG leads hardly weakens the averaged accuracy of automated anomaly detection [4, 5]. Nevertheless, no inferences can be made regarding the impact of information reduction on the detection quality of individual classes by different models. In addition, the leads from mobile patches usually differ from those of the standard ECG due to the different electrode placement. The information content is therefore changed and a different lead morphology is obtained. The lack of annotated mobile ECGs to train high-performance models raises the question of whether AF detection models trained on standard ECGs are transferable to mobile ECGs.

We therefore conduct two studies, as illustrated in Figure 1. The first is a validation study, investigating the influence of the number and selection of leads on the quality of AF detection. In favor of interpretability, we use an approach based on manual feature extraction and comprehensible decision tree ensembles as classifiers. The second is a prospective study, which investigates the transferability of a classifier trained on standard ECGs to mobile ECGs.

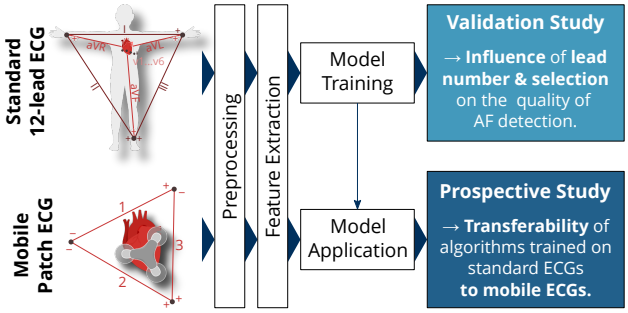


Figure 1. Designs of validation and prospective study including the respective data origin.

2. Methods

2.1. Data material

We combined the public and the hidden China Physiological Signal Challenge 2018 databases (CPSC2018, CPSC2018_2) [6] with the Physikalisch-Technische Bundesanstalt Diagnostic ECG XL database (PTB-XL) [7, 8], see Table 1. All databases included 12-lead ECGs and rhythm/anomaly annotations that were summarized as AF, normal sinus rhythm (N), and other anomalies (O). We discarded ECGs shorter than 30 s and cropped ECGs longer than 30 s (centered). ECGs exceeding 60 s were split into consecutive windows, leading to a total of $n = 2,478$ ECGs ($n_{AF} = 179$, $n_N = 359$, $n_O = 1,940$).

2.2. Signal processing

All ECGs were band-pass filtered between 0.3 Hz and 120 Hz and notch filtered at 60 Hz, using zero padding to avoid boundary effects. We applied QRS detection [4, 9] and iterative two-dimensional signal warping (i2DSW) [10, 11] to robustly extract RR intervals, beat templates, and morphological beat-to-beat changes. Noisy heart beats and abnormal RR intervals were rejected [10].

In addition to 4 global features (age, sex, sampling frequency, and gain), we extracted 186 features from each lead separately, including 27 heart rate variability (HRV) and 144 morphological features as well as 15 signal quality indices (SQIs).

a) HRV features: From the remaining normal RR intervals, we calculated standard heart rate metrics (mean, median, minimum, maximum) as well as statistical, geometric, nonlinear, and frequency-based HRV features [9].

b) Morphological features: In order to extract the template beat features, we determined the height, length, baseline slope, area, and skew of the P and T waves, as well as the QRS complex. Additionally, we determined the length and slope of PQ and QT segments. To account for morphological beat-to-beat changes, we extracted these features, for each individual beat and calculated the mean, median, and standard deviation (SD) across all beats of a 30 s window, as well as the inter-beat similarity of beat-segment waveform. The distribution of F waves was included as the mean, SD, and distribution skewness of F-F distances [12]. Furthermore, QT interval variability (QTV) features were extracted, including the SD of QT intervals (SDQT) and the QTV index (QTVi), as well as the T wave amplitude corrected measures cQTV, cSDQT, and cQTVi. Additionally, the mean, median, minimum, and maximum QT length were used. [9]

c) SQIs: We considered the entity of the signal quality using the rejection and filtration rates, various SQIs from the fecgsyn toolbox [13], and features to describe the frequency and shape of peaks. [9]

For application to mobile ECG recordings, we created a reduced feature set, comprising the HRV features, 50 beat-to-beat morphological variability features, and 2 SQIs per lead, which are less dependent on lead morphology and therefore suitable for the application to mobile ECGs.

Table 1. Number n of ECG recordings per database and class as well as mean and standard deviation (SD) of age and sex (f, female; m, male) of the corresponding subjects.

	Database			**Class**			**Total**
	CPSC2018 [6]		**PTB-XL**	**AF**	**N**	**O**	
	Public	Hidden	[7, 8]				
n	586	315	1,577	179	359	1,940	2,478
Age *Mean ± SD (years)*	71.5 ± 12.8	70.7 ± 14.9	66.3 ± 11.5	69.5 ± 12.8	40.8 ± 14.9	59.3 ± 14.3	57.6 ± 16.0
Sex *f : m*	35 : 65	52 : 48	11 : 89	28 : 72	33 : 67	35 : 65	34 : 66

2.3. Classification model

Based on the PhysioNet/CinC Challenge 2021 [5], we considered different lead sets for training, including single (I, II, III), 2- (II, V1), 3- (I-III), 4- (I-III, V1), 6- (I-III, aVR, aVL, aVF), and 12-leads. For improved interpretability, we trained binary decision tree ensembles for AF, N, and O detection, respectively. 5-fold cross validation (cv) was used for training and model validation. To determine the optimal model parameters, we applied a grid search, varying the boosting algorithm, the number of bins between 64 and 256, the number of learners between 128 and 1,024, and the number of splits between 2 and 4. The learning rate was set to 0.1. Each hyperparameter combination was tested with and without undersampling. The models were evaluated using the accuracy, the area under the precision-recall curve (AUPRC), and the F1 score.

For each model, the feature relevance was determined using the Matlab function *PredictorImportance*.

2.4. Application to mobile ECG recordings

The 3-lead model, generated from the reduced feature set, was applied to a preliminary subset of the prospective data from the TIMELY living lab study [14]. The long-term ECGs were recorded using the mobile 3-lead net_ECG patch (livetec Ingenieurbüro GmbH, Lörrach, Germany) with a sampling frequency of 800 Hz. For processing, the ECGs were split into windows of 30 s duration with 50 % overlap. Noisy and artifact affected windows were removed. To reduce misclassifications, a sliding median filter with a range of five windows was applied to the predictions.

3. Results

3.1. Validation study

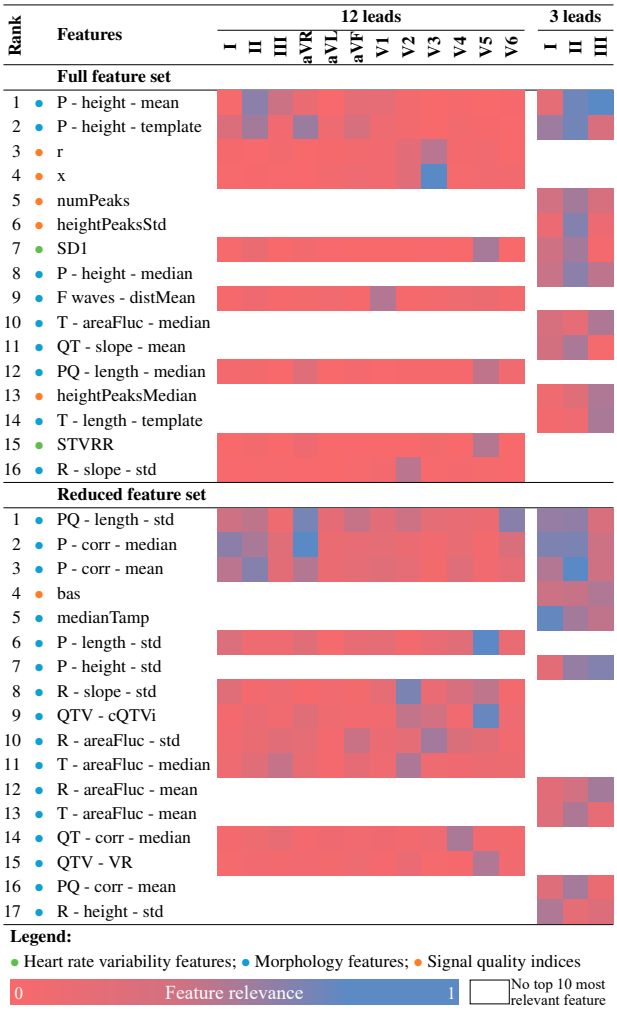
The highest F1 scores of 0.957 and 0.942 were achieved using 12 ECG leads, as detailed in Table 2. Conversely, the lowest F1 scores of 0.907 and 0.893 (full and reduced feature set) were observed with single-lead III. The 3-lead model exhibited a slightly inferior F1 score relative to the 12-lead model, with a decrease of 1.1 % and 2.1 % on the full and reduced feature sets, respectively. The F1 scores for the reduced feature set were between 1.5 % (lead I or lead III) and 3.2 % (lead II) below those for the full feature set. Notably, across all configurations, the F1 score was optimized when using GentleBoost and threshold optimization without undersampling.

Table 3 shows the most relevant features for the 12-lead and the 3-lead configuration, for both the full and the reduced feature sets. For both feature sets, P wave and PQ segment morphology features are particularly relevant, supplemented by SQI features mainly in the case of the full feature set and morphology features of other ECG segments mainly in the case of the reduced feature set. In the full feature set, the most relevant HRV feature for both, the 12-lead and the 3-lead configuration is the Poincaré feature *SD1*, while in the reduced feature set there is no HRV feature in the top 10 most relevant features.

Table 2. Optimal number of bins (n_b), learners (n_l) and splits (n_s) per lead configuration according to grid search with corresponding performance metrics. For all configurations, the F1 scores was maximized using GentleBoost and threshold optimization but no undersampling.

Leads	Hyperparameters			Performance					
				F1 score				Accuracy	AUPRC
	n_b	n_l	n_s	AF	N	O	∅ Macro	∅ Macro	∅ Macro
Full feature set									
12	256	768	4	0.96	0.93	0.98	0.957	0.980	0.976
6	256	640	4	0.96	0.92	0.98	0.951	0.977	0.972
4	128	896	3	0.94	0.93	0.98	0.948	0.978	0.969
3	192	1024	3	0.95	0.91	0.98	0.946	0.976	0.968
2	96	512	4	0.93	0.91	0.97	0.936	0.972	0.964
1: I	96	512	4	0.92	0.88	0.97	0.921	0.967	0.956
1: II	128	896	4	0.92	0.89	0.97	0.927	0.969	0.961
1: III	160	768	4	0.90	0.86	0.96	0.907	0.961	0.945
Reduced feature set									
12	224	1024	3	0.95	0.90	0.97	0.942	0.974	0.965
6	96	896	4	0.94	0.89	0.97	0.931	0.969	0.958
4	192	896	4	0.95	0.88	0.97	0.931	0.968	0.958
3	160	512	4	0.94	0.87	0.96	0.922	0.965	0.941
2	160	768	4	0.93	0.85	0.96	0.915	0.962	0.948
1: I	128	896	4	0.93	0.84	0.96	0.907	0.958	0.944
1: II	128	1024	4	0.93	0.81	0.96	0.897	0.955	0.931
1: III	128	1024	4	0.89	0.84	0.95	0.893	0.954	0.927

Table 3. Relative relevance of the ten most relevant features of the 12-lead configuration compared to the ten most relevant features of the 3-lead configuration, color-coded according to the legend.



3.2. Prospective study

Our algorithm was applied to 1,601 recordings available at the time of this preliminary prospective study. The recordings were from 187 patients with documented coronary artery disease and had a duration of 23:05 h ± 9:41 h (mean ± SD) each. ECG segments, automatically classified as AF, were manually reviewed by a specialist and a senior cardiologist. AF was detected in 5 patients, of whom 2 had no previous diagnosis. Based on these incidental findings, further diagnostic measures were initiated.

4. Discussion and Conclusion

In the validation study, we investigated the influence of ECG lead reduction on AF detection using manually ex-

tracted, interpretable features given into binary decision tree ensembles. With the full feature set, consisting of 5 global features and 186 features per lead, we achieved an F1 score of 0.957 with 12 leads, which was reduced by 1.1 % when limited to 3 leads. When restricted to a reduced feature set, primarily consisting of the HRV and fewer lead-dependent morphological variability features, the F1 score decreased by 1.5 % – 3.2 % and was still 0.893 (lead III) – 0.942 (12 leads), which corresponds to a very good classification performance. Depending on the lead configuration, the relevance of individual features changed, but without a recognizable trend. P wave or PQ segment morphology features and SQIs remain the most relevant, particularly in lead II, where F waves and P waves are particularly visible. The reduced relevance of these features in chest wall leads may be either due to electrophysiological causes or the fact that the features used are optimized for limb leads in particular. The high relevance of T wave, R peak, and QT segment features may be due to the deformation of these segments resulting from the superposition of F waves [9]. Similarly, the high relevance of the SQIs can be attributed to their necessity for the valuation of morphological features.

In the prospective study, we applied the 3-lead model, trained on the reduced feature set, to mobile long-term ECGs. We identified 2 patients with undiagnosed AF of whom further diagnostic measures were initiated. Due to lack of data annotation, it is not yet possible to make a definitive statement about our model's classification quality, particularly the specificity, on mobile ECGs. However, annotation of a subset of the data is under consideration for full model validation at the end of the TIMELY study.

In conclusion, we were able to show that the accuracy of AF detection using interpretable, manually extracted ECG markers and explainable decision tree ensembles hardly decreases with a reduction in the number of leads and is still at an F1 score of 0.946 or an accuracy of 0.976 for 3 leads. We have achieved reliable classifications and explanations across all derivative configurations. We have also shown that machine learning models trained on standard ECGs are transferable to mobile ECGs and are able to detect previously undetected AF. Our findings provide a foundation for the prospective utilization of such models in mobile applications, which is essential for the continuous monitoring of high-risk patients.

Acknowledgments

This study was supported by grants from the European Union's Horizon 2020 research and innovation program (TIMELY, No. 101017424).

References

[1] Chugh SS, *et al*. Worldwide epidemiology of atrial fibrillation: a global burden of disease 2010 study. Circulation 2014;129(8):837–847.

[2] Roth GA, *et al*. Global burden of cardiovascular diseases and risk factors, 1990–2019. J Am Coll Cardio 2020; 76(25):2982–3021.

[3] Padfield GJ, *et al*. Progression of paroxysmal to persistent atrial fibrillation: 10-year follow-up in the canadian registry of atrial fibrillation. Heart Rhythm 2017;14(6):801–807.

[4] Hammer A, Scherpf M, Ernst H, Weiß J, Schwensow D, Schmidt M. Automatic classification of full- and reduced-lead electrocardiograms using morphological feature extraction. In Computing in Cardiology 2021, volume 48. Brno (CZ), 2021; 1–4.

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

[6] Liu FF, *et al*. An open access database for evaluating the algorithms of ECG rhythm and morphology abnormal detection. J Med Imaging Health Infor 2018;8(7):1368–1373.

[7] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020;7(1):154.

[8] Wagner P, Strodthoff N, Bousseljot RD, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset, 2022.

[9] Hammer A, Malberg H, Schmidt M. Towards the prediction of atrial fibrillation using interpretable ECG features. In Computing in Cardiology 2022, volume 49. Tampere (FI), 2022; 1–4.

[10] Schmidt M, Baumert M, Porta A, Malberg H, Zaunseder S. Two-dimensional warping for one-dimensional signals—conceptual framework and application to ECG processing. IEEE Trans Signal Process 2014;62(21):5577–5588.

[11] 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.

[12] Hammer A, Malberg H, Schmidt M. Morphology features self-learned by explainable deep learning for atrial fibrillation detection correspond to fibrillatory waves. In Computing in Cardiology 2024, volume 51. Karlsruhe (GER), 2024; 1–4.

[13] Andreotti F, Behar J, Zaunseder S, Oster J, Clifford GD. An Open-source framework for stress-testing non-invasive foetal ECG extraction algorithms. Physiol Meas 2016; 37(5):627–648.

[14] Schmitz B, *et al*. Patient-centered cardiac rehabilitation by AI-powered lifestyle intervention – the timely approach. Atherosclerosis 2022;355:251, 1879-1484.

Address for correspondence:

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