Deep Learning for Amplified P-Wave Duration Annotation

Silvia Becker[1,2], Heiko Lehrmann[2], Amir Jadidi[3], Ajay Krishna[2], Till Keller[2,4], Dirk Westermann[2],
Thomas Arentz[2], Axel Loewe[1], Martin Eichenlaub[2]

[1] Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2] Department of Cardiology and Angiology, University of Freiburg, Bad Krozingen, Germany
[3] Department of Cardiology, Lucerne Cantonal Hospital, Lucerne, Switzerland
[4] Department of Cardiology, Justus Liebig University Giessen, Giessen, Germany

Abstract

Atrial cardiomyopathy (AtCM) is associated with new-onset atrial fibrillation (AF), higher AF recurrence rates after pulmonary vein isolation (PVI), and increased risk for ischemic stroke. Automated diagnosis of AtCM using electrocardiograms (ECGs) could enable non-invasive screening of large cohorts. The amplified P-wave duration (APWD) holds potential for diagnosing and staging AtCM. In this study, we propose a long short-term memory (LSTM) model to annotate APWD. The model's training involved two phases: initial pretraining with weak labels and subsequent training with expert labels. We investigated the effects of pretraining, trimming input signals, and upsampling on the absolute error between predictions and labels. The best-performing model was a bidirectional LSTM with 16 hidden units using pretraining, no trimming, and upsampling during training, resulting in absolute errors of 13.9 ± 24.9, 15.4 ± 17.4, *and* $18.2 \pm 19.8\,ms$ *for the P-wave onset, offset and duration, respectively. On the independent data set, errors were* 7.3 ± 7.4, 15.6 ± 16.5, *and* $16.5 \pm 21.1\,ms$, *accordingly. The model showed little systematic bias and generalized well to unseen data. In conclusion, this work demonstrates promising results for the automation of AtCM diagnosis, suggesting potential for improved screening efficiency, ultimately enabling improved patient management and outcome.*

1. Introduction

Atrial cardiomyopathy (AtCM) is linked to incident atrial fibrillation (AF), poor outcomes after pulmonary vein isolation (PVI), and ischemic stroke. By analyzing the amplified P-wave duration (APWD) on multi-channel electrocardiograms (ECGs), atrial conduction delay in patients with AtCM can be quantified non-invasively [1–3]. Currently, APWD is annotated manually, which is time-consuming and impractical for large-scale screening. Consequently, there is a need for automated APWD annotation for widespread adoption of this valuable metric. This work aims to develop an algorithm to automatically measure APWD and validate it against manual expert annotations.

2. Methods

Three data sets were utilized in this study: one for pretraining, one for training, and one for validation.

The pretraining data set comprised 129,302 sinus rhythm ECGs extracted from the MIMIC-IV database [4, 5], ensuring inclusion of only one ECG per patient. 214 ECGs from this database were excluded from pretraining for inclusion in the training data set. All ECGs were sampled at 500 Hz. Individuals were 57±20 years old, 53 % were male, and for 1 % sex was unknown.

The training data set included 1,044 ECGs from various patient groups representing different stages of AtCM: 314 ECGs from young individuals aged between 18 and 30 years from the MIMIC data set (214 ECGs, 500 Hz) [4, 5] and the PTB-XL database (100 ECGs, 500 Hz) [5–7], 212 ECGs acquired at 1 kHz from older cardiovascular patients at risk for AF (without diagnosed AF), 415 ECGs with sampling frequency of 500 Hz from patients diagnosed with AF, and 103 ECGs acquired at 500 Hz from AF patients with a left atrial thrombus. The mean age of this cohort was 56±22 years, 58 % were male.

Independent validation data set: This data set contained 60 ECGs from patients who underwent PVI. This cohort had a mean age of 63±10 years and 72 % of individuals were male. All ECGs of the training and validation data set were recorded at the Medical Center of the University of Freiburg, all patients provided written informed consent.

Weak labels for the P-wave on- and offset generated by ECGdeli [8] were used for pretraining. For training, high-quality labels of APWD were provided by an expert. Each data set was split into 70 % training, 10 % validation and 20 % test. For the training data set, cohorts were distributed equally across these splits to ensure balanced training. The independent data set included high-quality annotations, reflecting the consensus of three experts, achieving an inter-observer agreement greater than 0.9.

To evaluate the model's performance, the absolute errors for P-wave onset, offset, and duration were calculated. Moreover, an ablation study analyzing the effect of pretraining, trimming the input template to the assumed relevant region, and upsampling of the input, was conducted.

2.1. Data preprocessing

To eliminate P-wave variances between different heartbeats, a single beat template was created for each lead from each 10 second 12-lead ECG. Template generation consisted of R-peak detection, RR-interval calculation, bandpass (0.1–150 Hz) and notch (50/60 Hz) filtering, exclusion of RR-interval outliers, creation of snippets corresponding to individual heartbeats, PCA-based outlier removal, and per-lead averaging of the remaining snippets. These templates were the input for the models. The annotation process involved assigning a sample-wise classification for the P-wave class: labels were set to one for samples between the annotated P-wave onset and offset and zero elsewhere.

2.2. Network architectures

This study examined long short-term memory (LSTM) models, including both uni- and bidirectional LSTMs, with configurations ranging from 4 to 64 hidden units. The model architectures were structured with the following layers: a masking layer, an LSTM or bidirectional LSTM, followed by dropout (10 %), batch normalization, and a dense layer with a softmax activation function. Dropout and batch normalization were employed to prevent overfitting and to stabilize learning. Categorical cross-entropy loss was used for training with early stopping to avoid overfitting. After the pretraining, all layers except the dense layer were frozen. This layer was then retrained to fine-tune the model's final predictions. Hyperparameter optimization was conducted using the Optuna framework [9] to identify the optimal number of hidden units, optimizer (Adam, RMSprop, SGD), batch size, learning rate, and class weighting.

3. Results

The best-performing architecture was a bidirectional LSTM with the following specifications: 16 hidden units, class weighting of 20, batch size of 18, learning rate of 0.001, and Adam optimizer. This model was trained without trimming the input but incorporated upsampling during training to improve performance.

3.1. Upsampling and template trimming

Upsampling ECGs acquired at 500 Hz to 1 kHz, which constituted approximately 80 % of the training data set, slightly improved the P-wave duration error on the test set by -5.4 ± 0.7 %. However, when upsampling was applied during the pretraining phase, it slightly increased the error by 4.6 ± 1.4 %.

The absolute error on the test set for the P-wave onset, offset, and duration increased by 8.7 ± 13.5, 1.4 ± 5.3, and 4.5 ± 17.8 %, respectively, when trimming the template to the P-wave. Figure 1 depicts an input template, with the sections discarded by trimming highlighted in red. Trimming was conducted at the R-peak because the P-wave should always precede the QRS complex. This targeted trimming aimed to enhance model accuracy by focusing the analysis solely on the P-wave region but increased the error instead.

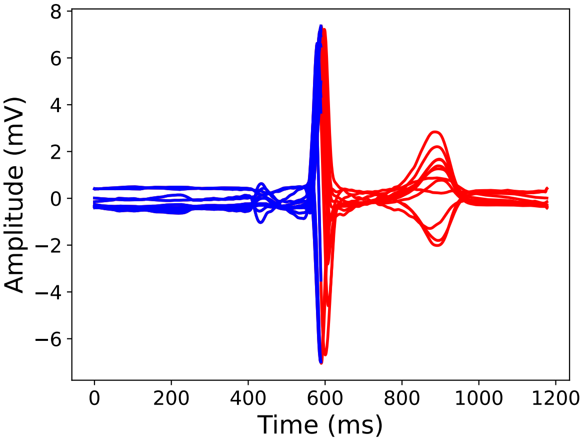

Figure 1. ECG template. This figure shows a 12-lead template that was used as the input for the model. The red parts indicate the discarded parts of the template when trimming the template to the region of interest.

3.2. Performance

Figure 2 illustrates the absolute errors of the final model with and without pretraining on the ECGs withheld from the training data set for testing. Pretraining reduced both the mean absolute error and the standard deviation. The absolute error for P-wave onset, offset, and duration were reduced by -25.3 ± -34.0, -25.2 ± -34.1, and -30.3 ± -37.7 %, respectively.

Bland-Altman analysis identified a bias of 1.48 ms, with a slight increase in error dispersion at higher mean values (Figure 3). Analysis of the model's performance on the

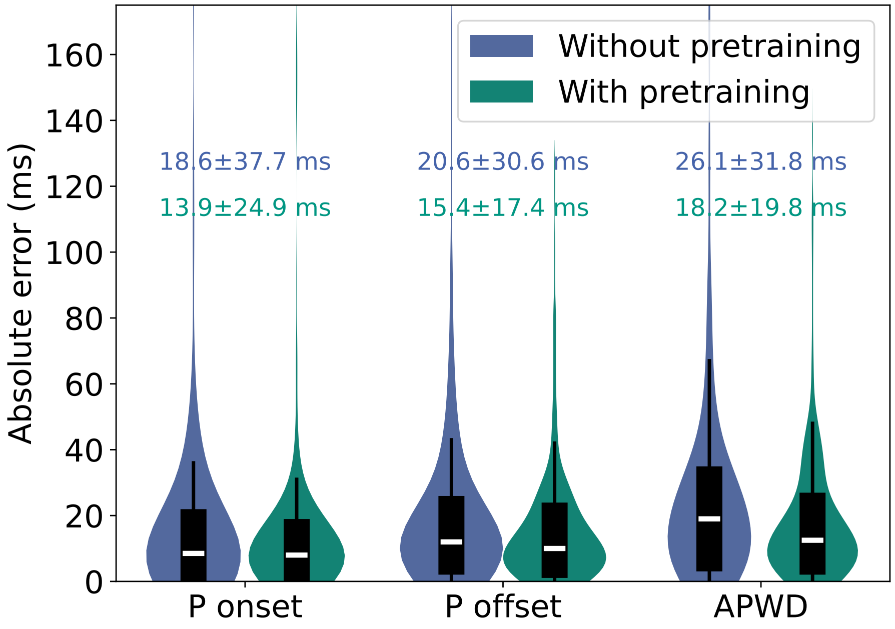

Figure 2. Absolute errors with and without pretraining. Blue violin plots represent errors for the P-wave onset, offset, and APWD without pretraining, while green plots represent errors with pretraining, respectively. Mean and standard deviations are indicated above each plot.

different subgroups in the test set revealed the results depicted in Table 1. The model performed best in the young individuals, followed by the patients with diagnosed AF, and performed worst in the subgroup at risk for AF and in patients with AF and left atrial thrombus.

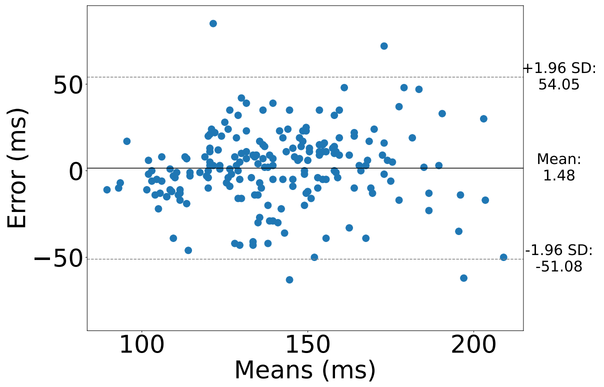

Figure 3. Bland-Altman plot for the test set and the best-performing model. Errors between expert annotations and model predictions for P-wave measurements are compared. The plot illustrates the relationship between measurement differences and their mean values.

To evaluate the model's generalization to unseen data, performance on the independent validation data set was analyzed. The absolute errors for P-wave on-, offset, and duration were 7.3 ± 7.4, 15.6 ± 16.5, and 16.5 ± 21.1 ms, respectively, see Figure 4. Notably, all errors on this data set were lower than on the test set.

	Onset	Offset	APWD
Young individuals	7.4 ± 7.1	10.8 ± 12.3	12.1 ± 11.9
At risk for AF	15.3 ± 24.6	13.6 ± 11.8	23.4 ± 25.5
AF diagnosis	12.0 ± 12.4	17.7 ± 15.9	18.9 ± 17.9
AF with left atrial thrombus	37.5 ± 59.1	23.9 ± 33.4	23.3 ± 26.3

Table 1. Absolute error in ms per subgroup

4. Discussion

The effect of trimming the input templates resulted in increased errors, suggesting a negative impact on model performance. Although trimming shortens the template and might prevent large annotation errors, such as annotation of P-wave offset within QRS complex or T-wave, the overall performance decreased. This indicates that the model benefits from the temporal information provided by the QRS complex and T-wave in relation to the P-wave. LSTMs are specifically designed to account for temporal

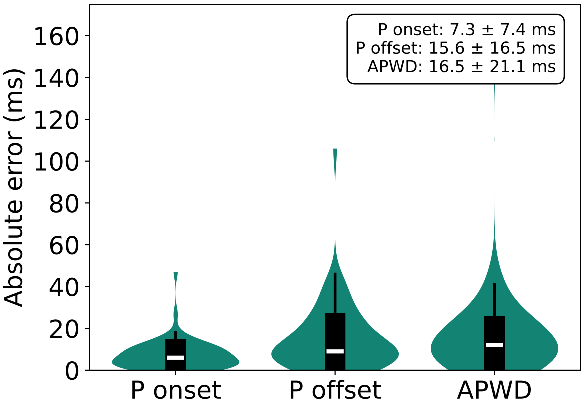

Figure 4. Absolute errors on the independent validation data set for P-wave onset, offset, and duration. Means and standard deviations are given in the legend.

relationships within data, and these relationships are better utilized when the entire template is included as input. Furthermore, the superior performance of the bidirectional LSTM compared to the unidirectional version highlights the value of capturing temporal dependencies in both directions for this classification task.

Upsampling all ECGs to 1 kHz had minimal impact on model performance, indicating that the model is not particularly sensitive to the temporal resolution of the input data. While upsampling was hypothesized to potentially enhance model accuracy by providing more detailed data, these results suggest that such increased resolution does not significantly affect performance during training. Moreover, the complexity introduced by upsampling during pretraining decreased the model's performance, possibly due to difficulty in parameter convergence.

The results of the Bland-Altman plot suggest that there is only little systematic bias in the model's predictions. However, the slight increase in error dispersion towards larger mean values aligns with the observation that the model's performance was somewhat reduced for subgroups with very prolonged P-waves, such as thrombus patients, compared to those with shorter P-waves, like in the young subgroup. This indicates that while the model performs well generally, improvements are necessary for accurately predicting very long P-waves where differentiation between low-amplitude parts of the P-wave and noise is difficult.

The results from the independent data set surpassed those of the test set, which could be attributed to the more homogeneous nature of the independent data set consisting only of AF patients undergoing PVI. The good results on this data set suggest that the model is capable of generalizing effectively to new data and is not overfitted to the training data set.

5. Conclusion

This study presents a promising approach for the automatic annotation of APWD, facilitating retrospective analysis and the screening of large cohorts. This development has the potential to significantly enhance non-invasive risk stratification of AtCM, ultimately contributing to improved patient management and outcomes.

Acknowledgments

This work was supported by the Leibniz ScienceCampus Digital Transformation of Research (DiTraRe).

References

[1] Jadidi A, Müller-Edenborn B, Chen J, et al. The duration of the amplified sinus-P-wave identifies presence of left atrial low voltage substrate and predicts outcome after pulmonary vein isolation in patients with persistent atrial fibrillation. JACC Clin Electrophysiol 2018;4(4):531–543.

[2] Mueller-Edenborn B, Minners J, Keyl C, et al. Electrocardiographic diagnosis of atrial cardiomyopathy to predict atrial contractile dysfunction, thrombogenesis and adverse cardiovascular outcomes. Sci Rep 2022;12(1):576.

[3] Huang T, Nairn D, Chen J, et al. Structural and electrophysiological determinants of atrial cardiomyopathy identify remodeling discrepancies between paroxysmal and persistent atrial fibrillation. Front Cardiovasc Med 1 2023;9:1–18.

[4] Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV. PhysioNet 2024;URL https://physionet.org/content/mimiciv/.

[5] Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[6] Wagner P, Strodthoff N, Bousseljot RD, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020;7(1):1–15.

[7] Strodthoff N, Mehari T, Nagel C, et al. PTB-XL+, a comprehensive electrocardiographic feature dataset. Sci Data 5 2023;10(1):279–279.

[8] Pilia N, Nagel C, Lenis G, et al. ECGdeli-an open source ECG delineation toolbox for MATLAB. SoftwareX 2021; 13:100639.

[9] Akiba T, Sano S, Yanase T, et al. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019; 2623–2631.

Address for correspondence:

Silvia Becker, publications@ibt.kit.edu
Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), Fritz-Haber-Weg 1, 76131 Karlsruhe, Germany