Test-Time Adaptation for a Generalizable Deep Learning-Based ECG Segmentation

Alaa Salama[1], Amar Kachenoura[1], Salman Almuhammad Alali[1], Guy Carrault[1], Lotfi Senhadji[1], Ahmad Karfoul[1]

[1] Univ Rennes, Inserm, LTSI-UMR 1099, Rennes, France

Abstract

Detecting the onsets and offsets of the P, QRS, and T waves in an ElectroCardioGram (ECG) is crucial for diagnosing cardiac pathologies. Existing ECG segmentation methods can be classified as either unsupervised or supervised. Since artefacts can obscure ECG wave boundaries, unsupervised approaches tend to underdetect ECG waves. In contrast, supervised approaches generally overdetect ECG waves. Another limitation of supervised methods is their lack of generalizability to unseen data. This study presents a new deep learning-based ECG segmentation pipeline that strikes a balance between improving the model's generalization and maintaining performance. More precisely, our pipeline involves two stages: i) a segmentation stage using a trained Multi-Head attention-based convolutional neural network-long short term memory model (ConvLSTM$_{MA}$), and ii) a test time adaptation stage to adapt the testing data to the target during inference. Using two PhysioNet databases (QT for training and LU for testing), a database-independent evaluation strategy was employed to compare our pipeline with one of the best supervised methods (based on self-attention) and two commonly used unsupervised ones (ECGDeli and ECGKit). This study demonstrates test time adaptation's potential to increase the models' generalizability to unseen data while reducing the False Detection Rate (FDR).

1. Introduction

Accurate segmentation of ElectroCardioGram (ECG) waves (P, QRS, and T) is mandatory for identifying cardiac abnormalities [1]. ECG segmentation waves remain challenging because the ECG is subject to noise that can distort its waveforms. This makes ECG wave identification tedious and leads to a high False Detection Rate (FDR) and/or to underdetection. ECG segmentation methods fall into two main categories: unsupervised and supervised. Unsupervised approaches encompass techniques such as Wavelet Transform (WT) [2], Fourier Transform (FT), and Empirical Mode Decomposition (EMD) [3]. Among supervised approaches, Deep Learning (DL) has been extensively applied to the automatic segmentation of ECG signals. For example, Malali et al. [4] proposed a Convolutional LSTM (ConvLSTM) model, which combines the capabilities of Convolutional Neural Networks (CNNs) and bidirectional LSTMs, along with a Self-Attention (SA) layer, leading to the ConvLSTM$_{SA}$ model.

Although these ECG segmentation methods have demonstrated promising performances, they treat the segmentation task as a binary classification problem for each ECG time sample, without considering the total number of segmented waves (P, QRS, and T). Additionally, isolated artifacts are likely to occur in conditions of low Signal to Noise Ratio (SNR). Since artifacts can obscure the onset and offset times of each ECG wave, unsupervised approaches tend to misdetect these critical times and/or underdetect a high number of ECG waves. Supervised approaches, typically DL-based, lack generalizability to unseen data and tend to overdetect ECG waves. This leads to a significant FDR in terms of the number of segmented waves. Indeed, in real-world scenarios, the testing data often comes from a different distribution than the training data, giving rise to the well-known data shift problem. Many solutions have been proposed to improve generalization ability: i) Fine-Tuning (FT) [5] adapts a pretrained model to a new dataset by retraining and updating primarily the high-level parameters, ii) Domain Adaptation (DA) [6] combines supervised loss on the source domain with unsupervised loss to align the source and target distributions, iii) Source Free Domain Adaptation (SFDA) [7] addresses the same issue without requiring access to source data, though it still relies on prior knowledge of the target domain and involves retraining the model, iv) Test Time Training (TTT) [8] shifts adaptation to the inference stage by fine-tuning the model on each target sample, leveraging a Y-shaped architecture that integrates an auxiliary self-supervised task and v) Test Time Adaptation (TTA) [9] operates directly at inference by optimizing an unsupervised objective, such as entropy, to adjust specific parameters (typically batch normalization layers) before gener-

ating predictions. TTA stands out as particularly advantageous over the other methods, since it offers a lightweight and adaptable solution that does not involve retraining, source data access, or architectural changes.

Thus, to cope with the limitations of unsupervised and supervised methods, a new two-stage DL-based ECG segmentation pipeline is proposed. The first stage is a trained Multi-Head (MH) attention-based CNN-LSTM model that deals with the problem of underdetection. The second is a TTA stage that increases the model's generalizability to unseen data. Besides, a supplementary performance criterion is provided for a comprehensive and valuable evaluation of the proposed pipeline, in addition to the traditional one based on standard classification metrics (accuracy, recall, precision, etc.). This new criterion evaluates the number of P, QRS, and T waves that were annotated by the clinician in the tested database against those detected by the proposed method. Finally, a database-independent evaluation strategy, using two databases from PhysioNet (QT [10] for training and LU [11] for testing) was exploited to compare our pipeline with one of the best supervised method ($\text{ConvLSTM}_{\text{SA}}$ [4]) and two commonly used unsupervised ones (ECGDeli [12] and ECGKit [13]).

2. Methodology

2.1. The $\text{ConvLSTM}_{\text{MA}}$-based pipeline

Data preprocessing: In this step, new features are extracted from each segment of the ECG signal. More precisely, as stated in the work of [4], a cubic smoothing technique is employed to smooth each ECG segment. The first derivative of the ECG segments is also calculated, providing information on the rate of change of the ECG waveform. Subsequently, absolute value normalization is applied to each ECG segment in order to address the effect of amplitude variation.

$\text{ConvLSTM}_{\text{MA}}$ model architecture: The proposed model consists of several key components, as shown in Figure 1. The input layer receives four distinct signals. The first and second inputs represent the first and second ECG leads. The third and fourth inputs correspond to the smoothed version and the first derivative of the first lead, respectively. Subsequent to the input layer, a convolutional layer comprising 64 filters of size 3 is used to enhance the model's ability to extract local features related to the onset and offset of the P, QRS, and T waves. Then, a bidirectional LSTM layer with 128 units is used to capture dependencies in both the past and future ECG sequences. In order to focus more intently on the critical parts of ECG signals, a multihead attention layer with three heads and a key length of 128 is utilized. Furthermore, a 128-cell bidirectional LSTM layer is incorporated. This layer employs the weights from the multihead attention layer to enhance the representation of the ECG data. Finally, the output layer consists of three channels, each corresponding to one of the three possible classes, namely P, QRS, and T waves.

2.2. Test-time adaptation

As pointed out previously, TTA addresses the data shift problem by adapting a trained model to unseen data during testing. Therefore, it doesn't need source data or any prior knowledge of that data. During the inference phase, TTA solely requires the target data to adapt the DL model before making a prediction. In the study, following the algorithm in [9], the batch normalization parameters of a DL model were adapted to each new test sample by minimizing the entropy of the predictions. Indeed, lower entropy indicates more confident predictions for each ECG wave (P, QRS, and T), while higher entropy correlates with the inability of the model to detect the wave. During inference, before predicting the P, QRS, and T waves in ECG, the model computes the gradient of the entropy loss with respect to the batch normalization affine parameters, namely the scale and shift. Backpropagation is used to calculate the derivatives of the loss for both parameters. These derivatives, termed gradients, are used to update the parameters in order to improve the model's ability to accurately predict each ECG wave. It is important to note that the optimization process performed with a single sample results in a trivial solution, simply assigning the prediction to the most probable class. Consequently, the adaptation was executed in batches, with a batch size set at 11.

2.3. Dataset and training strategy

Whatever the experiment in this study, a non-database-specific strategy is adopted. In other words, the studied pipeline models were trained using the QT database [10], and tested on the LU database [11]. Recall in brief that the QT database contains 105 recordings of 15 minutes in duration, with a sampling rate of 250 Hz. It comprises a wide variety of P, QRS, and T morphologies with real-world variability. Regarding the LU database, it consists of short signals of 10 seconds from 200 subjects, with 12-lead recordings sampled at 500 Hz. Both databases were annotated by cardiologists who marked the onset, peaks, and offset of P, QRS, and T waves. To ensure a fair comparison between supervised methods ($\text{ConvLSTM}_{\text{MA}}$ and $\text{ConvLSTM}_{\text{SA}}$) and unsupervised ones (ECGdeli and ECGkit), these latter were also evaluated exclusively on the LU database.

3. Experiments and results

Many current studies on ECG segmentation evaluate performance using standard classification metrics, i.e., Ac-

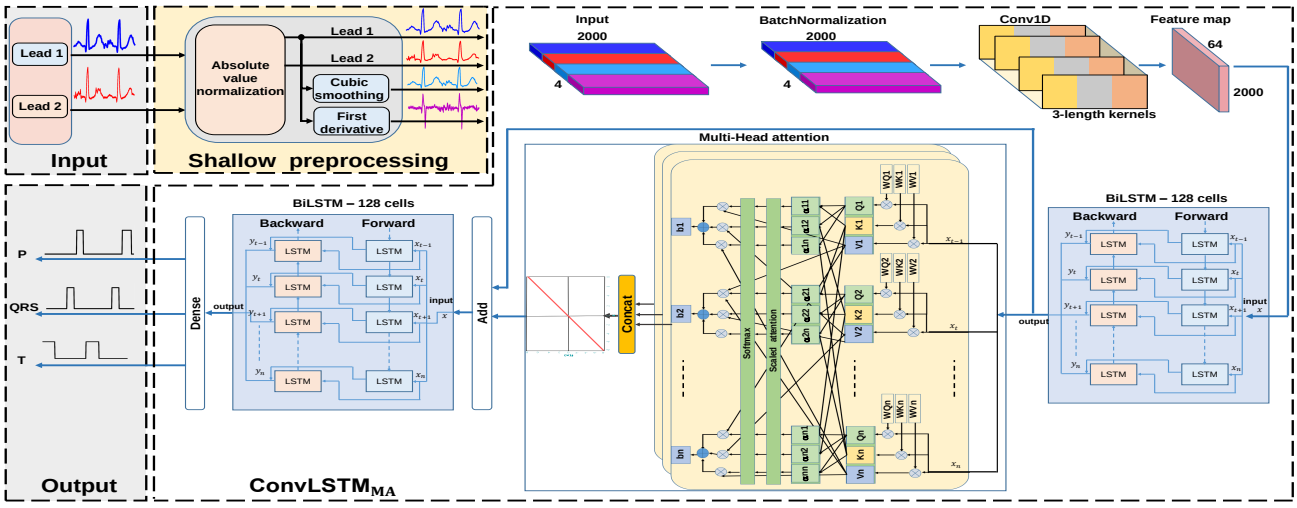

Figure 1. Proposed ConvLSTM$_{MA}$-based pipeline for ECG waves segmentation.



Figure 2. Test Time Adaptation strategy.

curacy, Recall, Precision, and F1-score, which are collectively referred to as Criterion 1 in this study. These metrics use binary classification to evaluate the ability to detect the onset and offset of P, QRS, and T waves. However, they are not accurate for determining the number of waves incorrectly detected. To address this issue, a novel criterion (Criterion 2) is hereby proposed. It assesses the FDR rate of each wave and each method that has been tested. It is given by: $FDR = \frac{(n_p - n_t)*100}{n_t}$, where n_p denotes the number of waves detected by the method, and n_t denotes the number of waves annotated by experts. As stated before, we evaluated both supervised and unsupervised pipelines using the LU dataset.

The results for each pipeline and wave type according to Criteria 1 and 2 are presented in Table 1. According to the classification criterion (Criterion 1), the supervised methods, ConvLSTM$_{MA}$ and ConvLSTM$_{SA}$, achieved higher segmentation quality than the unsupervised methods ECGDeli and ECGKit. This demonstrates the effectiveness of supervised learning in improving ECG signal segmentation. Specifically, if we focus on the F1-score metric (calculated as the harmonic average of recall and precision), the unsupervised methods ECGDeli and ECGKit show an approximate F1-score of 79% and 75% for the P wave, 80% and 69% for the QRS wave, and 77% for the T wave. The supervised methods achieved F1-scores ranging from 87% to 88% for the P wave, from 86% to 91% for the QRS wave, and from 87% to 90% for the T wave. The supervised methods, ConvLSTM$_{SA}$ and ConvLSTM$_{MA}$ seem similar. Note also that TTA improves performance slightly, whatever the segmented wave. In terms of the number of waves detected (criterion 2), the DL-based methods, ConvLSTM$_{SA}$ and ConvLSTM$_{MA}$, show overdetection of the P, QRS, and T waves. Compared with the unsupervised method, ECGDeli, this overdetection is more pronounced. The unsupervised method, ECGKit, suffers from underdetection of the P and T waves. Note that, ConvLSTM$_{MA}$ consistently shows less overdetection than ConvLSTM$_{SA}$. Interestingly, the TTA largely mitigates the overdetection of ConvLSTM$_{MA}$, TTA-ConvLSTM$_{MA}$ with FDR of 3.46%, 8.64%, and 6.50% for the P, QRS, and T waves, respectively. Overall, balancing both segmentation quality and error rates, the TTA-ConvLSTM$_{MA}$ pipeline demonstrates superior performance over addressed methods for all ECG wave types.

4. Conclusion

In this study, a new three-stage pipeline based on DL has been proposed for automatic ECG segmentation. This model comprises a preprocessing step followed by a segmentation step using the proposed ConvLSTM$_{MA}$ model, and a TTA stage to overcome the problem of false positives. The results obtained have shown that the

Table 1. Criteria 1 and 2 of the five tested methods.

Criteria	Metric	Model	P	QRS	T
Criterion 1	Accuracy	ECGDeli	0.9567	0.9634	0.9231
		ECGkit	0.9558	0.9405	0.9343
		$ConvLSTM_{SA}$	0.9798	0.9766	0.9596
		$ConvLSTM_{MA}$	0.9802	0.9815	0.9625
		$TTA\text{-}ConvLSTM_{MA}$	0.9809	0.9852	0.9685
	Recall	ECGDeli	0.9630	0.9542	0.8429
		ECGkit	0.7842	0.8259	0.7372
		$ConvLSTM_{SA}$	0.8251	0.9100	0.8826
		$ConvLSTM_{MA}$	0.8113	0.8714	0.8610
		$TTA\text{-}ConvLSTM_{MA}$	0.8489	0.9064	0.9117
	Precision	ECGDeli	0.6731	0.6976	0.7115
		ECGkit	0.7236	0.5909	0.8183
		$ConvLSTM_{SA}$	0.9321	0.8200	0.8603
		$ConvLSTM_{MA}$	0.9506	0.8938	0.8921
		$TTA\text{-}ConvLSTM_{MA}$	0.9221	0.9084	0.8868
	F1-score	ECGDeli	0.7924	0.806	0.7716
		ECGkit	0.7527	0.6889	0.7756
		$ConvLSTM_{SA}$	0.8753	0.8618	0.8710
		$ConvLSTM_{MA}$	0.8755	0.8825	0.8763
		$TTA\text{-}ConvLSTM_{MA}$	0.8840	0.9074	0.8991
Criterion 2	Number of Waves	Expert Annotation	1211	1215	1215
		ECGDeli	1241	1243	1227
		ECGKit	1167	1370	995
		$ConvLSTM_{SA}$	1420	1544	1580
		$ConvLSTM_{MA}$	1267	1384	1331
		$TTA\text{-}ConvLSTM_{MA}$	1253	1320	1294
	FDR [%]	ECGDeli	2.47	2.30	0.98
		ECGKit	-3.63	12.75	-18.10
		$ConvLSTM_{SA}$	17.25	27.07	30.04
		$ConvLSTM_{MA}$	4.62	13.90	9.54
		$TTA\text{-}ConvLSTM_{MA}$	3.46	8.64	6.50

$ConvLSTM_{MA}$ model offers better performance compared to ECGDeli, ECGKit, and $ConvLSTM_{SA}$ methods. In addition, the proposed approach emphasizes the usefulness of the multi-head attention mechanism for assigning different levels of importance to ECG signal features. Furthermore, the obtained results demonstrated the limitations of relying solely on classical classification metrics, as this can result in biased conclusions. This is particularly true for DL-based models, which generally suffer from a non-negligible rate of overdetection. Consequently, incorporating a supplementary evaluation criterion, such as the number of segmented waves, is imperative for ensuring a more robust evaluation process. The findings of the study show that the use of a TTA step results in a substantial reduction in the FDR. Subsequent endeavours will centre on the evaluation of the proposed pipeline on more extensive real-world databases and the exploration of the explainability of the proposed DL-based model.

5. Acknowledgments

This work was supported in part by the PEPR Santé Numérique; Projet DIIP-HEART: ANR-22-PESN-0018 and in part by the PrepRisC project of the ARED program of Région Bretagne, France.

References

[1] Goldberger AL, Goldberger ZD, Shvilkin A. Clinical electrocardiography: a simplified approach, expert consult: online and print, 8: clinical electrocardiography: A Simplified approach. Elsevier Health Sciences, 2012.

[2] Lenis G, Pilia N, Oesterlein T, Luik A, Schmitt C, Dössel O. P wave detection and delineation in the ecg based on the phase free stationary wavelet transform and using intracardiac atrial electrograms as reference. Biomedical Engineering Biomedizinische Technik 2016;61(1):37–56.

[3] Abderahman HN, Dajani HR, Groza VZ. Adaptive r-peak detector in extreme noise using emd selective analyzer. In 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA). IEEE, 2022; 1–6.

[4] Malali A, Hiriyannaiah S, Siddesh G, Srinivasa K, Sanjay N. Supervised ECG wave segmentation using convolutional LSTM. ICT Express 2020;6(3):166–169.

[5] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Advances in Neural Information Processing Systems 2014;27.

[6] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision. 2017; 2223–2232.

[7] Bateson M, Kervadec H, Dolz J, Lombaert H, Ayed IB. Source-free domain adaptation for image segmentation. Medical Image Analysis 2022;82:102617.

[8] Sun Y, Wang X, Liu Z, Miller J, Efros A, Hardt M. Test-time training with self-supervision for generalization under distribution shifts. In International Conference on Machine Learning. PMLR, 2020; 9229–9248.

[9] Wang D, Shelhamer E, Liu S, Olshausen B, Darrell T. Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv200610726 2020;.

[10] Laguna P, Mark RG, Goldberger AL, Moody GB. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ECG. Computers in Cardiology 1997;24:673–676.

[11] Kalyakulina AI, Yusipov II, Moskalenko VA, Nikolskiy AV, Kosonogov KA, Osipov GV, Zolotykh NY, Ivanchenko MV. Ludb: a new open-access validation tool for electrocardiogram delineation algorithms. IEEE Access 2020;8:186181–186190.

[12] Pilia N, Nagel C, Lenis G, Becker S, Dössel O, Loewe A. Ecgdeli-an open source ECG delineation toolbox for matlab. SoftwareX 2021;13:100639.

[13] Demski A, Llamedo MS. ECG-kit a matlab toolbox for cardiovascular signal processing. Journal of Open Research Software 2016;4:e215–e220.

Address for correspondence:

Ahmad Karfoul

LTSI-Inserm UMR 1099, Batiment 22, Campus de Beaulieu, Universite de Rennes, 35042 Rennes, France

ahmad.karfoul@univ-rennes.fr