Mitigating False Positives in ECG Segmentation Through Deep Learning-based Postprocessing

Alaa Salama
Univ. Rennes, Inserm, LTSI-UMR 1099
F-35000 Rennes, France
alaa.salama@univ-rennes.fr

Amar Kachenoura
Univ. Rennes, Inserm, LTSI-UMR 1099
F-35000 Rennes, France
amar.kachenoura@univ-rennes.fr

Salman Almuhammad Alali
Univ. Rennes, Inserm, LTSI-UMR 1099
F-35000 Rennes, France
salman.almuhammad-alali@univ-rennes.fr

Guy Carrault
Univ. Rennes, Inserm, LTSI-UMR 1099
F-35000 Rennes, France
guy.carrault@univ-rennes.fr

Lotfi Senhadji
Univ. Rennes, Inserm, LTSI-UMR 1099
F-35000 Rennes, France
lotfi.senhadji@univ-rennes.fr

Ahmad Karfoul
Univ. Rennes, Inserm, LTSI-UMR 1099
F-35000 Rennes, France
ahmad.karfoul@univ-rennes.fr

***Abstract*—The characteristics of P, QRS, and T ECG waves, as well as the intervals and segments between them, are crucial for the diagnosis of cardiac diseases. However, automatic ECG signal segmentation methods suffer from the False Positives (FP) issue caused by either those isolated artifacts that are mistakenly detected as events of interest or artifacts that hide target events of interest. In this paper, a shallow Deep Learning (DL)-based postprocessing stage is proposed to mitigate FP in ECG segmentation. The effectiveness of the proposed postprocessing stage is evaluated by applying it to a recent supervised approach, namely ConvLSTM, and two well-known unsupervised methods, namely ECGdeli and ECGKit, using the popular QT and LU databases of PhysioNet. The results obtained confirm the effectiveness of the proposed postprocessing in reducing the False Detection Rate (FDR). In addition, for a comprehensive analysis of segmentation quality, an additional evaluation criterion addressing the number of false detected waves is also suggested. The current study demonstrates the relevance of combining ECG segmentation approaches, especially supervised ones, with the proposed shallow DL-based postprocessing stage, in terms of both the accuracy and the number of detected waves.**

***Index Terms*—Electrocardiography; ECG segmentation; False detection rate**

I. INTRODUCTION

Accurate segmentation of different ECG waves (P, QRS and T) and wave intervals and segments (PR, QT, ST, etc) is mandatory; Since these features are commonly used by clinicians to determine cardiac abnormalities and are crucial for diagnosing cardiac pathologies or predicting adverse events [1]. Although many methods have been proposed to segment the ECG signal, this task remains challenging as the ECG signal is subject to various noise types that can drastically distort its waveform, making it difficult to identify these ECG waves and hence leading to a high FP segmentation rate. ECG segmentation methods can be classified into two main categories: unsupervised and supervised. Unsupervised approaches encompass techniques such as the Pan-Tompkins algorithm [2], Wavelet Transform (WT) [3], [4], Fourier Transform (FT) [5], and Empirical Mode Decomposition (EMD) [6], among others. Among supervised approaches, Deep Learning (DL) has been extensively applied to the automatic segmentation of ECG signals. For example, in [7], the authors introduced ECG-SegNet, a novel segmentation method based on two layers of bidirectional Long Short-Term Memory (LSTM). Malali et al. [8] 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 $\text{ConvLSTM}_{\text{SA}}$ model. In this study, the authors compared the latter model with the Hidden Markov Model (HMM) [9] and ECG-SegNet [7], demonstrating that $\text{ConvLSTM}_{\text{SA}}$ achieved superior performance.

Although these ECG segmentation methods have demonstrated promising accuracy, they consider the segmentation task as a binary classification problem for each ECG time sample, without accounting for the total number of segmented waves (P, QRS, and T). As a result, high classification accuracy can be achieved at the time sample level, while a high FDR is expected for the number of these segmented waves. Furthermore, in low SNR conditions, isolated artifacts are likely to occur. These artifacts can obscure the onset and offset of each ECG wave, leading unsupervised approaches to generally underdetect these critical time points accurately. Regarding supervised approaches, typical DL-based ones, they generally tend to overdetect ECG waves. To address the aforementioned issues, this paper makes two key contributions: (i) the introduction of a shallow DL-based postprocessing stage that combines CNNs and BiLSTM models, which can be integrated with both unsupervised and supervised ECG segmentation methods, and (ii) the suggestion to consider the number of detected waves as an additional evaluation criterion alongside segmentation accuracy, ensuring a more reliable assessment of segmentation quality. In this study, the impact of the proposed postprocessing on the ECG segmentation of two classical unsupervised approaches, namely ECGdeli [10]

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

and ECGKit [11], and an efficient supervised ConvLSTM$_{SA}$ [8] is evaluated using two open source databases QTDB [12] and LUDB [13], employing a non-database specific strategy.

II. Methodology

Assume at this stage that the ECG segmentation is performed using either an unsupervised or a supervised method. This segmentation result will then be introduced to the subsequent proposed DL-based postprocessing stage to reduce FDR as previously mentioned (see Fig. 1). Two key points are addressed in this section: (i) the strategy being adopted to train both the segmentation and the proposed DL-based postprocessing models, and (ii) the evaluation methodology considered for a comprehensive and full ECG segmentation analysis.

A. Dataset and training strategy & postprocessing model

Two databases from PhysioNet were exploited here to train and evaluate the studied pipelines. Annotations indicating the onsets, peaks, and offsets of the P, QRS, and T waves in each recording (one record per patient) in both databases were provided. The first database, namely QTDB [12], comprises 105 recordings of 15 minutes each, with two ECG leads. The sampling frequency is 250 Hz. The second database, the Lobachevsky University Electrocardiography Database (LUDB) [13], comprises 200 10-second, 12-lead ECG recordings, with a sampling frequency equal to 500 Hz.

It is important to note that, in this study, a non-database specific strategy was adopted to ensure the generalization of the proposed pipeline: all supervised processes were trained using the QTDB and tested on LUDB (see Fig 1). More precisely, each of the 105 ECG QTDB recordings has been split into segments of 2000 samples (8 seconds). After that, 70 % of the obtained segments were used to train the DL-based segmentation model, ConvLSTM$_{SA}$, while the remaining 30 % were used as the training set for the postprocessing DL model. Regarding the testing set (LUDB recordings), a 5-fold cross-validation technique was applied to test the DL-pipeline, namely segmentation + postprocessing (see testing stage, Fig 1). As far as the unsupervised methods (ECGdeli and ECGkit) were concerned, they were evaluated only on the LUDB to make results comparable.

The DL-based postprocessing model involves several components to extract valuable features from the segmented ECG waves, as shown in Figure 2. The input layer receives the three outputs (P, QRS, T) of the segmentation step. This input layer is followed by a 1D-CNN layer of 64 filters with a kernel size of 12 to extract local features. Then, a BiLSTM layer [14] with 64 cells is used to capture temporal dependencies. Finally, a dense layer is used to deliver three corrected ECG segments.

B. Evaluation criteria

As previously mentioned, a lot of existing ECG segmentation studies assess the performance of their methods using standard classification metrics (i.e., Accuracy, Recall, Precision, Specificity, and F1-score). The latter metrics are computed for each wave through a binary classification of the ECG samples. However, note that if these metrics are far from acceptable, the segmentation methods used are classified as inefficient. Importantly, achieving good segmentation performance based on these classification metrics does not necessarily ensure the effectiveness of the method used. In fact, these metrics are primarily designed to assess the accuracy of detecting the onset and offset of P, QRS, and T waves, but they are not robust for evaluating the number of incorrectly detected waves. To cope with this issue, we propose to consider the following complementary criteria.

Criterion 1 assesses standard metrics of classification, namely the F1 score, the accuracy, the recall, and the precision.

Criterion 2 assesses the FDR rate of each wave (P, QRS complex, and T) and each evaluated method, given by:

$$FDR = \frac{(n_p - n_t) * 100}{n_t} \tag{1}$$

where n_p denotes the number of waves detected by the tested method, and n_t denotes the number of waves annotated in the database.

Criterion 3 calculates the mean and the standard deviation (in [ms]) of the difference between the onset and offset of P, QRS and T waves estimated by the pipeline compared to their corresponding clinical annotations.

III. Experiments and results

For a comparative evaluation, the results obtained using Criteria 1, 2, and 3 are given, for each method and each wave, in Tables I and II, with and without proposed postprocessing, respectively.

Criterion 1 - Table I shows that the supervised method, namely ConvLSTM$_{SA}$, clearly outperforms the unsupervised methods, whatever is the calculated metric and the segmented wave (P, QRS, and T). According to Criterion 1, we can conclude that the postprocessing stage is unnecessary. However, when evaluating the methods using the number of detected waves and FDR (Criterion 2 - Table I), we notice that the supervised method has a high FDR, which means that it suffers from the overdetection of P, QRS, and T waves, while the ECGKit method tends to underdetection of P and T waves. Now, if we take a look at Criterion 3 - Table I, clearly ECGDeli and ECGKit are less effective at precisely detect the onset and offset of the segmented waves, especially the P and T waves, which present a weak SNR in comparison to QRS.

Applying a postprocessing stage (see Table II), clearly benefits the supervised method. We can see that after applying the postprocessing stage, ConvLSTM$_{SA}$ is still the best in terms of Criterion 1. More importantly, this postprocessing stage deals with overdetection that the supervised method suffers from (Criterion 2-Table II), as it provides a very limited FDR of -0.24%, 3.29%, and 4.85% for the P, QRS, and T waves, respectively. Additionally, the supervised method is still the best in terms of detecting the onset and the offset of all waves (Criterion 3 - Table II).

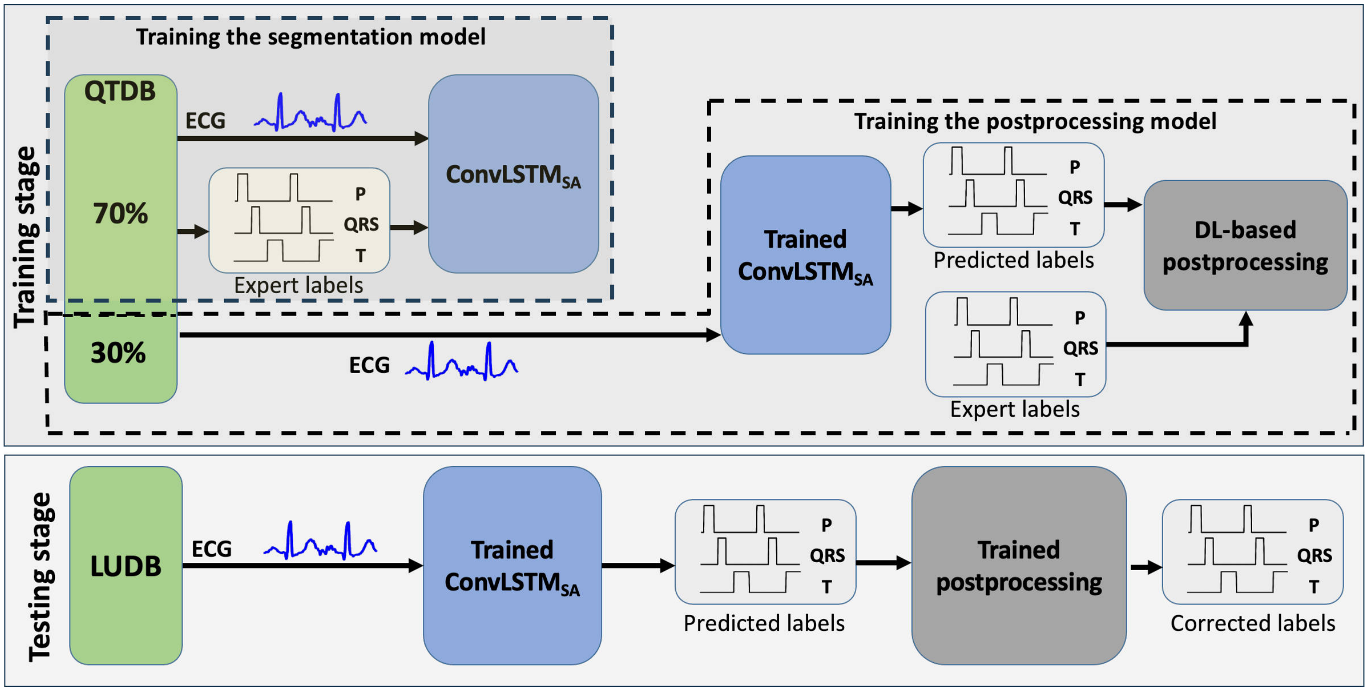


Fig. 1. Proposed ECG wave segmentation and postprocessing pipeline.

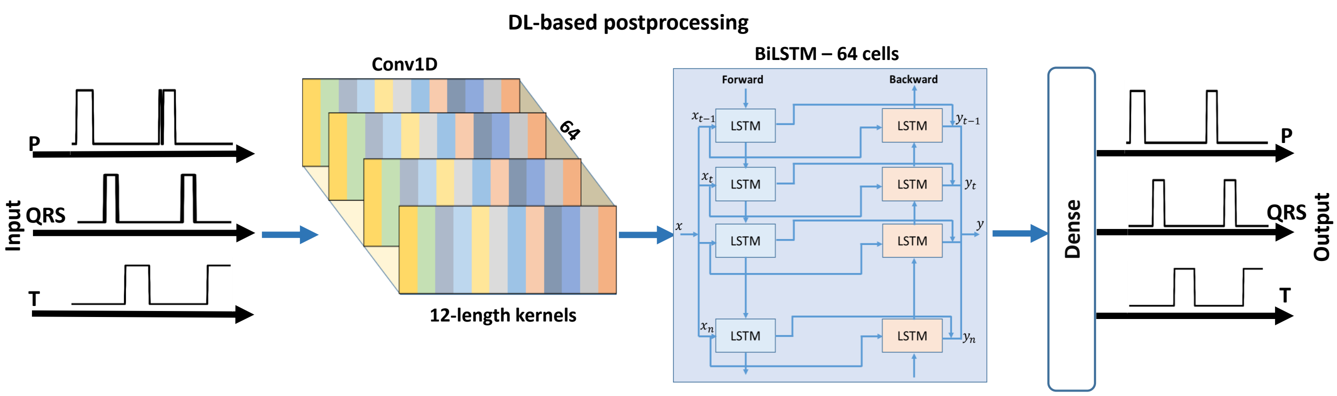


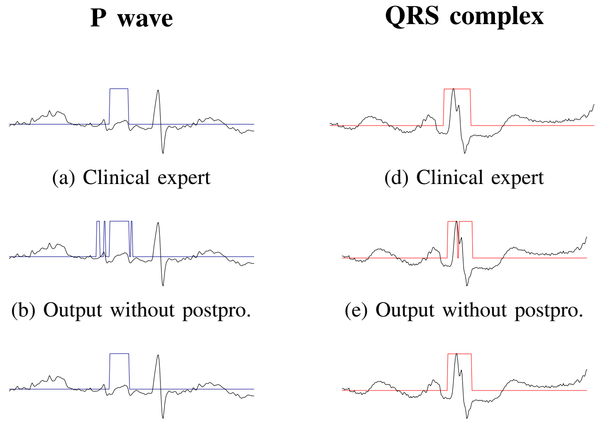
Fig. 2. Proposed model for DL-based postprocessing.

IV. Discussion

Reliable ECG segmentation is essential for diagnosing various cardiac diseases. In this study, a deep learning-based postprocessing method has been proposed, mitigating the FDR of the segmented ECG waves. The preliminary results showed that the DL-based method outperformed the standard unsupervised one in terms of classical statistical metrics (Criterion 1). We also demonstrated that while classical unsupervised methods suffer from the underdetection issue, the DL-based ones tend to overdetect the number of P, QR, and T waves, leading in turn to a high FDR. This emphasizes that evaluating ECG wave segmentation algorithms using only classical statistical metrics, as commonly practiced, is not fair and may lead to erroneous results. More precisely, low values of metrics calculated in Criterion 1 will indicate that the used method is inefficient. High performance in terms of these metrics does not necessarily guarantee that the method performs well overall. Therefore, in this study, we also propose evaluating the performance of the pipelines using Criteria 2 and 3 to ensure a comprehensive and reliable evaluation. The use of such criteria demonstrated that a postprocessing stage is mandatory to reduce the FDR, regardless of the approach used. Finally, we also confirmed that the ConvLSTM$_{SA}$ method, supported by a shallow DL-based postprocessing stage, outperformed the unsupervised methods, namely ECGDeli and ECGKit, and this for all segmented waves and the evaluation criteria considered.

V. Conclusion

In this study, a new DL-based postprocessing model designed to deal with the high FDR inherent to DL-based ECG segmentation methods is proposed. The results obtained showed



(a) Clinical expert (d) Clinical expert

(b) Output without postpro. (e) Output without postpro.

(c) Output with postpro. (f) Output with postpro.

Fig. 3. Example of P and QRS segmentation: (a), (d) expert annotation; (b), (e) without the proposed postprocessing (postopro.) stage; (c), (f) with the proposed postprocessing stage.

TABLE I
CRITERIA 1, 2, AND 3 OF THE THREE TESTED METHODS WITHOUT ANY POSTPROCESSING.

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
	Recall	ECGDeli	0.9630	0.9542	0.8429
		ECGkit	0.7842	0.8259	0.7372
		$ConvLSTM_{SA}$	0.8251	0.9100	0.8826
	Precision	ECGDeli	0.6731	0.6976	0.7115
		ECGkit	0.7236	0.5909	0.8183
		$ConvLSTM_{SA}$	0.9321	0.8200	0.8603
	F1-score	ECGDeli	0.7924	0.806	0.7716
		ECGkit	0.7527	0.6889	0.7756
		$ConvLSTM_{SA}$	0.8753	0.8618	0.8710
Criterion 2	Number of Waves	Expert Annotation	1211	1215	1215
		ECGDeli	1241	1243	1227
		ECGKit	1167	1370	995
		$ConvLSTM_{SA}$	1420	1544	1580
	FDR[%]	ECGDeli	2.47	2.30	0.98
		ECGKit	-3.63	12.75	-18.10
		$ConvLSTM_{SA}$	17.25	27.07	30.04
Criterion 3	Onset $m \pm \sigma$ [ms]	ECGDeli	13.70 ± 16.46	7.61 ± 8.16	21.48 ± 21.65
		ECGKit	14.15 ± 28.78	5.73 ± 10.55	15.70 ± 28.97
		$ConvLSTM_{SA}$	6.69 ± 8.74	5.85 ± 6.85	11.08 ± 13.04
	End $m \pm \sigma$ [ms]	ECGDeli	10.87 ± 10.63	11.76 ± 11.56	22.91 ± 30.58
		ECGKit	8.95 ± 24.78	7.33 ± 19.29	14.56 ± 28.82
		$ConvLSTM_{SA}$	4.75 ± 7.63	5.68 ± 5.85	8.98 ± 14.04

that the DL-based pipeline (DL-based segmentation + DL-based postprocessing) outperforms both the supervised and unsupervised methods, in terms of classical classification metrics, the FDR of segmented waves, and the mean error related to the onset and offset of P, QRS complex, and T waves. In addition, this study underlined that evaluating segmentation quality using only standard classification metrics is not sufficient. In fact, relying solely on these measurements can lead to misleading conclusions about the behavior of the method used, and incorporating additional evaluation criteria, such as the FDR of segmented waves, is mandatory to provide a comprehensive assessment.

TABLE II
CRITERIA 1, 2 AND 3 OF THE THREE TESTED METHODS WITH THE PROPOSED POSTPROCESSING STAGE.

Criteria	Metric	Model	P	QRS	T
Criterion 1	Accuracy	ECGDeli	0.9592	0.9629	0.9246
		ECGkit	0.9563	0.9411	0.9333
		$ConvLSTM_{SA}$	0.9798	0.9772	0.9595
	Recall	ECGDeli	0.9606	0.9514	0.8355
		ECGkit	0.7759	0.8235	0.7369
		$ConvLSTM_{SA}$	0.8198	0.9117	0.8826
	Precision	ECGDeli	0.6874	0.6958	0.7201
		ECGkit	0.7312	0.5944	0.8127
		$ConvLSTM_{SA}$	0.9380	0.8243	0.8598
	F1-score	ECGDeli	0.8013	0.8037	0.7735
		ECGkit	0.7529	0.6904	0.7730
		$ConvLSTM_{SA}$	0.8748	0.8651	0.8708
Criterion 2	Number of Waves	Expert Annotation	1211	1215	1215
		ECGDeli	1241	1243	1211
		ECGKit	1166	1354	995
		$ConvLSTM_{SA}$	1208	1255	1274
	FDR [%]	ECGDeli	2.47	2.30	-0.32
		ECGKit	-3.71	11.44	-18.10
		$ConvLSTM_{SA}$	-0.24	3.29	4.85
Criterion 3	Onset $m \pm \sigma$ [ms]	ECGDeli	13.59 ± 16.37	7.33 ± 8.41	21.65 ± 24.51
		ECGKit	14.21 ± 28.84	6.10 ± 12.82	16.63 ± 29.04
		$ConvLSTM_{SA}$	6.55 ± 9.26	6.11 ± 8.17	11.45 ± 14.95
	End $m \pm \sigma$ [ms]	ECGDeli	9.70 ± 10.73	12.26 ± 12.06	21.91 ± 31.64
		ECGKit	9.08 ± 24.77	9.10 ± 25.32	14.33 ± 28.87
		$ConvLSTM_{SA}$	4.63 ± 7.60	6.04 ± 8.47	10.21 ± 17.64

REFERENCES

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

[2] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[3] G. Lenis, N. Pilia, T. Oesterlein, A. Luik, C. Schmitt, and O. Dössel, "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*, vol. 61, no. 1, pp. 37–56, 2016.

[4] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A wavelet-based ecg delineator: evaluation on standard databases," *IEEE Transactions on biomedical engineering*, vol. 51, no. 4, pp. 570–581, 2004.

[5] I. Murthy and U. Niranjan, "Component wave delineation of ecg by filtering in the fourier domain," *Medical and Biological Engineering and Computing*, vol. 30, pp. 169–176, 1992.

[6] H. N. Abderahman, H. R. Dajani, and V. Z. Groza, "Adaptive r-peak detector in extreme noise using emd selective analyzer," in *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6, IEEE, 2022.

[7] H. Abrishami, C. Han, X. Zhou, M. Campbell, and R. Czosek, "Supervised ecg interval segmentation using lstm neural network," in *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, pp. 71–77, The Steering Committee of The World Congress in Computer Science, Computer …, 2018.

[8] A. Malali, S. Hiriyannaiah, G. Siddesh, K. Srinivasa, and N. Sanjay, "Supervised ECG wave segmentation using convolutional LSTM," *ICT express*, vol. 6, no. 3, pp. 166–169, 2020.

[9] N. Hughes, L. Tarassenko, and S. J. Roberts, "Markov models for automated ecg interval analysis," *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[10] N. Pilia, C. Nagel, G. Lenis, S. Becker, O. Dössel, and A. Loewe, "Ecgdeli-an open source ECG delineation toolbox for matlab," *SoftwareX*, vol. 13, p. 100639, 2021.

[11] A. Demski and M. S. Llamedo, “ECG-kit a matlab toolbox for cardiovascular signal processing,” *Journal of Open Research Software*, vol. 4, pp. e215–e220, 2016.

[12] P. Laguna, R. G. Mark, A. L. Goldberger, and G. B. Moody, “A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ECG,” *Computers in Cardiology*, vol. 24, pp. 673–676, 1997.

[13] A. I. Kalyakulina, I. I. Yusipov, V. A. Moskalenko, A. V. Nikolskiy, K. A. Kosonogov, G. V. Osipov, N. Y. Zolotykh, and M. V. Ivanchenko, “Ludb: a new open-access validation tool for electrocardiogram delineation algorithms,” *IEEE access*, vol. 8, pp. 186181–186190, 2020.

[14] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.