MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230960

Towards Accurate Search for Neonatal Heartbeat: Weighted Algorithm for Reliable ECG Analysis of Premature Infants

Jessica RAHMAN[a,1], Aida BRANKOVIC[a], Mark TRACY[a,b,c], Robert HALLIDAY[b,c] and Sankalp KHANNA[a]
[a] *CSIRO Australian e-Health Research Centre, Australia*
[b] *Neonatal Intensive Care Unit, Westmead, Sydney, Australia*
[c] *University of Sydney, Australia*
ORCiD ID: Jessica Rahman https://orcid.org/0000-0001-9930-241X, Aida Brankovic https://orcid.org/0000-0001-7978-575X, Mark Tracy https://orcid.org/0000-0001-5648-468X, Sankalp Khanna https://orcid.org/0000-0001-6125-8871

Abstract. Accurate identification of the QRS complex is critical to analyse heart rate variability (HRV), which is linked to various adverse outcomes in premature infants. Reliable and accurate extraction of HRV characteristics at a large scale in the neonatal context remains a challenge. In this paper, we investigate the capabilities of 15 state-of-the-art QRS complex detection implementations using two real-world preterm neonatal datasets. As an attempt to improve the accuracy and reliability, we introduce a weighted ensemble-based method as an alternative. Obtained results indicate the superiority of the proposed method over the state of the art on both datasets with an F1-score of 0.966 (95% CI 0.962-0.97) and 0.893 (95% CI 0.892-0.894). This motivates the deployment of ensemble-based methods for any HRV-based analysis to ensure robust and accurate QRS complex detection.

Keywords. Electrocardiogram, premature infant, R wave detection, ensemble model, artificial intelligence

1. Introduction

Extremely preterm infants (born at 28 weeks or less) and very low birth weight (VLBW) babies (<1500g) have a higher risk of death and permanent disabilities. Physiological characteristics derived from Electrocardiogram (ECG), such as heart rate variability (HRV) can provide crucial information relating to adverse outcomes [1]. With the advent of big data and artificial intelligence (AI), sophisticated techniques using HRV markers from large-scale ECG signals are becoming popular in detecting adverse outcomes ahead of a clinical diagnosis [2]. The first step of HRV analysis is the accurate detection of R waves from the QRS complex, which indicates the heartbeat. This is a challenging signal detection process given the common issues of ECG noise related to the physical size of the chest wall of a micro premie. Over the years, many advanced computational techniques have been developed to efficiently detect QRS

[1] Corresponding Author: Jessica Rahman, email: jessica.rahman@csiro.au.

complex from large-scale ECG data. However, the outcomes of these techniques can vary widely depending signal characteristics, collection methods and pre-processing techniques. In a scoping review by Latremouille et al. [1], it was reported that a large number of studies analysing neonatal HRV did not report the techniques and tools used to handle the ECG data and R wave identification. Therefore, reliable and accurate extraction of HRV characteristics at scale in the neonatal context remains a challenge.

To tackle this issue, we evaluate and compare the performance of several state-of-the-art QRS complex detection algorithms on two real-world premature infant ECG datasets. We also introduce and evaluate a weighted ensemble algorithm that leverages the best-performing of these algorithms to deliver consistently superior performance.

2. Methods

2.1. Datasets

Two datasets consisting of ECG waveforms from premature infants were used in this study. The first one was the publicly available Preterm Infant Cardio-Respiratory Signals (PICS) database [3]. The dataset contains ECG signals collected from ten premature infants collected at 250/500 Hz. The total duration of the dataset was ~ 440 hours. R waves from QRS complexes were extracted using a modified Pan-Tompkins algorithm [4]. Then, the peaks were visually inspected by the researchers to remove artifacts and any erroneous peaks to determine the ground truth.

The second dataset contains 16 ECG data snippets collected at the Neonatal Intensive Care Unit (NICU) at Westmead Hospital, Sydney, Australia. The ECG signals are a subset from a de-identified dataset described in Jani et al. [5], collected at a sampling rate of 1000 Hz (Ethics approval number 2021/ETH00824). The total duration of this dataset was ~ 52 hours. R wave detection was done using the HRV 2.0 module of ADInstruments LabChart (Dunedin, New Zealand). These results were then checked by one researcher and a manual process of updating missed peaks and erroneous peaks was conducted. Finally, the updated annotations were reviewed by an expert neonatologist to finalise the ground truth. Both datasets were recorded from bedside patient monitors, using Intellivue MP70 device (Philips Medical Systems).

2.2. Data Pre-processing and QRS Complex Detection Benchmarks

Raw ECG data was pre-processed using a 0.5 high-pass fifth-order Butterworth filter. Then the powerline interference (50 Hz) was removed. In this study, a total of 15 state-of-the-art QRS complex detection methods [6-16] were analysed. Implementations of these algorithms in the Python biosignal processing toolboxes, Biosppy [17] and Neurokit [18], were employed. All algorithms were executed using their default settings to conduct a fair comparison of the methods. An additional step of peak correction was done with a maximum tolerance of ten samples.

2.3. Proposed Weighted Ensemble Method

As an attempt to improve the robustness and accuracy of the individual methods, we employed the theoretical foundations of the Condorcet jury theorem used in ensemble

learning [19] where state-of-the-art methods were potential ensemble members. Considering that some of the methods are likely to perform better than others here we introduced a weighted approach that takes the form of Logistic Regression (LR) with lasso regularization instead of *majority voting*. To ensure that adding more 'voters', i.e. methods, will increase the probability that the majority decision is correct we considered only the QRS detection methods that had individual performance greater than 0.7 in F1-score. Training of this weighted ensemble model was done using two-thirds of the data, and one-third was used in testing. Stratified sampling was used to encounter the imbalance of positive and negative instances in the dataset. The analysis was performed using the Python scikit-learn machine learning library.

2.4. Performance Evaluation and Statistical Analysis

To assess model performance, we used F1-score, the harmonic mean of precision (i.e. positive predictive value) and recall (i.e. sensitivity). F1-score is a better evaluation measure than accuracy for class-imbalanced data and when the false negatives and false positives are important to consider. To assess the robustness of the benchmark methods and the proposed ensemble method, we compared their performance against ground truth (manually annotated peaks) using bootstrap sub-sampling of 20-minute segments. The process was repeated 30 times. Confidence intervals (CIs) were computed with two-sample paired t-tests. All analyses were performed using Python version 3.7.

3. Results

The optimal value of the regularisation term was obtained with the cross-validated grid- search and was C=0.001 for both considered detests. It resulted in the ensemble comprising of 3 methods for PICS and 8 for the Westmead NICU dataset. Table 1 shows the average F1-Score of 30 bootstrap subsamples and their corresponding 95% confidence interval for all 15 methods and the LR model using an ensemble of top-scoring methods.

The results show that the proposed ensemble approach is superior overall, achieving a mean F1-score of 0.966 (95% CI 0.962-0.97) for PICS database, and 0.893 (95% CI 0.892-0.894) for the Westmead NICU database. Considering individual methods, Rodrigues [13] method using the Neurokit implementation resulted in the highest F1-score of 0.965 (95% CI 0.961-0.969) for PICS dataset, while the default method in Neurokit (based on the steepness of the absolute gradient of the signal) achieved the highest F1- score of 0.869 (95% CI 0.868-0.87) for the Westmead NICU database. Obtained results indicate that the state-of-the-art methods are data-sensitive and hence not robust. Though the available evidence shows a lack of robustness, further analysis on a number of different datasets is needed to reach a firm conclusion.

Table 1. Evaluation measures of QRS complex detection methods using two databases.

		PICS Database		**Westmead NICU Database**	
Algorithm	**Implementation**	**Mean F1**	**CI**	**Mean F1**	**CI**
Christov	Biosspy	0.606	[0.592,0.621]	0.77	[0.768,0.771]
Engzee	Biosppy	0.629	[0.625,0.634]	0.83	[0.83,0.831]
Gamboa	Biosppy	0.584	[0.575,0.592]	0.867	[0.866,0.868]

Neurokit (Default)	Neurokit	0.767	[0.761,0.772]	0.869	[0.868,0.87]
Pan-Tompkins	Neurokit	0.29	[0.279,0.301]	0.331	[0.331,0.332]
Martinez	Neurokit	0.25	[0.247,0.253]	0.836	[0.835,0.837]
Christov	Neurokit	0.643	[0.637,0.648]	0.579	[0.574,0.584]
Gamboa	Neurokit	0.584	[0.575,0.592]	0.867	[0.866,0.868]
Elgendi	Neurokit	0.561	[0.558,0.563]	0.829	[0.827,0.830]
Kalidas	Neurokit	0.001	[0.001,0.002]	0.006	[0.0061,0.0063]
Rodrigues	Neurokit	0.965	[0.961,0.969]	0.674	[0.672,0.676]
Zong	Neurokit	0.69	[0.683,0.697]	0.561	[0.557,0.565]
Nabian	Neurokit	0.481	[0.477,0.486]	0.5	[0.5,0.501]
Hamilton	Neurokit	0.327	[0.323,0.332]	0.583	[0.582,0.584]
Promac	Neurokit	0.73	[0.725,0.735]	0.744	[0.743,0.745]
Ensemble LR		**0.966**	[**0.962,0.97**]	**0.893**	[**0.892,0.894**]

4. Discussion

The results provide us with some interesting insights into the different algorithms for detecting QRS complexes. It can be noticed that individual algorithms perform differently depending on the dataset. For example, the algorithm with the best performance on the PICS dataset (Rodrigues) was outperformed by 8 of the 14 other algorithms on the Westmead dataset. Some algorithms performed consistently well or consistently poorly across both datasets. Thus, it is difficult to come to a resolution on which is the best technique to choose for different datasets. The proposed weighted ensemble approach proved to offer superior performance across both datasets proving the efficacy of ensemble approaches in tackling datasets where individual methods demonstrate high variability. However, a firm conclusion cannot be drawn without further robust investigation on a number of different datasets.

Another point to note is that the widely used Pan-Tompkins [4] algorithm, performed very poorly on both datasets. It is possible that in the previous literature, the algorithm was modified significantly to adapt to the neonatal data. However, without further details on the approach, achieving similar results with other datasets is not possible. In addition, the performance using the Kalidas algorithm got poor results for both datasets, thus implying that significant modification to the algorithm is necessary to make the approach suitable for the given datasets. In absence of ground truth and information on necessary adoption approaches, the ensemble technique can be applied to identify the best-performing algorithms and combine their results to obtain a more robust approach that performs well across different datasets. This approach is therefore useful to facilitate accurate HRV analysis from large-scale ECG signal data.

5. Conclusions

In this paper, we evaluated 15 state-of-the-art QRS complex detection implementations using two preterm neonatal datasets. A weighted ensemble technique using Logistic Regression was applied which outperformed the individual methods for both datasets.

This study suggests using the ensemble-based approach to ensure consistent performance across multiple datasets where individual methods deliver inconsistent performance. Suggested future work in this area is to investigate the monotonicity of the ensemble approach by exploring a broader set of QRS complex detection methods.

References

[1] Latremouille S, Lam J, Shalish W, Sant'Anna G. Neonatal heart rate variability: a contemporary scoping review of analysis methods and clinical applications. BMJ Open. 2021 Dec;11(12):e055209, doi: 10.1136/bmjopen-2021-055209.

[2] McAdams RM, Kaur R, Sun Y, Bindra H, Cho SJ, Singh H. Predicting clinical outcomes using artificial intelligence and machine learning in neonatal intensive care units: a systematic review. Am J Perinatol. 2022 May;13:1-5, doi: 10.1038/s41372-022-01392-8.

[3] Gee AH, Barbieri R, Paydarfar D, Indic P. Predicting bradycardia in preterm infants using point process analysis of heart rate. IEEE Trans Biomed Eng. 2016 Nov;64(9):2300-8, doi: 10.1109/TBME.2016.2632746.

[4] Pan J, Tompkins WJ. A real-time QRS detection algorithm. IEEE Trans Biomed Eng. 1985 Mar;3:230-6, doi: 10.1109/TBME.1985.325532.

[5] Jani P, Lowe K, Hinder M, Galea C, D'Çruz D, Badawi N, Tracy M. Liberal hemoglobin threshold affects cerebral arterial pulsed Doppler and cardiac output, not cerebral tissue oxygenation: a prospective cohort study in anemic preterm infants. Transfusion. 2019 Oct;59(10):3093-101, doi: 10.1111/trf.15452.

[6] Christov II. Real time electrocardiogram QRS detection using combined adaptive threshold. Biomed Eng Online. 2004 Dec;3(1):1-9, doi: 10.1186/1475-925X-3-28.

[7] Engelse WA, Zeelenberg C. A single scan algorithm for QRS-detection and feature extraction. Comput Cardiol. 1979 Sep;6(1979):37-42.

[8] Lourenço A, Silva H, Leite P, Lourenço R, Fred AL. Real Time Electrocardiogram Segmentation for Finger based ECG Biometrics. In Biosignals 2012 Feb (pp. 49-54), doi: 10.5220/0003777300490054.

[9] Gamboa H. Multi-modal behavioral biometrics based on HCI and electrophysiology (Doctoral dissertation, Universidade Técnica de Lisboa).

[10] Martínez JP, Almeida R, Olmos S, Rocha AP, Laguna P. A wavelet-based ECG delineator: evaluation on standard databases. IEEE Trans Biomed Eng. 2004 Mar;51(4):570-81, doi: 10.1109/TBME.2003.821031.

[11] Elgendi M, Jonkman M, DeBoer F. Frequency bands effects on QRS detection. Pan. 2010 Jan;5(15):1-5, doi: 10.5220/0002742704280431.

[12] Kalidas V, Tamil L. Real-time QRS detector using stationary wavelet transform for automated ECG analysis. In 2017 IEEE 17th international conference on Bioinformatics and Bioengineering (BIBE) 2017 Oct 23 (pp. 457-461). IEEE, doi: 10.1109/BIBE.2017.00-12.

[13] Rodrigues T, Samoutphonh S, Silva H, Fred A. A Low-Complexity R-peak Detection Algorithm with Adaptive Thresholding for Wearable Devices. In 2020 25th International Conference on Pattern Recognition (ICPR) 2021 Jan 10 (pp. 1-8). IEEE, doi: 10.1109/ICPR48806.2021.9413245.

[14] Zong W, Heldt T, Moody GB, Mark RG. An open-source algorithm to detect onset of arterial blood pressure pulses. In Comput Cardiol. 2003 2003 Sep 21 (pp. 259-262). IEEE, doi: 10.1109/CIC.2003.1291140.

[15] Nabian M, Yin Y, Wormwood J, Quigley KS, Barrett LF, Ostadabbas S. An open-source feature extraction tool for the analysis of peripheral physiological data. IEEE J Transl Eng Health Med. 2018 Oct;6:1-1, doi: 10.1109/JTEHM.2018.2878000.

[16] Hamilton P. Open source ECG analysis. In Comput Cardiol. 2002 Sep 22 (pp. 101-104). IEEE, doi: 10.1109/CIC.2002.1166717.

[17] Carreiras C, Alves AP, Lourenc¸o A, Canento F, Silva H, Fred A. BioSPPy: Biosignal Processing in Python; 2015–. [Online; accessed]. Available from: https://github.com/PIA-Group/BioSPPy/.

[18] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SA. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behav Res Methods. 2021 Feb;2:1-8, doi: 10.3758/s13428-020-01516-y.

[19] Caritat MJAN, et al. Essai sur l'application de l'analyse a` la probabilite´ des de´cisions, rendues a` la pluralite´ des voix. 1785.