Comparative Study on the Generalization Ability of Machine Learning Algorithms for PPG Quality Assessment

Santiago Mula[1], Roberto Zangróniz[1], Óscar Ayo-Martín[2], José J Rieta[3], Raul Alcaraz[1]

1 Research Group in Electronic, Biomedical and Telecom. Eng., Univ. of Castilla-La Mancha, Spain
2 Department of Neurology, University Hospital of Albacete, Spain
3 BioMIT.org, Electronic Engineering Department, Universitat Politecnica de Valencia, Spain

Abstract

Photoplethysmogram (PPG) signals are often too noisy to accurately measure heart rate. Methods for PPG quality assessment based on machine learning (ML) concepts have gained considerable attention, but they have been commonly developed on datasets recorded under specific conditions and not tested on others. Hence, the aim of the present work is to compare their performance and generalization ability on datasets representative of different conditions.

Three common classifiers, such as decision tree (DT), support vector machine (SVM), and K-Nearest Neighbor (KNN), were trained on a publicly available dataset and tested on two external datasets, also freely available. The obtained results reported a different generalization ability for each classifier. Whereas DT was unable to generalize for any external dataset, SVM and KNN were only able for an external dataset presenting PPG signals acquired with the same recording system and then with a morphology similar to those in the training one. Hence, the selection of datasets for training and validation seems a key aspect to be taken into account to know where a machine learning method can be reliably used for PPG quality assessment.

1. Introduction

A significant challenge using Photoplethysmography (PPG) for continuous heart rate monitoring is its unreliable accuracy, particularly in uncontrolled real-world settings [1]. The inaccuracy in heart rate extraction from PPG signals can be due to individual variations, physiological processes or external factors. Examples of individual differences are body mass, age, sex, and skin melanin levels, which affect the absorption of light. Physiological processes, like breathing, and environmental factors, like measurement site (e.g., wrist, fingertip, ear, forehead), local temperature, ambient light, and movement artifacts, also impact accuracy [2]. As a result, signal quality must be carefully assessed before relying on the measured heart rate data.

There has been extensive research on the development of quality assessment methods for PPG signals. Many of the methods proposed to date use machine learning techniques which use time-frequency features extracted from the PPG signal. The work by Guo et al.[3] approach PPG artifact detection as a 1D segmentation problem, training on the PPG DaLiA [4] dataset and validating on the WESAD [5] and TROIKA [6] datasets. Those three datasets are focused on daily-life activities and physical exercise. The work by Lutin [7] is done on the TROIKA dataset. The work by Pereira [8] compares several methods on recording from ICU patients.

Despite extensive research on developing algorithms for assessing the quality of PPG signals, there is a lack of consistency in evaluation methodology, leading to varied results across different datasets, validation approaches, and performance metrics. This inconsistency hinders comparison and understanding of how different datasets reflect real-world inaccuracies, such as measurement location, sensor wavelength, physical activity or individual variations. To properly use current approaches and develop better ones, a thorough analysis of these differences is essential to identify the limitations of existing methods for PPG quality assessment.

The goal of this work is to compare the performance and generalization ability of common machine learning methods for PPG quality assessment on datasets representative of different challenges a real-world application of PPG for continuous heart rate monitoring would face.

2. Materials and methods

2.1. Datasets

For the training and evaluation of machine learning models, PPG segments must be labeled as clean or noisy. This process can be done through manual annotation or

by comparison with a golden standard. The golden standard for heart rate estimation is the electrocardiogram, or ECG. An example of reference-based labeling is the work by Lutin [7]. Other works use manual labeling [8] [3].

This work employed reference-based labeling. PPG peaks and ECG R peaks were detected with standard methods such as Pan-Tompkins, the difference between heart rate was calculated and signals were divided into 5-second segments, which were labeled as noisy when the difference surpassed a threshold of 3 beats-per-minute, or clean otherwise.

To ensure the reference ECG heart rate is reliable, each ECG segment has been classified as such with the method described by Huerta et al. [9], to avoid a misclassification of the PPG as noisy when the ECG is noisy and the PPG is not, as the heart rates will still differ.

In order to use this quality labeling of the PPG signals, only datasets containing simultaneous PPG and ECG recordings were used. Moreover, to facilitate reproduction of the obtained results, only freely available datasets were considered. The following datasets were used:

- PPG DaLiA contains 15 recordings of 15 different subjects, aged 21 to 55 years with an average of 30.6 years and a variance of 5.95 years. Recordings are around 150 minutes long on activities common on daily-life. The PPG has been recorded with a Empatica E4, with two green LEDs, two red LEDs, and two photodiodes with a sampling frequency of 64 Hz. The ECG has been measured with a RespiBAN and sampled at 700 Hz [4].
- WESAD is composed of recordings of 15 subjects aged 24 to 25 years. Recordings are around 100 minutes long and include daily-life activities, but are recorded on a controlled environment. It uses the Empatica E4 and the RespiBAN as well [5].
- BIDMC contains 53 recordings of 53 subjects and is a subset of the MIMIC II dataset [10] [11], composed of recordings of fingertip-PPG on ICU patients [12] [11]. The PPG and ECG are both sampled at 125 Hz.

All PPG signals have been resampled to 256 Hz and band-pass filtered between 0.7 and 10.0 Hz with an order 3 Butterworth filter. The ECG signals were not resampled to a common frequency because Both the ECG quality assessment method [9] and the Pan-Tompkins peak detection method used are sampling-rate independent.

2.2. Methods

The conventional machine learning approach to classification involves the feature extraction and feature selection steps before actual classification. A range of time-domain features have been employed, including statistical measures such as mean, variance, standard deviation, and root mean square of all samples in a segment, as well as those same features of the first and second derivatives of the signal. Additionally, frequency-domain features have been used, including spectral power by band for 64 bands of 2 Hz, and the mean, variance, standard deviation, and root mean square of all bands spectral power. This feature extraction works in the same or similar manner as previous works [8] [7] [3].

The experiments have been repeated with all combinations of groups of features with the goal of finding the features that achieved the maximum sensitivity and specificity in the training dataset for the best classifier. The groups of features were the time-domain features of the signal, those of the first derivative, those of the second derivative, the spectral power band with 4, 8, 16, 32, 64 and 128 bands with the corresponding mean and deviation of power density for all bands. The experiments were conducted with the features normalized by the largest feature value, with the largest feature value becoming 1 and all values being contained in the 0 to 1 range. Non-normalized features performed better and all results shown here use features without normalization. Since the results on the external evaluations showed no difference between optimizing feature selection for each classifier and using common features for all classifiers, the same features were used for all classifiers.

The best results used the time-domain standard deviation and standard deviation of spectral power by band, with 64 bands and a sample frequency of 256 Hz, each spectral power density divided by the maximum power density of all bands.

The machine learning classification methods used are Decision Tree (DT) [13] [14], Support Vector Machine (SVM) [15] [16] [14] and K-Nearest Neighbors (KNN) [14].

2.2.1. Experimental setup

Training is performed on the WESAD dataset, and external evaluation is performed on the PPG DaLiA and BIDMC datasets. The PPG DaLiA dataset is similar to WESAD in both device, an Empatica E4 as PPG, and setting, wrist PPG and daily life activities. The BIDMC dataset, on the other hand, is a subset of the MIMIC II dataset, recorded on ICU patients with a fingertip PPG.

In the evaluation metrics, a positive is defined as a clean segment and a negative a noisy segment, a true positive is a clean segment correctly classified as clean. Accuracy is the rate of true positives and true negatives over the total segments. Sensitivity is the rate of true positives over total positives and specificity the rate of true negatives over total negatives.

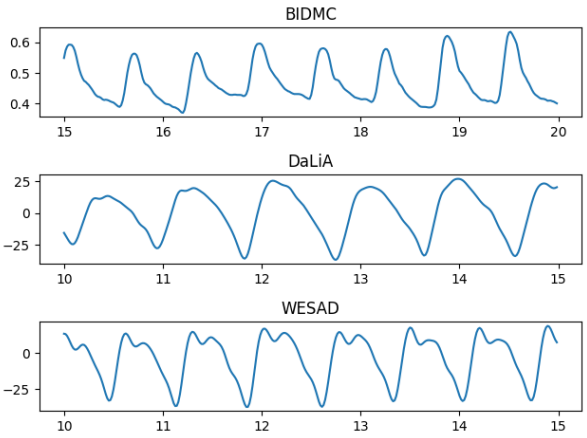

Figure 1. Clean waveforms of each dataset

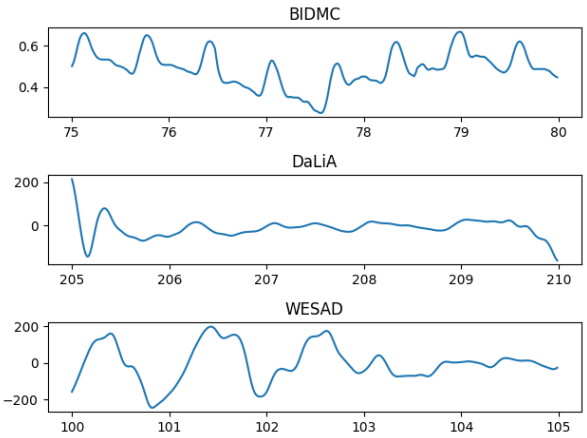

Figure 2. Noisy waveforms of each dataset

3. Results

After discarding segments with noisy ECG, training was performed on 17,201 segments of 5 seconds from the WESAD dataset, accounting for 23 hours and 54 minutes. The first external validation was made on 23,564 segments from the PPG DaLiA dataset, accounting for 32 hours and 44 minutes. The second external validation was made on 4,268 segments from BIDMC, accounting for six hours. WESAD had a clean/total segment ratio of 42.2%, PPG DaLiA a 29.2% and BIDMC a 74.1%.

Table 1 displays classification results obtained by the three machine learning classifiers in the training and testing phases. As can be seen, different methods exhibit a similar behavior.

The DT-based model reached 100% accuracy (Acc) on training, with a 30% drop in sensitivity (Se) and specificity (Sp) in DaLiA and an unbalanced Se-Sp on BIDMC of 31-84%. The SVM-based model reached an Acc of 82% on training, with a 3% performance drop on Se-Sp in DaLiA and 78-21% Se-Sp in BIDMC. Finally, the KNN-based model yielded an Acc of 85% on training and DaLiA, with an Se-Sp unbalance on BIDMC of 27-96%. Overall, all methods show a performance degradation on the second external validation, which is a dataset of a very different nature of the training dataset. The method with a greater generalization was the SVM, but is was still unreliable on the second external validation.

	Dataset	Acc	Sens	Spec
DT	WESAD (train)	1.0000	1.0000	1.0000
DT	DaLiA	0.7070	0.7115	0.7056
DT	BIDMC	0.7081	0.3126	0.8466
SVM	WESAD (train)	0.8233	0.8104	0.8362
SVM	DaLiA	0.7852	0.7888	0.7765
SVM	BIDMC	0.7830	0.2096	0.9839
KNN	WESAD (train)	0.8498	0.8581	0.8415
KNN	DaLiA	0.7739	0.7878	0.7400
KNN	BIDMC	0.7781	0.2629	0.9586

Table 1. Results by method and dataset.

4. Discussion

Whereas the DT-based model showed significant overfitting on the training dataset, the SVM-based and KNN-based models reported a successful performance on the dataset with a PPG morphology similar to that in the training dataset. Those same methods do not generalize to another dataset containing PPG signals with a different morphology, where they are unreliable. This result suggests that previous algorithms proposed for PPG quality assessment cannot be reliably used in every context. As well, the development of new algorithms for the same purpose should pay attention to the settings their training and validation datasets are representative of, because it could limit the scenarios where they could be used reliably.

Finally, some limitations of this work deserve to be commented. Firstly, only two different validation datasets were analyzed. They are only representative of two opposite settings and environments, and additional datasets obtained in different scenarios should be analyzed in the future. Secondly, only machine learning methods have been tested, leaving out the each day more common deep learning methods. A wider comparison of methods for PPG quality assessment should be considered in a further study. Lastly, since no information about synchronization between PPG and ECG signals is provided in the analyzed datasets, this step was manually conducted by aligning the heart rates of both signals. This approach could slightly impact on quality labelling of the PPG signals.

5. Conclusion

The generalization ability of common machine learning classifier for PPG quality assessment has shown notable differences under a unique experimental setup. Whereas DT was unable to generalize for any external dataset, SVM and KNN were only able for an external dataset presenting PPG signals acquired with the same recording system and with a morphology similar to those in the training one. Hence, the selection of datasets for training and validation must be taken into account to know where a machine learning method can be reliably used for PPG quality assessment.

Acknowledgments

This research has received financial support from Daiichi Sankyo SLU and from public grants PID2021-00X128525-IV0, PID2021-123804OB-I00, and TED2021-130935B-I00 of the Spanish Government 10.13039/501100011033 jointly with the European Regional Development Fund (EU), SBPLY/21/ 180501/000186 from Junta de Comunidades de Castilla-La Mancha, and AICO/2021/286 from Generalitat Valenciana.

References

[1] Pankaj, Kumar A, Komaragiri R, Kumar M. A review on computation methods used in photoplethysmography signal analysis for heart rate estimation. Archives of Computational Methods in Engineering May 2021;29(2):921–940. ISSN 1886-1784.

[2] Fine J, Branan KL, Rodriguez AJ, Boonya-ananta T, Ajmal, Ramella-Roman JC, McShane MJ, Coté GL. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. Biosensors 2021;11(4). ISSN 2079-6374.

[3] Guo Z, Ding C, Hu X, Rudin C. A supervised machine learning semantic segmentation approach for detecting artifacts in plethysmography signals from wearables. Physiological Measurement dec 2021;42(12):125003.

[4] Reiss A, Indlekofer I, Schmidt P, Van Laerhoven K. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. Sensors 2019;19(14). ISSN 1424-8220.

[5] Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356923, 2018; 400–408.

[6] Zhang Z, Pi Z, Liu B. Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. IEEE Transactions on Biomedical Engineering 2015;62(2):522–531.

[7] Lutin E, Biswas D, Simoes-Capela N, Van Hoof C, Van Helleputte N. Learning based quality indicator aiding heart rate estimation in wrist-worn ppg. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). 2021; 7063–7067.

[8] Pereira T, Gadhoumi K, Ma M, Liu X, Xiao R, Colorado RA, Keenan KJ, Meisel K, Hu X. A supervised approach to robust photoplethysmography quality assessment. IEEE Journal of Biomedical and Health Informatics March 2020; 24(3):649–657.

[9] Huerta A, Martinez Rodrigo A, Gonzalez V, Quesada A, Rieta J, Alcaraz R. Quality assessment of very long-term ecg recordings using a convolutional neural network. In 2019 E-Health and Bioengineering Conference (EHB). 11 2019; 1–4.

[10] Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG. Multiparameter intelligent monitoring in intensive care ii: A public-access intensive care unit database*. Critical Care Medicine May 2011;39(5):952–960. ISSN 0090-3493.

[11] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. Physiobank, physiotoolkit, and physionet. Circulation 2000;101(23):e215–e220.

[12] Pimentel MAF, Johnson AEW, Charlton PH, Birrenkott D, Watkinson PJ, Tarassenko L, Clifton DA. Toward a robust estimation of respiratory rate from pulse oximeters. IEEE Transactions on Biomedical Engineering 2017;64(8):1914–1923.

[13] Breiman L, Friedman J, Stone C, Olshen R. Classification and Regression Trees. Taylor & Francis, 1984. ISBN 9780412048418.

[14] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011; 12:2825–2830.

[15] Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. Liblinear: A library for large linear classification. Journal of Machine Learning Research 2008;9:1871–1874.

[16] Chang CC, Lin CJ. Libsvm: A library for support vector machines. ACM Trans Intell Syst Technol may 2011;2(3). ISSN 2157-6904.

Address for correspondence:

Santiago Mula Muñoz
Campus Universitario S/N. 16071 - Cuenca (España)
Instituto de Tecnologías Audiovisuales (ITAV)
santiago.mulamunoz@uclm.es