Quality Assessment of Photoplethysmography Signals For Cardiovascular Biomarkers Monitoring Using Wearable Devices

Felipe M. Dias[1], Marcelo A. F. Toledo[1], Diego A. C. Cardenas[1], Douglas A. Almeida[1], Filipe A. C. Oliveira[1], Estela Ribeiro[1], Jose E. Krieger[1], Marco Antonio Gutierrez[1]

[1]Heart Institute (InCor) – Clinics Hospital
University of Sao Paulo Medical School (HCFMUSP)
Sao Paulo – SP – Brazil

f.dias@hc.fm.usp.br, marcelo.arruda@hc.fm.usp.br

diego.cardona@hc.fm.usp.br, douglas.andrade@hc.fm.usp.br

filipe.acoliveira@hc.fm.usp.br, estela.ribeiro@hc.fm.usp.br

j.krieger@hc.fm.usp.br, marco.gutierrez@incor.usp.br

Abstract. *Photoplethysmography (PPG) is a non-invasive technique widely used to monitor cardiovascular parameters such as heart rate. However, its reliability can be compromised by factors like motion artifacts. In this study, we extracted 27 statistical features from PPG signals and trained multiple machine learning models (XGBoost, CatBoost, Random Forest) to assess signal quality. Using a publicly available dataset of PPG time series, we evaluated model performance using sensitivity, positive predictive value, and F1-score. Our best model (CatBoost) achieved 94.7%, 95.9%, and 95.3% for these metrics, respectively. These results are comparable to state-of-the-art approaches but relying on relatively simple models.*

1. Introduction

Photoplethysmography (PPG) is a non-invasive technology that measures changes in blood volume in the microvascular bed of tissue, and is widely used in pulse oximetry (SpO2) devices to assess cardiovascular health [Alian and Shelley 2014] [Reisner et al. 2008]. In addition to pulse oximetry, PPG can potentially be used to measure other important parameters such as heart rate, respiratory rate, and other physiological parameters over time [Charlton et al. 2022]. With its non-invasive nature and ability to provide continuous monitoring, PPG has become an important tool for monitoring cardiovascular health and diagnosing various cardiovascular conditions [Mejía-Mejía et al. 2022].

Generally speaking, PPG measures changes in the blood volume of vascular tissues by shining light over a peripheral tissue and measuring the amount of light that is absorbed and scattered. The attenuated light is detected by an optical sensor, which records the changes in light intensity over time [Mejía-Mejía et al. 2022] [Nitzan and Ovadia-Blechman 2022]. The PPG signal consists of an Alternating Current (AC) component, which is the high-intensity component caused by the light absorption

of hemoglobin in pulsatile arterial blood, and a Direct Current (DC) component, which is the low intensity component caused by changes in other tissues components and non-pulsatile arterial blood [Mukkamala et al. 2022]. The AC component of the PPG signal reflects the changes in blood volume due to the oscillations in blood cell aggregation and blood flow related to changes in arterial blood pressure. Specifically, during systole, when the heart pumps blood and arterial blood pressure increases, there is an increased absorbance of light, leading to a higher AC component of the PPG signal. During diastole, when the heart is filling with blood and arterial blood pressure decreases, there is a decreased absorbance of light, leading to a lower AC component of the PPG signal [Nitzan and Ovadia-Blechman 2022].

The quality of PPG signals obtained from wearable devices is a major concern [Fine et al. 2021]. PPG signals can be affected by various sources of noise, including motion artifact, probe-tissue interface disturbance, such as pressure between the PPG sensor and the skin, baseline interference due to respiration and body movement, low and high frequency noise, and type of sensor and location of the measurement [Elgendi 2012] [Li et al. 2018]. In this context, [Moscato et al. 2022] showed that physical activity affected PPG signal quality in a way that, during rest, 94% of the heartbeats were considered of good quality compared to only 9% during physical activity and suggested that healthy subjects have a better signal quality (44% of good quality) compared to oncological patients (13% of good quality). These factors can negatively impact the PPG signal analysis and hinder the extraction of meaningful features and biomarkers, or even act as confounding factors, invalidating their usage or interpretation. In Fig. 1, we show an illustrative example of a good PPG signal segment, where physiological measurements (e.g., heart rate) can be easily extracted, and a bad PPG signal where such measurements are not reliable.

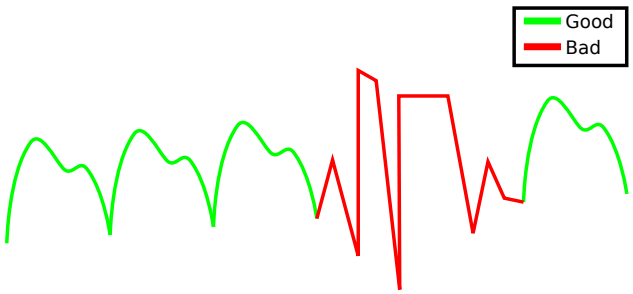

Figura 1. Individual PPG heartbeat signal for two different quality levels: Good and Bad.

Therefore, developing a good signal quality detector for PPG signals is crucial in order to ensure that the signals obtained from wearable devices are of high quality and suitable for analysis. However, the lack of labeled and publicly available datasets of signal quality assessment is a major issue, as it makes it difficult to train and validate the performance of signal quality detectors.

One of the earliest attempts to approach signal quality in PPG signal was proposed by [Elgendi 2016]. In his work, he proposed a recommendation for visual quality assessment annotation of individual heartbeats following three quality levels: (1) excellent, for PPG signals where systolic and diastolic peaks can be clearly detected; (2) acceptable, for

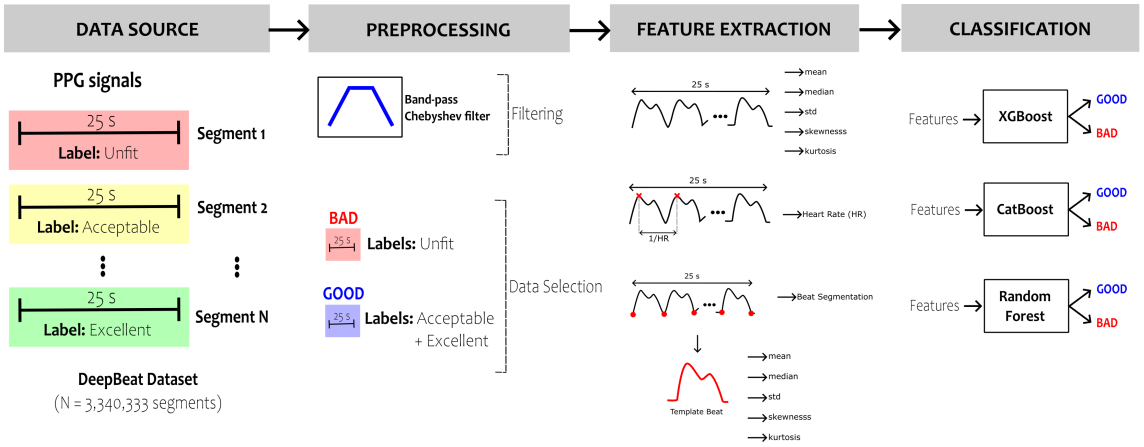

Figura 2. General structure of the proposed methodology to classify PPG signal quality.

PPG signals where heart rate can still be estimated even though the diastolic peak is not salient; and (3) unfit, for noisy PPG signals where systolic and diastolic peaks cannot be distinguished.

Recently, [Torres-Soto and Ashley 2020] created a public PPG quality dataset and proposed the DeepBeat algorithm that used segments of 25 seconds of PPG signals as input to predict signal quality, along with Atrial Fibrillation (AF) rhythm classification. [Mohagheghian et al. 2022] used multiple databases to assess the quality of PPG signals by extracting statistical and morphological features from the signals. [Moscato et al. 2022], using a private dataset, employed a combination of features extracted features both from the accelerometers and PPG waveform to estimate signal quality. [Dias et al. 2022] used the MIMIC-II [Goldberger et al. 2000] dataset and employed a template-based method to assess the quality of PPG signals. They detected the beats in each window of the signal, estimated a template average beat, and computed Pearson's correlation coefficient between each beat and the template average beat. Windows whose mean correlation was lower than a given threshold were considered to have poor signal quality. The field of automatic assessment of PPG signals for cardiovascular biomarkers monitoring is still in its early stages, and there is much room for improvement and further research.

In this study, we propose a signal quality assessment method for PPG signals by extracting 27 statistical features from 25 s segments of these signals. These features are fed to a machine learning model (XGBoost, CatBoost, or Random Forest) which classify the segment as either good or bad quality. The proposed method aims to reduce the effect of noisy beats on PPG signals, improving further analysis and applications.

2. Materials and Methods

In this section, we describe the publicly available dataset used to classify the PPG signal quality (A). Thereafter, our proposed methodology for signal quality classification is described, based on feature extraction of PPG signals (B, C) and the use of three algorithms for the classification step (D). The general structure of the proposed method to classify PPG signal quality is shown in Fig. 2.

2.1. Data source

We used a publicly available dataset provided by Stanford University [Torres-Soto and Ashley 2020] named DeepBeat. This dataset is composed of three different types of signals. In the first part of the dataset, they collected data using a wrist PPG wearable device (Simband), sampled at 128 Hz, from subjects with confirmed Atrial Fibrillation diagnosis, performing elective cardioversion or stress tests. For the second part of the dataset, they generated synthetic physiological signals of sinus rhythms and atrial fibrillation rhythms, adding noise components to this synthetic dataset, simulating high-quality signals with low noise or low-quality signals with high noise. The third and last part of the dataset is composed of PPG data from the 2015 IEEE Signal Processing Cup, to include signals from healthy subjects. The DeepBeat dataset provides signals partitioned into segments of 25 s and split into training, validation, and test partitions, avoiding data leakage, i.e., samples from the same subject don't appear in the training, validation, and test sets. To provide the labels for each 25 s segments, the authors of this dataset used the Elgendi's quality assessment [Elgendi 2016] recommendation and labeled 1,000 randomly selected segments. A separate model was trained with these 1,000 labeled segments, predicting quality labels for all the remaining segments of the dataset. In summary, the DeepBeat dataset provides 25 s segments labeled by a model proposed by the dataset authors into three classes, i.e., Excellent, Acceptable and Unfit. Table 1 show the number of 25 s segments for each class, distributed on each partition, available on DeepBeat dataset.

Tabela 1. Summary of the DeepBeat dataset.

Class	Training	Validation	Test	Total
Excellent	550,702	124,995	3,246	678,943
Acceptable	281,024	64,647	2,032	347,703
Unfit	1,972,208	329,140	12,339	2,313,687
Total	2,803,934	518,782	17,617	3,340,333

With this approach, this dataset doesn't score individual heartbeats, as proposed in Elgendi's quality assessment [Elgendi 2016], and the authors didn't make it clear what criteria they used to determine if a given segment was defined as excellent, acceptable, or unfit. Besides, the authors didn't specify which dataset a particular segment corresponds to, i.e., the one collected from AF subjects, the synthetic, or the healthy subjects dataset. Furthermore, most of the data was labeled by a quality assessment model, which means that the labels are not entirely reliable, and the metrics of this model's performance were not presented in [Torres-Soto and Ashley 2020]. Regardless of these constrains, this is the only publicly available dataset on this matter, to the best of our knowledge.

2.2. Preprocessing

In our work, we decided to merge the labels excellent and acceptable signals (see Table 1) into one unique label, resulting in a binary class, i.e., Good (excellent and acceptable) and Bad signals (Unfit). All the PPG signal 25 s segments were filtered using a 4th-order Chebyshev type II bandpass filter of 0.5 – 10 Hz.

2.3. Feature extraction

To extract features from each of the PPG segments, we performed the following steps, as illustrated in Figure 3:

1. *Step 1:* Extracted Heart Rate (HR) and 5 statistical features (mean, median, standard deviation, skewness, and kurtosis) from the full 25 s segments;
2. *Step 2:* Performed beat segmentation using the Multi-Scale Peak and Trough Detection (MSPTD) [Bishop and Ercole 2018] algorithm for each segment, resulting in N beats per segment;
3. *Step 3:* Defined a Template Beat as the average beat of the segment;
4. *Step 4:* Extracted 5 statistical features (mean, median, standard deviation, skewness, and kurtosis) from the Template Beat;
5. *Step 5:* Computed the area within $\pm$ 1 std of the Template Beat according to the remaining beats;
6. *Step 6:* Computed the Dynamic Time Warping (DTW) distance for each individual beat with the Template Beat ($\mathbf{DTW} = [DTW_1, DTW_2, \ldots, DTW_N]$) and extracted 5 statistical features from the resulting $\mathbf{DTW}$ vector;
7. *Step 7:* Computed the Euclidean distance for each individual beat with the Template Beat ($\mathbf{E} = [E_1, E_2, \ldots, E_N]$) and extracted 5 statistical features from the resulting $\mathbf{E}$ vector;
8. *Step 8:* Computed Pearson's correlation for each individual beat with the Template Beat ($\rho = [\rho_1, \rho_2, \ldots, \rho_N]$) and extracted 5 statistical features from the resulting ρ vector.

These steps result in a set of $n = 27$ features that were used to perform the quality assessment of the PPG signals. Table 2 displays a summary of the features extracted.

2.4. Classification (D)

We used three traditional and well-known machine learning algorithms for the quality assessment of the PPG signals, being them: XGBoost [Chen and Guestrin 2016]; CatBoost [Prokhorenkova et al. 2017]; and Random Forest [Breiman 2001]. XGBoost (eXtreme Gradient Boosting) [Chen and Guestrin 2016] is a gradient boosting algorithm used for classification and regression tasks, capable to handle complex patterns in data, and deliver high model performance, providing efficient computation for large datasets. Likewise, CatBoost (Categorical Boosting) [Prokhorenkova et al. 2017] is another gradient boosting algorithm designed to handle categorical features. Finally, Random Forest [Breiman 2001] is an ensemble learning algorithm based on the concept of decision trees, but instead of using a single decision tree, Random Forest combines multiple decision trees to make predictions in a more robust and accurate manner, providing a balance between performance and interpretability. For these three methods, we employed their default hyperparameter values.

The DeepBeat dataset [Torres-Soto and Ashley 2020] already provide the data split in three sets: training, validation and testing, as described on Table 1. Our results were generated using the testing set. To evaluate the employed models, we assessed three different metrics, including Sensitivity (Se), Positive Predicted Value (PPV), and F1-score (F1). Sensitivity measures the proportion of actual positive samples that are correctly predicted, positive predictive value measures the proportion of predicted positive samples

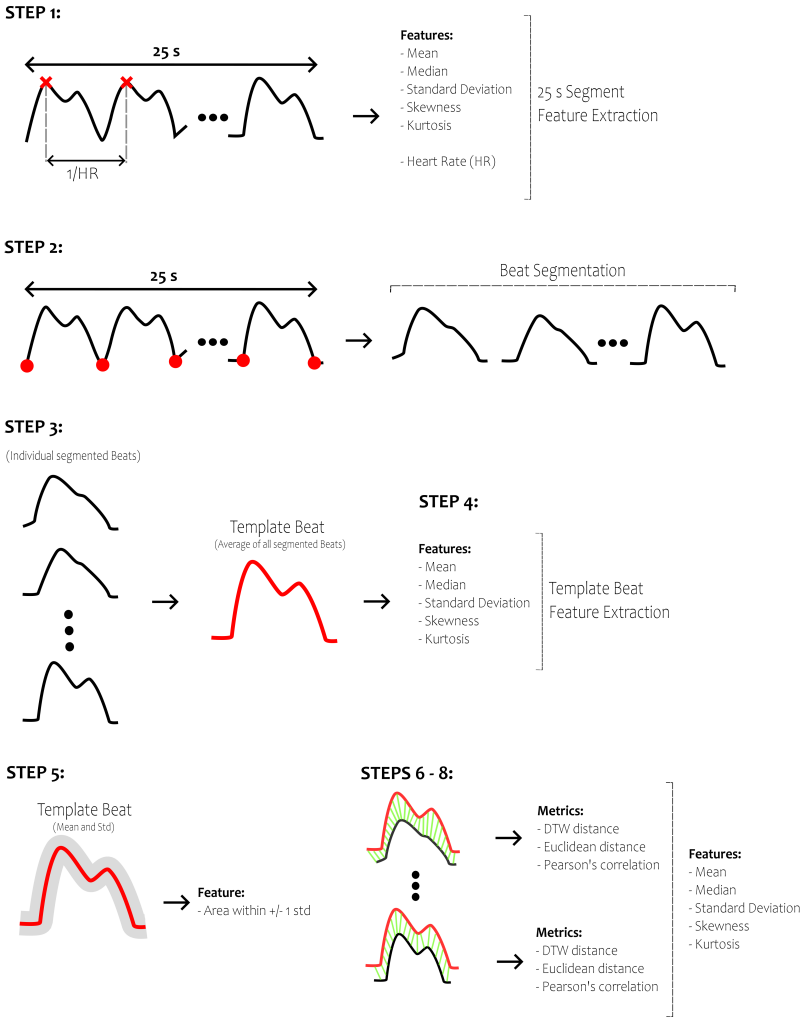

Figura 3. Overview of the 8 feature extraction steps.

that are correct, and F1-score is a single performance metric that balances sensitivity and positive predictive value.

2.5. Experimental setup

The experiments were performed in Python (3.8.10) with the support of the libraries scikit-learn (1.1.3), XGBoost (1.6.1), numpy (1.19.5), scipy (1.8.1), and catboost (1.1.1).

3. Results

Table 3 provides the performance results of the three proposed models for PPG quality assessment and the comparison with the results found in the current state-of-the-art works. We achieved Sensitivity (Se), Positive Predictive Value (PPV), and F1-score of 94.4%, 95.6%, and 95.0%, respectively, using the XGBoost method. Additionally, using Random Forest, we obtained 93.7%, 91.3%, and 92.5% for Se, PPV, and F1-score, respectively. Our best results were obtained using CatBoost, achieving Se of 94.7%, PPV of 95.9%, and F1-score of 95.3%.

Figures 4a, 4b and 4c displays the confusion matrix for the three proposed algorithms on the test set.

Tabela 2. Summary of the features extracted.

Index	Feature name
0	Mean of the full 25 s segments
1	Median of the full 25 s segments
2	Standard deviation of the full 25 s segments
3	Skewness of the full 25 s segments
4	Kurtosis of the full 25 s segments
5	Heart Rate
6	Mean of the template
7	Median of the template
8	Standard deviation of the template
9	Skewness of the template
10	Kurtosis of the template
11	Area within $\pm$ 1 std of the template
12	Mean DTW distance
13	Median DTW distance
14	Standard deviation DTW distance
15	Skewness DTW distance
16	Kurtosis DTW distance
17	Mean Euclidean distance
18	Median Euclidean distance
19	Standard deviation Euclidean distance
20	Skewness Euclidean distance
21	Kurtosis Euclidean distance
22	Mean Pearson's correlation
23	Median Pearson's correlation
24	Standard deviation Pearson's correlation
25	Skewness Pearson's correlation
26	Kurtosis Pearson's correlation

Tabela 3. Comparison of the performance results of our three proposed algorithms for PPG quality assessment and the related state-of-the-art works.

Algorithm	Dataset	**Se**(%)	**PPV**(%)	**F1**(%)
Our method: CatBoost	DeepBeat (test set)	94.7	95.4	95.0
Our method: XGBoost	DeepBeat (test set)	94.6	95.2	94.9
Our method: Random Forest	DeepBeat (test set)	94.6	95.0	94.8
[Torres-Soto and Ashley 2020]	DeepBeat (test set)	97.6	97.4	97.5
[Mohagheghian et al. 2022]	Multiple datasets	86.2	98.4	91.9

Finally, we show the feature importance measures on Figure 5, providing insights into which features are most important in predicting the outcome for the three proposed models.

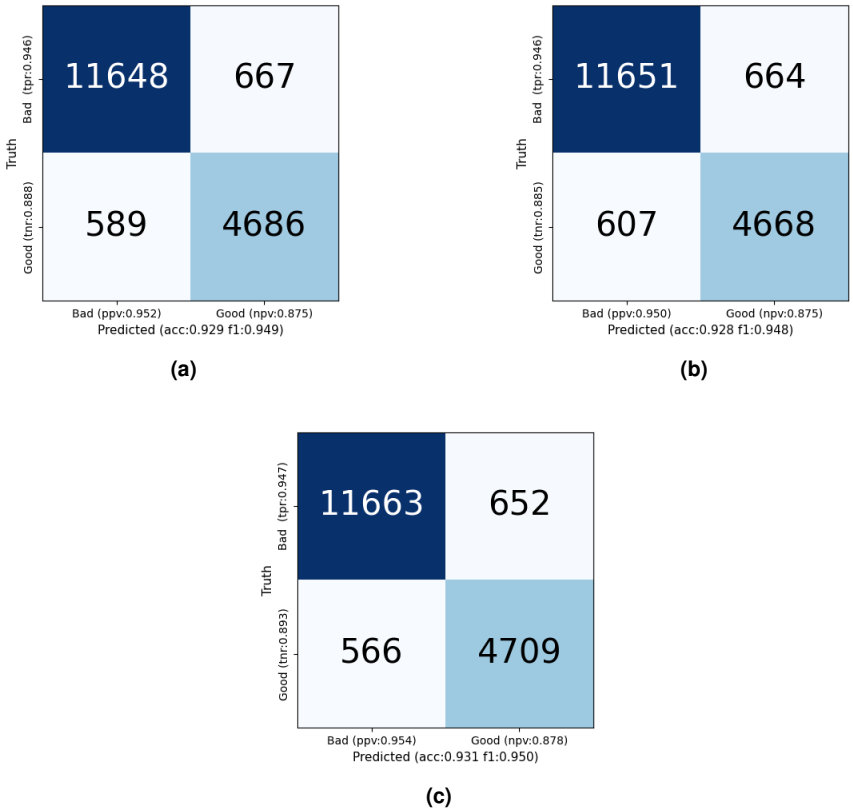

Figura 4. Confusion Matrix of the three proposed algorithms on the test set: (a) XGBoost; (b) Random Forest and (c) CatBoost.

4. Discussion

In this study, we propose a promising alternative method to evaluate the quality of photoplethysmography (PPG) signals by comparing it with other methodologies proposed in the literature. It is worth noting that only a few studies have proposed methodologies for assessing PPG signal quality, mostly using different datasets. Therefore, comparing our results with those of other studies is challenging.

We obtained our best results using the CatBoost algorithm, achieving Sensitivity (Se), Positive Predictive Value (PPV), and F1-score of 94.7%, 95.9%, and 95.3%, respectively. Compared to other feature-based approach, i.e., Mohagheghian et al. [Mohagheghian et al. 2022], we obtained a superior F1-score. On the other hand, compared to the deep learning approach proposed by Torres-Soto and Ashley [Torres-Soto and Ashley 2020], we achieved competitive results using a simpler approach.

Moreover, features 5 and 22, Heart Rate and Mean Pearson's correlation respectively, are indicated as the most important for our CatBoost algorithm. XGBoost algorithm considers the Mean Pearson's correlation feature as the most relevant. Random Forest algorithm also considers this feature as the most relevant. It seems that Mean Pearson's correlation feature carry important information on the quality of the PPG signal. Even though this feature is significant for all algorithms, Random Forest and CatBoost algorithms still consider other features on their predictions.

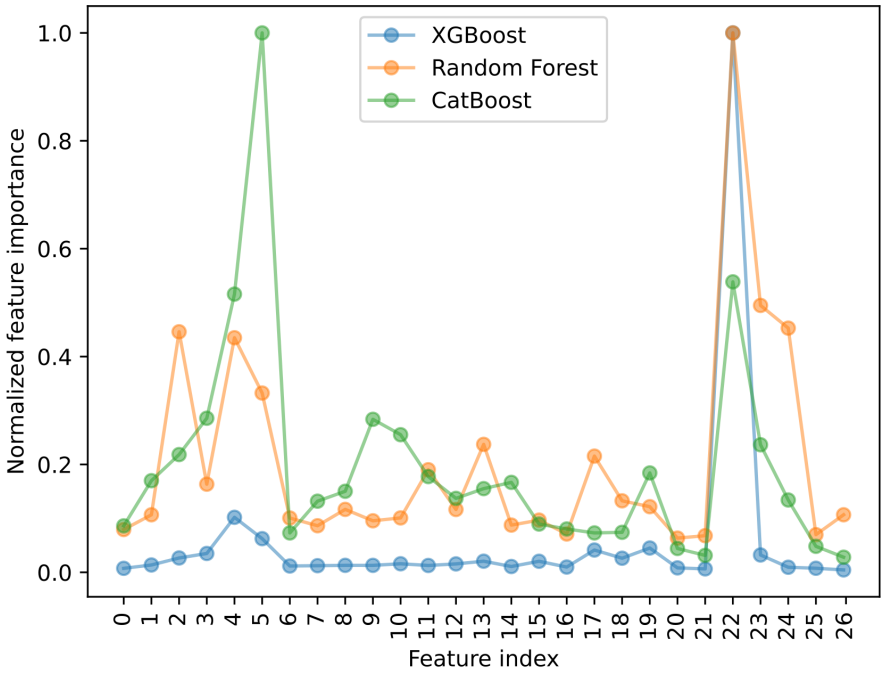

Figura 5. Feature importance measures for the three proposed algorithms.

Our results, which used a less complex model than the state-of-the-art, demonstrate that Machine Learning models are advantageous for creating a pipeline for assessing PPG signal quality, reducing noise and improving the reliability of subsequent analyses and applications on PPG devices used for remote, non-invasive, and continuous monitoring.

Since PPG signals can be affected by various noise sources due to motion disturbances and acquisition condition, which can reduce their morphological quality, it consequently can impact the accuracy of the information retrieved [Moscato et al. 2022].

PPG quality assessment can also help to identify faulty wearable devices and prevent false measurements. This is important in clinical settings, where accurate and reliable measurements are critical for making informed decisions about patient care. This is especially important for the monitoring of cardiovascular biomarkers, as inaccurate or unreliable measurements can have severe implications for the diagnosis, treatment, and management of cardiovascular diseases.

The automatic signal quality evaluation technique proposed here aims to increase the reliability of the PPG parameters and expand its practical applicability. Future works on PPG quality assessment for cardiovascular biomarkers monitoring using wearable devices should consider validate the proposed methods on a larger and more diverse dataset, including patients with different medical conditions, ages, and ethnicities, to ensure their effectiveness and reliability. To do so, it is necessary the development of well-annotated datasets that include a wide range of signal qualities, including low and high-quality signals, and different types of noise sources, such as motion artifacts, ambient light, and

physiological noise.

5. Conclusion

The proposed signal quality assessment of PPG signals can help improve the accuracy and reliability of physiological parameters, such as respiratory rate and heart rate, in wearable devices by reducing the impact of unfavorable factors such as motion disturbances. This is important in the context of continuous monitoring and to ensure the wide applicability of PPG signals in various applications.

In conclusion, our proposed method of PPG signal quality assessment using statistical features shows promising results. However, the limitations of the available database used in this study need to be addressed for a fairer evaluation. Further improvement and validation of the method is necessary with a larger and more diverse dataset, which would lead to a more robust and practical PPG signal quality assessment for various applications.

Acknowledgements

This study was financially supported in part by Foxconn Brazil and the Zerbini Foundation as part of the research project "Machine Learning in Cardiovascular Medicine".

Competing interests

The authors declare no competing interests.

Referências

Alian, A. A. and Shelley, K. H. (2014). Photoplethysmography. *Best Practice & Research Clinical Anaesthesiology*, 28(4):395–406.

Bishop, S. M. and Ercole, A. (2018). Multi-scale peak and trough detection optimised for periodic and quasi-periodic neuroscience data. In Heldt, T., editor, *Intracranial Pressure & Neuromonitoring XVI*, pages 189–195, Cham. Springer International Publishing.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Charlton, P. H., Kyriacou, P. A., Mant, J., Marozas, V., Chowienczyk, P., and Alastruey, J. (2022). Wearable photoplethysmography for cardiovascular monitoring. *Proceedings of the IEEE*, 110(3):355–381.

Chen, T. and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Dias, F. M., Costa, T. B., Cardenas, D. A. C., Toledo, M. A. F., Kriger, J. E., and Gutierrez, M. A. (2022). A machine learning approach to predict arterial blood pressure from photoplethysmography signal. In *2022 Computing in Cardiology (CinC)*, volume XX, pages 1–4.

Elgendi, M. (2012). On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25.

Elgendi, M. (2016). Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4).

Fine, J., Branan, K. L., Rodriguez, A. J., Boonya-ananta, T., Ajmal, Ramella-Roman, J. C., McShane, M. J., and Coté, G. L. (2021). Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors*, 11(4).

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220.

Li, S., Liu, L., Wu, J., Tang, B., and Li, D. (2018). Comparison and noise suppression of the transmitted and reflected photoplethysmography signals. *BioMed research international*, 2018.

Mejía-Mejía, E., Allen, J., Budidha, K., El-Hajj, C., Kyriacou, P. A., and Charlton, P. H. (2022). 4 - photoplethysmography signal processing and synthesis. In Allen, J. and Kyriacou, P., editors, *Photoplethysmography*, pages 69–146. Academic Press.

Mohagheghian, F., Han, D., Peitzsch, A., Nishita, N., Ding, E., Dickson, E. L., DiMezza, D., Otabil, E. M., Noorishirazi, K., Scott, J., Lessard, D., Wang, Z., Whitcomb, C., Tran, K.-V., Fitzgibbons, T. P., McManus, D. D., and Chon, K. H. (2022). Optimized signal quality assessment for photoplethysmogram signals using feature selection. *IEEE Transactions on Biomedical Engineering*, 69(9):2982–2993.

Moscato, S., Palmerini, L., Palumbo, P., and Chiari, L. (2022). Quality assessment and morphological analysis of photoplethysmography in daily life. *Front. Digit. Health*, 4:912353.

Mukkamala, R., Hahn, J.-O., and Chandrasekhar, A. (2022). 11 - photoplethysmography in noninvasive blood pressure monitoring. In Allen, J. and Kyriacou, P., editors, *Photoplethysmography*, pages 359–400. Academic Press.

Nitzan, M. and Ovadia-Blechman, Z. (2022). 9 - physical and physiological interpretations of the ppg signal. In Allen, J. and Kyriacou, P., editors, *Photoplethysmography*, pages 319–340. Academic Press.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2017). Catboost: unbiased boosting with categorical features.

Reisner, A., Shaltis, P., McCombie, D., Asada, H., Warner, D., and Warner, M. (2008). Utility of the Photoplethysmogram in Circulatory Monitoring. *Anesthesiology*, 108(5):950–958.

Torres-Soto, J. and Ashley, E. A. (2020). Multi-task deep learning for cardiac rhythm detection in wearable devices. *npj Digital Medicine*, 3(1):1–8.