Signal Quality Assessment of Photoplethysmogram Signals Using Hybrid Rule- and Learning-Based Models

Avaliação da Qualidade do Sinal de Sinais de Fotopletismografia Usando Modelos Híbridos Baseados em Regras e Aprendizado

Evaluación de la Calidad de Señal de Señales de Fotopletismografía Utilizando Modelos Híbridos Basados en Reglas y Aprendizaje

Giovani Lucafó, Pedro Freitas, Rafael Lima, Gustavo da Luz, Ruan Bispo, Paula Rodrigues, Frank Cabello, Otavio Penatti

Samsung R&D Institute

Autor correspondente: Giovani Decico Lucafó
E-mail: g.lucafo@samsung.com

Abstract

Photoplethysmography signals are crucial for a wide range of applications and, therefore, high-quality PPG signals are crucial to describe the cardiorespiratory status accurately. Motion artifacts can impair PPG-based applications, especially when these signals are recorded via wearable devices. Taking that in consideration, some researchers had proposed few methods for assessing the quality of these signals. Some rule- and learning-based approaches for PPG signal are available to determine the quality of the signal. In this paper, we propose a tradeoff between these two approaches by introducing a hybrid model that employs both learning and decision rules to determine the quality of the signal.
Keywords: Quality Assessment; Photoplethysmography; Deep Learning

Resumo

Sinais de fotopletismografia são cruciais para uma ampla gama de aplicações e, portanto, sinais PPG de alta qualidade são cruciais para descrever o estado cardiorrespiratório com precisão. Artefatos de movimento podem prejudicar aplicativos baseados em PPG, especialmente quando esses sinais são registrados por meio de dispositivos vestíveis. Levando isso em consideração, alguns pesquisadores propuseram poucos métodos para

avaliar a qualidade desses sinais. Algumas abordagens baseadas em regras e aprendizado para o sinal PPG estão disponíveis para determinar a qualidade do sinal. Neste artigo, propomos uma troca entre essas duas abordagens, introduzindo um modelo híbrido que emprega regras de aprendizado e decisão para determinar a qualidade do sinal.
Descritores: Avaliação de Qualidade; Fotopletismografia; Aprendizado Profundo

Resumen

Las señales de fotopletismografía son cruciales para una amplia gama de aplicaciones y, por lo tanto, las señales PPG de alta calidad son cruciales para describir con precisión el estado cardiorrespiratorio. Los artefactos de movimiento pueden interrumpir las aplicaciones basadas en PPG, especialmente cuando estas señales se registran a través de dispositivos portátiles. Teniendo esto en cuenta, algunos investigadores han propuesto algunos métodos para evaluar la calidad de estas señales. Algunos enfoques de aprendizaje y basados en reglas para la señal PPG están disponibles para determinar la calidad de la señal. En este documento, proponemos una compensación entre estos dos enfoques, introduciendo un modelo híbrido que emplea reglas de aprendizaje y decisión para determinar la calidad de la señal.
Descriptores: Evaluación de la calidad; Fotopletismografía; Aprendizaje profundo

Introduction

Heart rate (HR) is an important indicator of cardiovascular healthiness, and electrocardiography (ECG) instruments (e.g., Holter monitor) have consistently be managed it. Although ECG provides a precise description of HR, it requires cumbersome electrodes to be in direct contact with the human skin, what restricts its applicability to daily use applications. More recently, wearable instruments based on Photoplethysmography (PPG) have pushed for HR tracking in considerably easier

settings than the ECG, resulting in devices of increasing popularity for continuous HR monitoring.

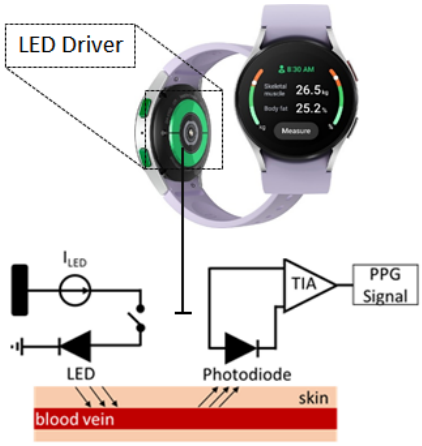


Figure 1 – Working principle of reflective PPG

The operational principle of a reflective PPG is illustrated in Figure 1. It consists of a light emitting diode (LED) driver, a photodiode, a trans-impedance amplifier (TIA), and a low-pass filter. The LED driver that conducts currents up to a few dozens of milliamps typically manages the energy consumption of the sensor. The intensity of the reflected light from the LED is modulated by the heartbeats, with the reflected light increasing during diastole and diminishing during systole. By extracting the peaks of the PPG waveform, HR can be calculated by considering the inter-beat interval (IBI) which may be computed as the time difference between two consecutive peaks (or consecutive valleys).

In general, PPG signals can be severely damaged by several factors, especially motion artifacts. The excess of movement of the PPG sensor can cause a morphological disrupt of the waveform, affecting future analysis on the signal. Taking this in consideration, many works nowadays features the analysis of signal quality with the goal of indexing and repairing disrupted signals. The existing works are modeled using solely one of the learning- or decision-based approach (2; 3). To our knowledge, there are no works which model the signal quality assessment using a hybrid approach. Therefore, in this work, we seek to fill this gap by proposing a hybrid solution which combines advantages of both approaches in order to achieve a cheaper model in terms of computational complexity while keeping the benefits of learning methods.

Materials and Methods

The dataset used for the paper comprises 56 subjects containing data from PPG collected with a sampling rate of 25Hz lasting 45-60 minutes per subject, which required manual annotation by experts for generating the signal quality labels. The PPG signals are split into windows of 3 seconds length, corresponding to 75 samples each with an overlap of 5 samples. The process of quality annotation can be performed using an annotation tool such as that depicted in Figure 2. This tool allows the annotation by visual inspection of the signal, distinguishing between high and low quality samples taking the concurrently measured ECG signal as reference. To create this dataset, the study was conducted on four clinical sites and the Institutional Review Boards (IRBs) documents were considered for each site.

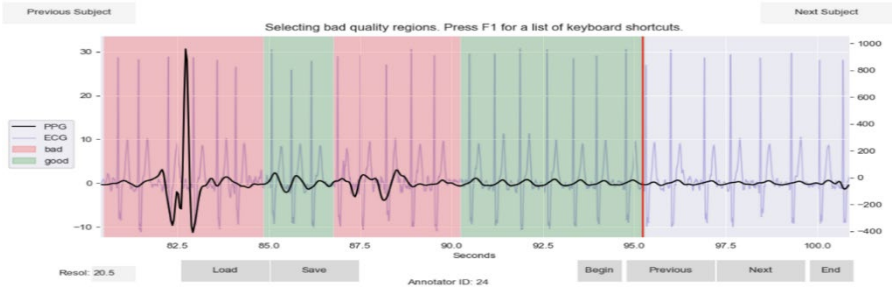


Figure 2 – Annotation tool.

During this visual signal quality assessment process, we noticed a strong variation of amplitudes of the PPG signal, which ended up significantly degrading the signal quality. This led to the conclusion that the majority of the high quality signal is situated below a specific interval range, meaning that PPG windows having high amplitude variance tend to present low quality signal, while windows having lower amplitude variances can have either low or high quality, as depicted in Figure 3. From that, we conducted a threshold analysis to determine which value could be used to automatically discard every signal window with variance value above this given limit. This method aims to distinguish the PPG signal based on the max-to-min amplitude, reducing the number of required CNN calls through bypassing low quality data with a simple decision rule.

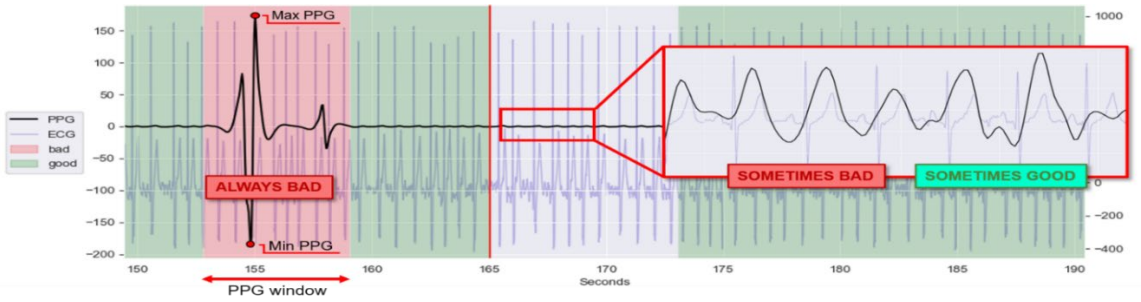
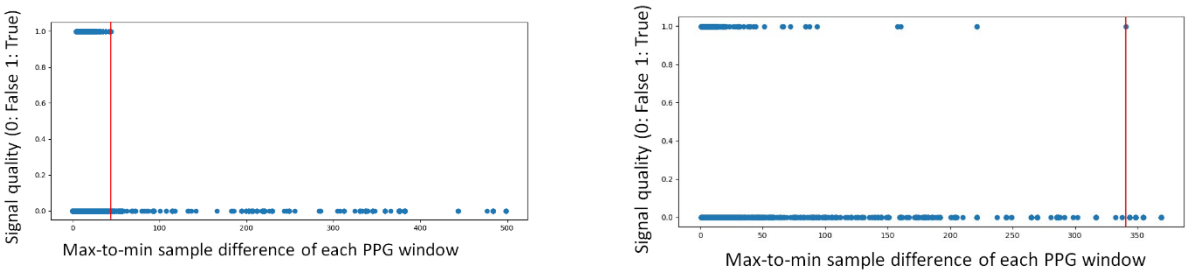


Figure 3 – Representation of the PPG classification by max-to-min amplitude difference.

Two thresholding methods were tested on the annotated dataset aiming to find the best threshold τ capable of eliminating a considerable amount of low quality signal without excessively discarding low quality windows in the decision process. Maintaining a high coverage of high quality windows (number of high quality windows covered by the threshold over the total number of windows of the dataset) allows for minimum impact on the training. The candidate methods implemented to find the optimal threshold value are the Last Value and Nearest Neighbors Search Thresholding, inspired by (5).

Last Value Thresholding

We proposed the Last Value Thresholding (LVT) approach to determine the threshold by setting the maximum D_i^{HQ} (Min-to-max PPG difference of the i-th good quality window) as the delimiter. LVT guarantees that the coverage of D_i^{HQ} will be 100%, where coverage is amount of preserved good windows in relation to all good windows. The main weakness LVT is its sensitiveness to outliers. This means that if the dataset has an abnormally high D_i^{HQ} value, the threshold will keep a considerable amount of bad quality PPG windows that ideally would be discarded from the classification due to their large amplitudes. This problem may be best visualized in the Figure 4, where the threshold was pushed away by an outlier and did not had too much effect on discarding obvious low quality PPG windows.



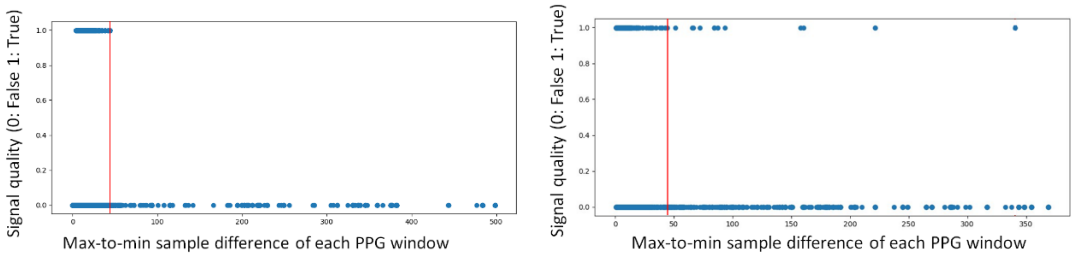
(a) Low variance *(b) High variance*

Figure 4 – LVT on a low (a) and a high (b) variance good quality windows. LVT is only suitable when all good windows are constrained in very limited range of signal intensity. The red vertical line represents the threshold.

Nearest Value Thresholding

The second method considered was the Nearest Neighbor Search (NNS) (5) which performs a search among clusters of D_i^{HQ} flexibilized by the utilization of different search radii. As implemented, all of the D_i^{HQ} values are sorted (D_k^{HQ}) and the distance between D_k^{good} and D_{k+1}^{good} is calculated and compared with the radius. When the distance calculated is greater than the radius, the threshold is defined as D_k^{HQ}.

The referred method showcased an advantageous robustness to data outliers from the threshold analysis, as mentioned in the Last Value Thresholding method, as shown in Figure 5. Because the majority of D_i^{HQ} are situated in a specific range, this method can retain almost every high quality window. For experimental purposes, different values of radii were tested on every subject. Each radius was chosen by empirical analysis during the first experiments on the NNS method.



(a) Low variance *(b) High variance*

Figure 5 – NNS method on a low (a) and high variance (b) of min-to-max PPG difference. The NNS is robust to different variances, as the optimal threshold remains stable despite of the data variance. The red vertical line represents the threshold.

Proposed Signal Quality Classifier

The proposed Signal Quality Classifier (SQC), depicted in Figure 6, has a hierarchical structure composed of an initial decision stage with the purpose of discarding PPG windows that are most likely to be of low quality based on the empirical observation of a given valid signal interval. The second stage of the SQC consists of the classification of the PPG window by a designed CNN.

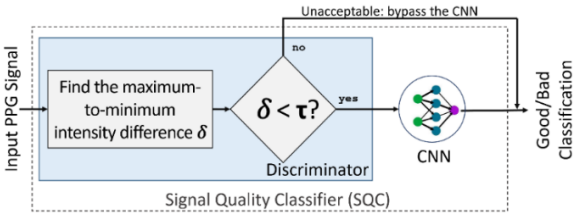


Figure 6 – Proposed signal quality classifier.

Discriminator: The idea behind the discriminator is to analyze each PPG window provided as input and decide, based on its difference between maximum and minimum values δ, if the whole window will be considered to be of low quality or possibly a high quality one, as previously presented in the Materials and Methods section. This decision is assisted by the threshold τ chosen from the earlier mentioned methods. The results regarding the most empirically reliable thresholds will be further discussed in the Results section.

CNN: The idea behind this stage is to process the data considered to be of possibly high quality based on the discrimination stage described above. The stage acts on a more limited interval, when a more refine analysis is requested to detect subtle waveform differences of the signal, in order to classify it. This type of analysis was found to be more suitable for learning-based methods.

Experimental Setup

For experimental purposes, two CNN models were implemented using Keras 2.8.0 and Python 3.8. These CNNs were designed as illustrated in Figure 7 and Figure 8, in order to compare the results of different models in the same signal.

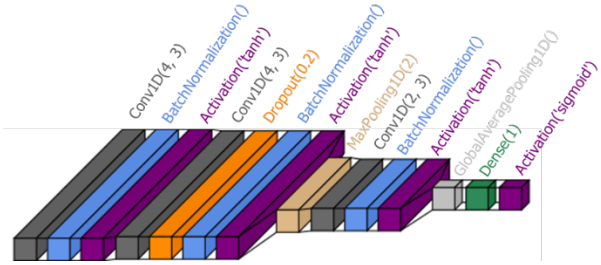


Figure 7 - Architecture of CNN1

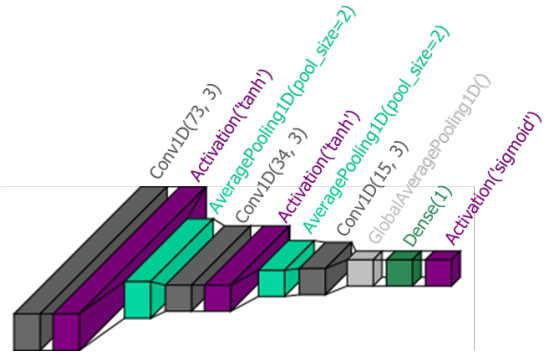


Figure 8 - Architecture of CNN2

The main difference between each model is the number of channels and the intermediate pooling and normalization layers. CNN1 employs a batch normalization step between convolutional layers. On the other hand, CNN2 performs an average pooling that compresses the data in half using the average function. This pooling is also applied between every pair of consecutive convolutional layers. Both CNNs are a compact alternatives for signal processing applications, inspired by (6; 7).

Each of the CNNs was tested with and without the discriminator, as a way of analyzing the individual impact of the discrimination stage on the overall signal classification performance. The proposed model uses CNN 1 and CNN2 individually combined with the discriminator stage and will be referred to as H1 and H2, respectively. Additionally, Hao's rule-based method (2), hereby referred to as Hao, is used for comparison purposes.

Results

The thresholding parameter was determined using Last Value Thresholding (LVT) and the Nearest Neighbor Search Thresholding (NNS) methods. For each of the methods, the results are depicted in Table 1 using the mean of the thresholds calculated from each subject.

Table 1 - Threshold calculated by each method.

Method	Threshold	Standard Deviation	Coverage (%)
LVT	49.87432	76.83759	1.00000
NNS (R = 1)	11.68113	5.72458	0.90800
NNS (R = 3)	15.53522	6.72193	0.96326
NNS (R = 5)	17.92992	8.06483	0.99921
NNS (R = 7)	19.22866	8.22255	0.99927
NNS (R = 10)	20.38929	8.87949	0.99939

It was possible to observe that the LVT method returned the highest value of the threshold mean due to always considering the maximum D_i^{HQ} value. As this method showed high susceptibility to outliers, we observed that the standard deviation of the threshold measurements were considerably high. In counterpart, The NNS tests showed lower values of standard deviation, meaning that these measurements tend to be more reliable than the LVT. Unlike the LVT, the NNS's thresholds discard some high quality PPG windows to discriminate in a more precise way the low quality PPG windows. With that in mind, it is important to notice that all of the NNS's methods tested achieved decent coverages, especially with the usage of radiuses 3, 5, 7, and 10, meaning that the impact of losing some high quality windows will be minimum.

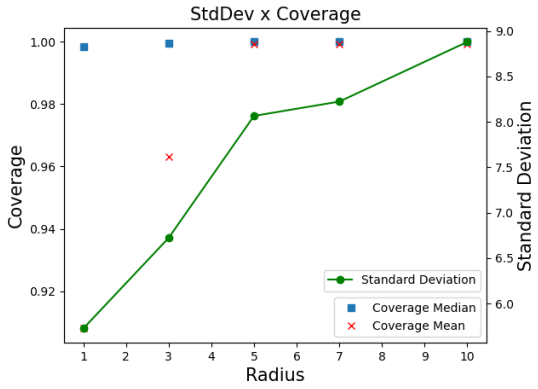
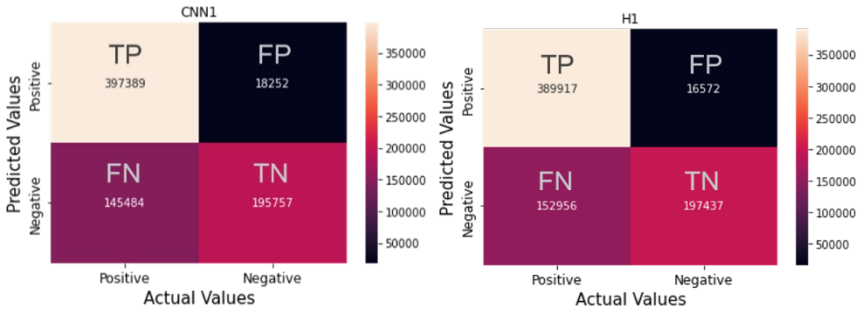
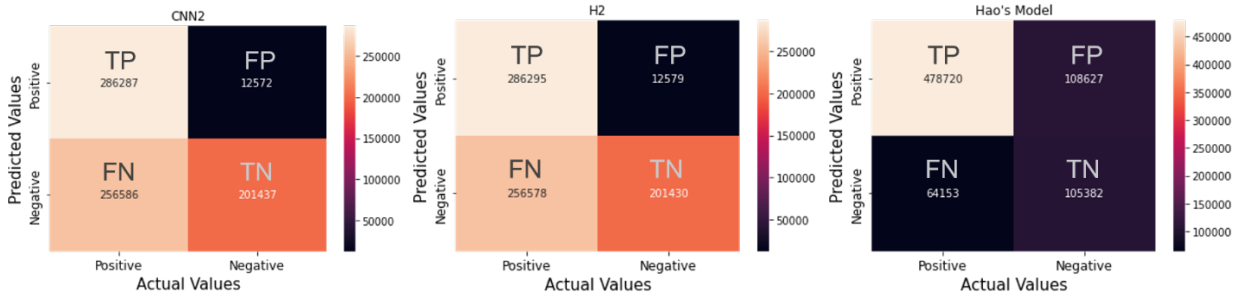


Figure 9 – Coverage versus Standard Deviation using NNS.

As depicted in Figure 9, the coverage mean reaches its optimal value with radius set to seven, considering that the increase of the standard deviation from radii 5 to 7 is negligible. Therefore, the optimum threshold obtained was 19.88266 computed through NNS (R = 7).



(a) CNN1 (b) H1



(c) CNN2 (d) H2 (e) Hao

Figure 10 – Confusion matrices of each model tested on learning-based classification stage presenting the classification of each PPG window.

After the first stage of signal discrimination, the PPG windows that were considered as possibly of high quality are served as input to the learning-based classification stage.

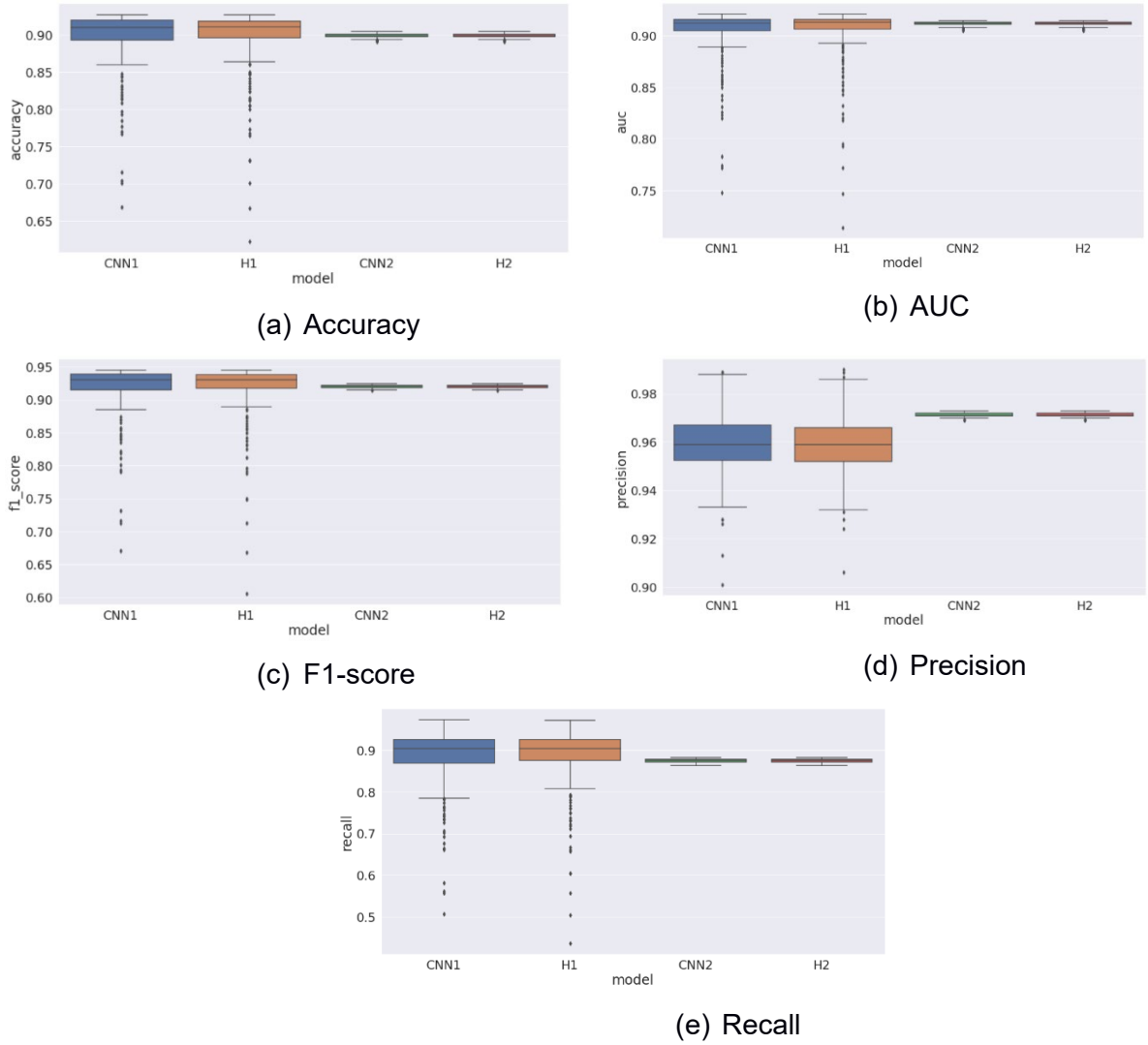
In this part of the classification process, two train-test methodologies were applied to compare the models' performance over the dataset: the Leave-One-Subject-Out (LOSO) and the train-test methodology. The LOSO methodology separates only one subject of the dataset for testing while the remaining subjects are used for training the model. This procedure was done for every subject and the final metrics were computed based on the compilation of the results from every subject in Table 2.

The CNN2 tended to misclassify more high quality PPG windows than CNN1, as it is possible to observe in Figure 10. That means that CNN2 achieved a lower overall recall (the amount of high quality windows that were predicted as indeed of high quality over the total amount of high quality windows) of 0.527. In bold we can see the best results for each column of the Table 2.

Table 2 – Mean results obtained from Leave-One-Subject-Out protocol. H1 and H2 have very few impact on the metrics, but reduce the number of calls to the learning-based method.

	Accuracy		AUC		F1-Score		Precision		Recall	
	Mean	**STD**	**Mean**	**STD**	**Mean**	**STD**	**Mean**	**STD**	**Mean**	**STD**
CNN1	**0.786**	0.169	**0.815**	0.097	0.806	0.174	0.953	0.059	0.738	0.221
H1(ours)	0.778	0.177	0.812	0.100	0.796	0.184	**0.956**	0.054	0.724	0.231
CNN2	0.644	0.167	0.721	**0.092**	0.646	0.185	0.946	**0.051**	0.519	**0.209**
H2(ours)	0.644	0.167	0.721	**0.092**	0.646	0.185	0.946	**0.051**	0.519	**0.209**
Hao	0.771	**0.120**	0.599	0.111	**0.809**	**0.139**	0.812	0.123	**0.847**	0.213

The discriminator did not impact significantly on the metrics despite having indeed pre-eliminated noisy PPG windows. That can be explained by the relatively small number of discarded PPG windows compared to the total amount of windows. The impact of the discriminator was noticed to be more relevant when it comes to execution time and energy consumption which will be discussed further.



(a) Accuracy

(b) AUC

(c) F1-score

(d) Precision

(e) Recall

Figure 11 – Boxplots of each metric calculated from the second train-test methodology.

The second cross-validation methodology employs a traditional fixed train-test (FX-TT) split, where the split is performed by separating subjects for training and testing. Out of the 56 subjects, 45 were used for training and 11 were separated for testing. The full pipeline of each model was executed 315 times because of the random variations of the training performance while using the GPU, and the results are depicted in Figure 11. The first point to take into consideration is that, as Hao's model is decision-rule-based in its entirety, the results do not vary from one execution to another. Because of that, it was not included in the variance analysis.

We noticed that, for the majority of the metrics, both CNNs as well as their hybrid versions performed similarly, with a slight advantage to the CNN1 architecture, except for the precision metric, where the CNN2 has a considerable advantage as it can be observed in Figure 11-d. Also, the CNN2 architecture behaved in a considerably more stable way, possibly being a more reliable option than CNN1, considering that the test data was collected from different people with consequenlty different PPG waveform patterns.

Table 3 – Results obtained from FX-TT protocol.

	Accuracy		AUC		F1-Score		Precision		Recall	
	Mean	**STD**	**Mean**	**STD**	**Mean**	**STD**	**Mean**	**STD**	**Mean**	**STD**
CNN1	0.898	0.036	0.904	0.022	0.919	0.036	0.959	0.012	0.886	0.066
H1(ours)	**0.899**	0.038	0.905	0.024	0.919	0.039	0.959	0.011	**0.886**	0.068
CNN2	**0.899**	**0.002**	**0.911**	**0.001**	**0.920**	**0.002**	**0.971**	**0.001**	0.875	**0.004**
H2(ours)	**0.899**	**0.002**	**0.911**	**0.001**	**0.920**	**0.002**	**0.971**	**0.001**	0.874	**0.004**
Hao	0.806	n/a	0.819	n/a	0.824	n/a	0.905	n/a	0.757	n/a

As shown in Table 3, CNN2 and its hybrid version, H2, presented the best results on the FX-TT protocol, followed by CNN1 and H1. This corroborates the previous evidence that the discriminator did not have a considerable impact on the training as it was expected. Besides, it was possible to notice its impact when it comes to execution time and energy consumption of the pipeline. The execution time of the discriminator is $5.603\,.\,10^{-5}s$ on a Linux main1 5.4.0-80-generic #90-Ubuntu SMP machine with 16 GB of RAM and AMD EPYC 7742 64-Core Processor. When compared to both CNN1 and CNN2 with an execution time of $0.0355\,s$ and $0.0501\,s$, respectively, the discriminator was responsible for only $0.111 - 0.157\,\%$ of the total time of an inference. This means that, for every PPG window that was eliminated by the discriminator, approximately 99.84 - 99.88% of inference time is saved.

Table 4 – Neural models sizes and their energy consumption.

Model	Size in # of parameters	Size in disk (Kb)	Energy per inference
CNN1	137	35.1	4.92e-05 J
CNN2	9,333	58.1	2.57e-03 J

A total of 25,317 bad segments were discarded by the discriminator, corresponding to 9.96% of all low-quality segments or 3.27% of the total number of segments. These values indicate that only 747,636 out of 772,953 segments would be processed by CNN. That corresponds to 25,317 fewer CNNs' calls (saving 3.27% of calls) resulting on 14.9 – 21.1 minutes saved depending on the considered CNN architecture. Table 4 indicates the energy consumption per inference by each CNN. Because of these results, we estimate that the CNN energy consumption when running on an ARM Cortex-A55 CPU and the proposed two-stage SQC reduced the energy consumption corresponding to an energy saving of 3.27%. This estimate was made using the Keras-Spiking framework (8), assuming a negligible energy consumption of the discriminator, and considering the whole dataset.

Conclusion

This paper proposes a hybrid approach for PPG signal quality assessment, aiming at reducing power consumption, which is highly desirable for constrained devices, like wearables. We propose the introduction of a simple discriminator step, which filters out bad quality signals thus reducing the number of calls to more complex signal quality classification methods. We evaluated two approaches for implementing the discriminator and two learning-based approaches for signal quality classification. Experiments indicate that the proposed approach have negligible influence on the quality metrics, while reducing the estimated energy consumption for PPG signal quality assessment.

References

1. A Quality Assessment System for PPG Waveform. Hao, Jiang, and Gao Bo. 2021, 2021 IEEE 3rd International Conference on Circuits and Systems (ICCS).
2. Optimal signal quality index for photoplethysmogram signals. Elgendi, Mohamed. 2016, Bioengineering 3.4, p. 21.
3. Extending the battery lifetime of wearable sensors with embedded machine learning. Fafoutis, X., Marchegiani, L., Elsts, A., Pope, J., Piechocki, R., & Craddock. 2018, IEEE 4th World Forum on Internet of Things (WF-IoT).

4. A survey of convolutional neural networks: analysis, applications, and prospects. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. 2021, IEEE transactions on neural networks and learning systems.
5. 1-D convolutional neural networks for signal processing applications. Kiranyaz, S., Ince, T., Abdeljaber, O., Avci, O., & Gabbouj, M. 2019, ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
6. Keras-Spiking. www.nengo.ai/keras-spiking. [Online]
7. Cross-validation. Refaeilzadeh, Payam, Lei Tang, and Huan Liu. 2009, Encyclopedia of database systems 5, pp. 532-538.
8. Accelerated visual context classification on a low-power smartwatch. Conti, F., Palossi, D., Andri, R., Magno, M., & Benini, L. 2016, IEEE Transactions on Human-Machine Systems 47.1, pp. 19-30.
9. A novel method for accurate estimation of HRV from smartwatch PPG signals. Bhowmik, Tanmoy, Jishnu Dey, and Vijay Narayan Tiwari. 2017, 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp. 109-112.