Computers in Biology and Medicine 145 (2022) 105430

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep convolutional neural network-based signal quality assessment for photoplethysmogram

Hangsik Shin, Ph.D.

Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, 05505, Republic of Korea

ARTICLE INFO

Keywords:
Convolutional neural network
Deep learning
Photoplethysmogram
Motion artifacts
Pulse quality assessment
Signal quality assessment

ABSTRACT

Quality assessment of bio-signals is important to prevent clinical misdiagnosis. With the introduction of mobile and wearable health care, it is becoming increasingly important to distinguish available signals from noise. The goal of this study was to develop a signal quality assessment technology for photoplethysmogram (PPG) widely used in wearable healthcare. In this study, we developed and verified a deep neural network (DNN)-based signal quality assessment model using about 1.6 million 5-s segment length PPG big data of about 29 GB from the MIMIC III PPG waveform database. The DNN model was implemented through a 1D convolutional neural network (CNN). The number of CNN layers, number of fully connected nodes, dropout rate, batch size, and learning rate of the model were optimized through Bayesian optimization. As a result, 6 CNN layers, 1,546 fully connected layer nodes, 825 batch size, 0.2 dropout rate, and 0.002 learning rate were needed for an optimal model. Performance metrics of the result of classifying waveform quality into 'Good' and 'Bad', the accuracy, specificity, sensitivity, area under the receiver operating curve, and area under the precision–recall curve were 0.978, 0.948, 0.993, 0.985, 0.980, and 0.969, respectively. Additionally, in the case of simulated real-time application, it was confirmed that the proposed signal quality score tracked the decrease in pulse quality well.

1. Introduction

With the development of bio-signal measurement technology, the dream of measuring bio-signals anytime is becoming a reality. Hospitals are providing services that can measure bio-signals in advance. They can use these bio-signals for diagnosis while waiting for medical treatment for emergency patients [1,2]. The number of cases of self-measurement of bio-signals using personal bio-signal measuring devices at home for healthcare is increasing [3]. The spread of health care using wearable devices or home devices is making bio-signal measurement and monitoring a common practice in daily life. In such user-led monitoring, the most important consideration is to secure the reliability of the measured signal. Unlike an environment in which measurement locations and methods are controlled and measurements are performed by professionals, user-led measurement may significantly increase the frequency of occurrence of various measurement errors due to environmental noise, user error, and motion artifacts. Of these, motion artifacts defined as artifacts caused by a change in the contact state of the sensor due to user's movement. The importance of motion artifact mitigation continues to grow due to the increasing outdoor use of PPGs. Noises may significantly degrade the quality of the measured signal, which may cause distortion of the transmitted physiological information, ultimately leading to misdiagnosis. Therefore, with the spread of bio-signal measurement and monitoring into daily life, the importance of signal quality assessment of the measured bio-signal is continuously increasing. Here, signal quality, also called pulse quality in photoplethysmogram (PPG), refers to how much a signal containing noise retains the original information of the signal. One of the most frequently studied bio-signals in signal quality measurement is the photoplethysmography wave usually installed in wrist-type wearable devices. The PPG is a biosignal that non-invasively measures the change in blood volume at the periphery of the human body through absorption, reflection, and scattering characteristics of light in human tissues [4]. The photoplethysmogram is being used for a variety of medical and health management purposes, such as for measuring pulse rate [4], assessing autonomic nerve activity [5,6], analyzing vascular stiffness [7,8], assessing surgical pain [9–11], estimating cuff-free blood pressure [12–15], and measuring oxygen saturation [4,16]. Due to the advantage of providing cardiovascular information simply without burdening the human body, photoplethysmography is being used in many wearable devices, including Apple watch, Fitbit, and Samsung Galaxy watch. In addition, it is used in medical devices such as Watch PAT (Itamar

E-mail address: hangsik.shin@amc.seoul.kr.

https://doi.org/10.1016/j.compbiomed.2022.105430
Received 5 January 2022; Received in revised form 18 February 2022; Accepted 23 February 2022
Available online 22 March 2022

Medical Ltd., Israel), a portable PSG device for measuring and diagnosing snoring or sleep apnea in a home environment, and ViSi Mobile Patient Monitor (Sotera Wireless, Inc., CA, USA), a wearable patient monitor. Various approaches have been introduced to evaluate the quality of photoplethysmography. In most studies, the quality of photoplethysmography waves is manually labeled as 'Good'/'Bad' or 'Acceptable'/Unacceptable'. The quality is automatically classified using various algorithms and the classification accuracy is evaluated. The most traditional method is to evaluate waveform quality through key feature values derived from the waveform or to use template matching. In a study by Orphanidou et al. [17], pulse quality was evaluated with a sensitivity of 91% and a specificity of 95% using heart rate and template matching. Elgandi [18] has compared the pulse quality index based on perfusion, kurtosis, skewness, relative power, non-stationarity, zero crossing, and entropy and revealed that skewness shows the best performance with an F1 score of about 75%.

Vadrevu et al. [19] have classified the quality of 38,620 pulse segments with an accuracy of 97.8% based on features such as global and local maximum amplitude, zero-crossing, and autocorrelation. Reddy et al. [20] have evaluated the quality of the PPG signal with an accuracy of about 93.2% using a classification criterion centered on the absolute value and change of amplitude. In addition, Alam et al. [21] have evaluated the quality of PPG with an accuracy of 96.5% using kurtosis and template-based correlation. There are studies that have evaluated PPG signal quality using pattern recognition or machine learning. Li and Clifford [22] have applied dynamic time warping and multi-layer perceptron (MLP) to classify the quality of PPG with an accuracy of about 95%. Pereira et al. [23] have applied support vector machine (SVM) and shown that waveform quality can be classified with an accuracy of about 95%. More recently, studies using machine learning to evaluate the quality of PPG waveform have been introduced. Roy et al. [24] have evaluated the signal quality with an accuracy of 95.8% based on a self-organizing map using entropy and statistical features. Roh [25] has classified 49,561 pulse segments using a recurrence plot and 2D convolution and shown a sensitivity of about 96.0%, a specificity of 99.0%, and an area under the curve (AUC) of 0.994. Gao et al. [26] have suggested an LSTM-based real-time remote PPG waveform quality assessment method and shown an average accuracy of 79.7%. Results of the PPG quality assessment studies so far suggest the possibility of evaluating the quality of PPG waveform with an accuracy of more than 95% based on various PPG waveform characteristics. However, in most previous studies, it was difficult to generalize the proposed algorithm or secure the stability because they used a small amount of data. In fact, several studies were performed based on only hundreds of minutes of data or thousands of pulses [18,21,22]. Even in studies using relatively many records [17,20,25], only about 30,000–50,000 pulse segments were used for algorithm development and validation. In addition, most previous studies detected features through waveform preprocessing to classify signal quality [17–23]. This feature-based waveform quality classification method is unsuitable for future big data analysis because computational complexity in the preprocessing process may increase and feature detection errors might be linked to waveform quality evaluation errors. The goal of this study was to present a high-performance and reliable model that could overcome limitations of conventional PPG waveform quality evaluation technology. Based on PPG big data, we developed and verified a deep learning model that could omit the preprocessing process and evaluate PPG quality. To this end, we labeled a PPG signal of about 30 GB from the MIMIC III waveform database, developed a convolutional neural network (CNN)-based deep learning model, and evaluated and verified its performance.

2. Materials and methods

2.1. Data

This study was performed using the MIMIC-III Waveform Database. The MIMIC III project was approved by Institutional Review Boards of the Beth Israel Deaconess Medical Center (Boston, MA, USA) and the Massachusetts Institute of Technology (Cambridge, MA, USA). The requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was de-identified. The MIMIC III waveform database consisted of 67,830 records acquired from about 30,000 intensive care unit (ICU) patients. Each record contained the original electrocardiogram, PPG, and arterial blood pressure signals. The MIMIC-III waveform database consisted of records obtained from bedside monitors. It could not identify the patient. It also had delay between signals being measured. Nevertheless, it included signals measured from various types of measurement devices with the advantage of showing measurement results including measurement problems in various actual clinical environments. Data used in this study had record numbers in the range of 3000060–3001002. A total of 458 cases and 28.9 GB of PPG were used.

2.2. Data segmentation and labeling

Before evaluating the quality of the signal, it is necessary to determine the period and frequency of quality evaluation. Since there were no special criteria for selecting the quality evaluation cycle, in this study, 5 s was arbitrarily determined as the quality evaluation cardiac cycle on the premise that at least one cycle was included. Accordingly, segments were generated by dividing the entire PPG signal into data segments having a length of 5 s without overlapping.

Each segment was labeled 'Good' or 'Bad' using an in-house graphical user interface. At this time, the labeling was pre-screened according to the 'Bad' criterion and the incorrectly created label was corrected manually. The segment that satisfied at least one of the following conditions was labeled 'Bad':

Condition 1. The number of zero-crossing exceeded twice the signal length in seconds.

Condition 2. The absolute value of the normalized signal contained a value greater than 2.58 (2-sigma)

Condition 3. The entropy of the normalized signal was greater than 5.

Condition 4. The kurtosis of the normalized frequency spectrum exceeded 2.4.

Initial annotating was performed by a total of four researchers. Cross-validation and final correction of annotation were performed by another two researchers. When annotating was performed, non-uniform beat intervals, parasitic peaks, severe baseline drift due to motion artifacts, signal loss, and other inaccurate beat shapes were classified as 'Bad'. As a result, a segment of about 4.9 GB was classified as 'Bad', which corresponded to about 17% of the total data (4.9/28.9 GB/GB). Fig. 1(a) shows an example of a waveform annotated 'Good', while Fig. 1(b) shows an example of a waveform annotated 'Bad'.

2.3. Machine-learning model

We used a machine-learning model based on CNN, a neural network that could be used to derive results from a convolution operation using a multidimensional kernel [27]. CNN can automatically learn high-level features that capture the structured information and semantic context in the multi-dimensional input data with spatiotemporal connectivity such as image or sequential data [28]. Our CNN model consisted of convolutional layers, followed by a max-pooling layer and fully connected layers. The input layer of the proposed model was $N \times 625$, reflecting N segments and the number of samples in a 5 s segment at sampling rate of 125 Hz.

We performed Bayesian optimization to optimize the number of convolutional layers, dropout rate, and learning rate. Bayesian optimization was performed for the number of convolutional layers of layers of 2–6, dropout rate of 0–0.5, number of nodes for dense layer of 256 to 2,048, batch size of 10 to 1,000, and learning rate of 0.001–0.01. Convolutional layers were designed to have different filters and kernel sizes

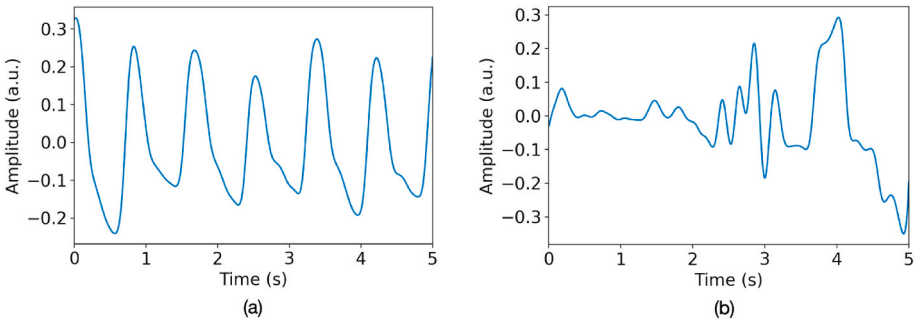

Fig. 1. An example of the quality of the photoplethysmogram waveform. (a) 'Good' quality segment. (b) 'Bad' quality segment.

for each layer. The first convolution layer contained 32 kernels of 10×1 in size. The second convolution layer contained 64 kernels of 6×1 in size. The third convolution layer contained 128 kernels of 4×1 in size. The fourth to sixth convolutional layers equally contained 256 kernels of 2×1 in size. Stride was 2 for all convolutional layers. The convolutional layer was initialized with He initialization [29]. Rectified linear unit (ReLU) was used as an activation function. Batch normalization was adapted after activation. The output of the last convolutional layer went into a fully connected layer by means of flattening. The fully connected layer included dropout [30]. The ReLU was applied as an activation function. This layer was fed into a Softmax function with two class-labeled output. We used the Adam (adaptive moment estimate) [31] for cost optimization in model development with a 0.9 exponential decay rate of the moving average gradient (β_1) and a 0.999 exponential decay rate of the moving average of the squared gradient (β_2. We developed and validated the proposed CNN model with a 3.8 GHz Intel Core i7-8700 processor, 64 GB 1,600 MHz DDR3 RAM, NVIDIA Geforce GTX 3090, and Python 3.6.7, Anaconda, Tensorflow 2.0.

2.4. Optimization: bayesian optimization

A neural network usually called as a 'black-box' in the sense that while it can approximate any functions, studying its structure will not provide any insights on the structure of the function being approximated. Bayesian optimization is often used when optimizing an unknown 'black-box' function. It is applied to solve the hyper parameter tuning problem [32]. In Bayesian optimization, instead of searching all grids, only some parameters are randomly observed. Parameter optimization is then performed based on the random search that selects the best parameter among randomly selected parameters within a specific range. Grid search has few opportunities to try various important parameters because both unimportant and important parameters need to be observed. However, since a random search is not limited to the grid, it has more opportunity to examine important parameters probabilistically. Equation (1) gives the Bayesian optimization formula. In this case, it is assumed that x is a bounded domain and $f(x)$ is a black-box function that does not know what the output is:

$$x^* = \underset{x \in X}{\operatorname{argmin}} f(x) \tag{1}$$

Bayesian Optimization works in the following way:

Process 1. Estimate the function $f(x)$ with a Gaussian process prior using the previously observed data, $D = [(x_1, f(x_1)), (x_2, f(x_2)) \ldots (x_n, f(x_n))]$.

Process 2. The function $f(x)$ is applied as the acquisition function (decision rule) of the next observation point $(x_{n+1}, f(x_{n+1}))$.

Process 3. Add the newly observed $(x_{n+1}, f(x_{n+1}))$ to D and repeat Process 1 until appropriate stopping criteria are reached.

2.5. Validation: classification performance

Five-fold cross validation was used for model generation and validation. This validation method divides the entire data into five groups having the same size. Of these five groups, four are used for model development (development set) and the remaining one group is used for model test (test set) so that data sets used for development and the test set do not overlap. This procedure is repeated five times by changing the test set. Cross validation can improve the reliability of results by allowing all data to be used for training or testing once. In model development, 25% of the data in the development set were used in the validation set to minimize overfitting. As a result, for each fold, 60% of the total data were used in the training set, 20% in the validation set, and 20% in the test set. The average value obtained in 5-fold was then calculated, and considered as the performance of the model. If ratios between data classes constituting the training set, validation set, and test set are not uniform, the model classification result might be biased toward a specific class. This may occur even when there is a significant difference in the amount of data between classes. Therefore, when creating a data set for model development, ratios between classes were kept constant for each training set, validation set, and test set by applying stratified k-fold. In addition, a weight was assigned to each class during learning so that a larger weight could be given when learning through a class with a smaller number of classes. The training was performed using validation loss as a metric. The value in the iteration where the validation loss no longer decreased while the training was repeated 20 times was judged as an optimized value.

2.6. Validation: real-time application

To evaluate the real-time applicability of the finally selected model, validation was performed by simulating real-time data input using a separate dataset that was not used for model development. The data set was arbitrarily selected from records containing both normal and abnormal waveforms. Real-time simulation simulates data real-time input and outputs signal quality evaluation results every second. Finally, the signal quality score was obtained by scaling the Softmax output from (0–1) to a value in the range of 0–100 and moving averaged scaled value of the previous 5 s results.

2.7. Statistical analysis

The performance of the developed model was evaluated in terms of accuracy, sensitivity, specificity, positive predictivity value, and area under the curve. Accuracy is the rate at which the classifier correctly judges the case as 'Good' (negative) or 'Bad' (positive). Sensitivity is the rate at which the classifier correctly classifies the case where the actual signal quality is 'Bad'. Specificity means the rate at which the classifier correctly classifies the case where the signal quality is 'Good'. Positive predictivity value refers to the rate at which the actual quality is 'Bad' among cases where the classifier determines that the signal quality is 'Bad'. To evaluate the overall performance, we derived the receiver operating characteristic (ROC) curve and precision–recall curve. The area under the receiver operating characteristic (AUROC) is a

performance metric for evaluating binary classification models. The AUROC is calculated as the area under the ROC curve. The ROC curve shows the trade-off between the true positive rate and the false positive rate across different decision thresholds. The precision–recall curve shows the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds. The area under the precision–recall curve (AUPRC) is also a useful performance metric for imbalanced data in a problem setting, where the researcher is concerned about finding positive examples.

3. Results

3.1. Machine learning model

Table 1 shows the parameter list and the search range to be optimized through Bayesian optimization with finally selected values. Bayesian optimization was repeated for a total of 40 times. The average parameter value from the top three results showing the best AUC was calculated and determined as the optimal parameter value. As a result of optimization, the proposed model was confirmed to have the best performance in the CNN architecture of 6-layer with the number of fully connected nodes of 1,546, batch size of 825, learning rate of 0.002, and dropout rate of 0.2. Fig. 2 shows the structure of the CNN model optimized through Bayesian optimization.

3.2. Classification performances

Table 2 shows the confusion matrix that classifies 'Good' or 'Bad' using the optimized model. The proposed model correctly classified 1,558,795 (97.8%) out of 1,594,014 segments and incorrectly classified 35,219 (2.2%) segments. Table 3 shows model performance evaluation metrics such as accuracy, sensitivity, predictivity, positive predictivity value, AUROC, and AUPRC derived from these results. Fig. 3(a) & (b) show ROC curve and the precision–recall curve of the model, respectively. In an AUROC curve (Fig. 3(a)), for an ideal model, the true positive ratio is 1 (top of the plot) and the false positive ratio is 0 (left of the plot). This highlights that the best possible classifier that achieves perfect skill is the top-left of the plot (coordinate 0,1). A skillful model is represented by a curve that bows towards a coordinate of (1,1). However, in an AUPRC curve (Fig. 3(b)), an ideal model has precision of 1.0 (top of the plot) and false positive ratio of 1.0 (left of the plot). A skillful model is represented by a curve that bows towards a coordinate of (1,1). In our results, the AUROC curve (Fig. 3(a)) and the AUPRC curve (Fig. 3 (b)) extremely curved to the upper left corner and the upper right corner, respectively, meaning that the model performance was excellent. As shown in Table 3, the developed model had an accuracy of 0.978, a sensitivity of 0.948, a specificity of 0.993, a positive predictivity value of 0.985, an AUROC of 0.980, and an AUPRC of 0.969, demonstrating very good performance in the ROC curve and the precision–recall curve.

Table 1
List of parameters and search ranges optimized by Bayesian optimization with optimized values (Bayesian optimization was performed a total of 40 times. The optimized value was obtained by averaging the top three parameter values when the maximum performance was achieved based on the area under the receiver operating curve (AUROC)).

Parameter	Search range	Optimized value
# CNN layers	[2,6]	6
# Fully-connected nodes	[256, 2,048]	1,546
Batch size	[0,1,10]	825
Dropout rate	[0, 0.5]	0.2
Learning rate	[0.001, 0.01]	0.002

3.3. Real-time signal quality assessment

Fig. 4 shows real-time signal quality assessment results for some sections of a random record. In Fig. 4, the shade means the signal quality's 'Bad' section assessed by an experienced researcher. The results confirmed that in the section with good signal quality, the signal quality score was high. However, at the interval where the waveform was distorted, the signal quality score was significantly reduced.

These results show that the signal quality assessed by the proposed method could be an alternative of the conventional manual signal quality assessment method.

4. Discussion

In this study, we evaluated the quality of photoplethysmogram without detecting feature points of the photoplethysmogram waveform based on an artificial intelligence signal quality assessment model developed using about 30 GB big data and validated the performance of the developed model. As a result, the developed model showed a very high performance (AUC = 0.980) in signal quality assessment of 5 s length PPG segments. Existing waveform quality evaluation studies have focused on evaluating the waveform quality based on characteristics of detected waveform features. However, to detect characteristics of a waveform, good waveform quality is necessary, which is the" begging the question" problem because "a good quality waveform is required as a prerequisite for evaluating the quality of the waveform". In addition, since these methods critically depend on the accuracy of a feature detection other than the waveform quality assessment method, assessment results may vary depending on external factors other than the actual waveform quality. Because the original waveform is used as is to evaluate the quality of the waveform without an additional preprocessing process, the waveform quality assessment without feature detection proposed in this study has the advantage of preventing disturbance due to external factors in the waveform quality evaluation process. Moreover, the signal quality assessment omitting the feature detection process can reduce the time and labor required to manually detect the feature. The approach of analyzing the original waveform using 'deep learning' for signal quality assessment also has the potential to apply 'unknown' or 'hidden features' that have not been used in existing rule-based waveform quality assessment studies. In the case of a photoplethysmogram wave, it basically has an amplitude of arbitrary unit. The amplitude and shape of the acquired signal can vary according to measurement and environmental conditions. Therefore, deriving results using data obtained under various conditions is an effective technique to reduce deviation caused by variability of the photoplethysmography. In addition, unlike the previous study, where only about 50,000 pulse segments were used, in this study, 459 cases, over 130,000 min, and about 30 GB of big data were used. Therefore, it is thought that the reliability of the proposed technology could be guaranteed.

Results of this study are expected to be applied to devices that measure PPG to detect error sections before extracting information from PPG signals and excluding error sections from analysis, or to save power consumption by temporarily stopping measurement in a noisy environment. However, as this study was about a model for classifying segments not temporally continuous, verification of whether the continuity of the trend of change could be maintained in continuous segments (that is, whether discontinuous results would not appear in continuous segments) has not been sufficiently performed. In addition, since the subject information of the data used in this study is not identified, results of the study might be affected by subject's age, sex, disease characteristics, and so on. Thus, model development and verification using more data to generalize this is required. In addition, in this study, the segment length for learning and classification was set to 5 s to compare model performance with previous studies. However, in future studies, the quality assessment of waveforms longer or shorter than 5 s should be

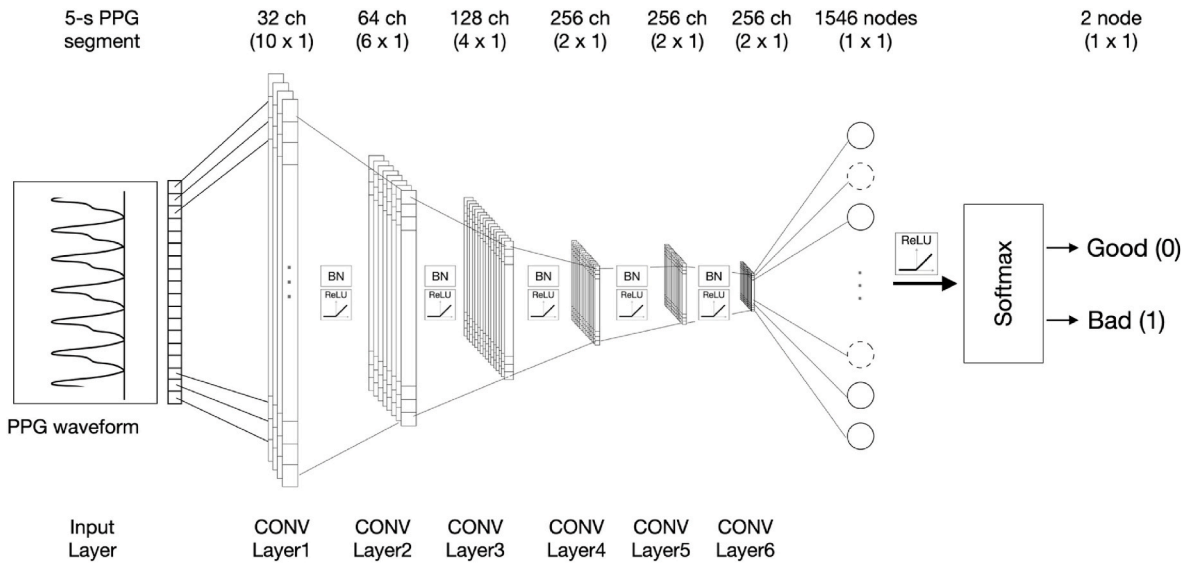

Fig. 2. Architecture of the proposed convolutional neural network (CNN) for photoplethysmogram waveform quality assessment. BN: batch normalization; ReLU: Rectified Linear Unit.

Table 2
Confusion matrix classifying 'Good' or 'Bad' using the optimized model.

Confusion Matrix		Predicted		Total
		Bad	Good	
Actual	Bad	499,942	7,601	507,543
	Good	27,618	1,058,853	1,086,471
Total		527,560	1,066,454	1,594,014

Table 3
Classification performance of the proposed signal quality assessment model by each performance metric in the test set (N = 1,594,014, 5-s length segment; AC: Accuracy; SE: Sensitivity; SP: Specificity; PPV: Positive Predictivity Value; AUROC: Area Under the Receiver Operating Characteristic Curve; AUPRC: Area Under the Precision–Recall Curve).

AC	SE	SP	PPV	F1 score	AUROC	AUPRC
0.978	0.948	0.993	0.985	0.969	0.980	0.969

considered in terms of performance and usability optimization. To minimize the delay in real-time application, it is advantageous to evaluate the quality with the shortest segment. Therefore, it is necessary to consider the minimum analysis interval length that can assess the waveform quality without performance degradation. As an approach to overcome the above-mentioned limitations and improve performance, the addition of a bi-directional layer that combines temporally before and after segments based on a model that assesses waveform quality for each segment could be considered. Moreover, real-time verification of the developed waveform assessment method should be performed. Furthermore, for continuous improvement of waveform quality assessment technology, it is necessary to develop an online learning platform that uses the developed algorithm to perform initial screening of waveform quality and then continuously update the model with results corrected by the researcher.

Table 4 shows performance of the signal quality assessment method proposed in this study in comparison with results of previous studies. It is difficult to specify the state of art technology of the PPG signal quality assessment algorithm because there is no standardized dataset. In addition, evaluation criteria are different for each researcher. Annotation such as 'good' and 'bad' also depends on the subjective judgment of the researcher. For example, annotation for a 5-s length segment may vary depending on which good segment criterion is used among various cases such as "when all pulsations in the segment are 'good'" or "when more than half of the pulsations are 'good' ". On the other hand, when annotating the analysis interval with a single pulse, unlike annotating

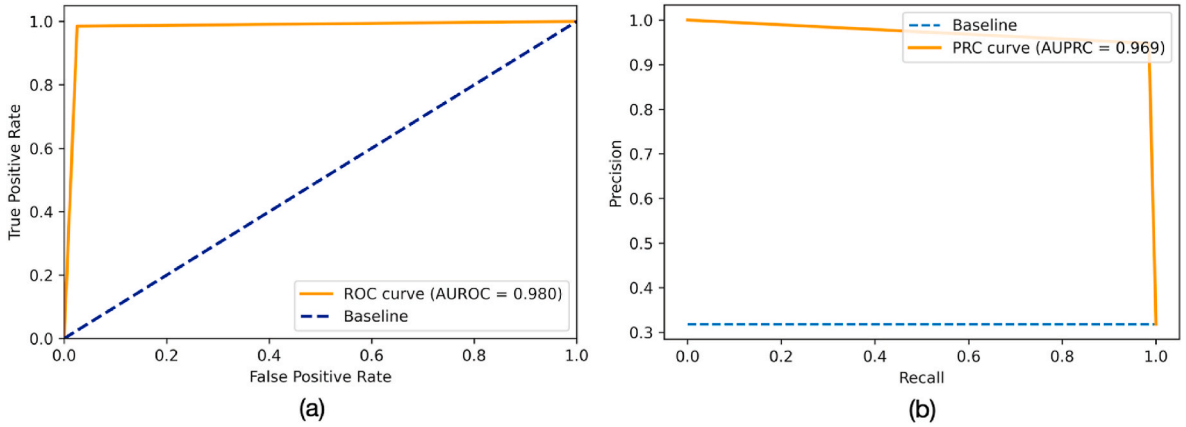

Fig. 3. Performance curves: (a) Receiver operating characteristic (ROC) curve, (b) Precision–recall curve. AUROC: area under the receiver operating curve; AUPRC: area under the precision–recall curve.

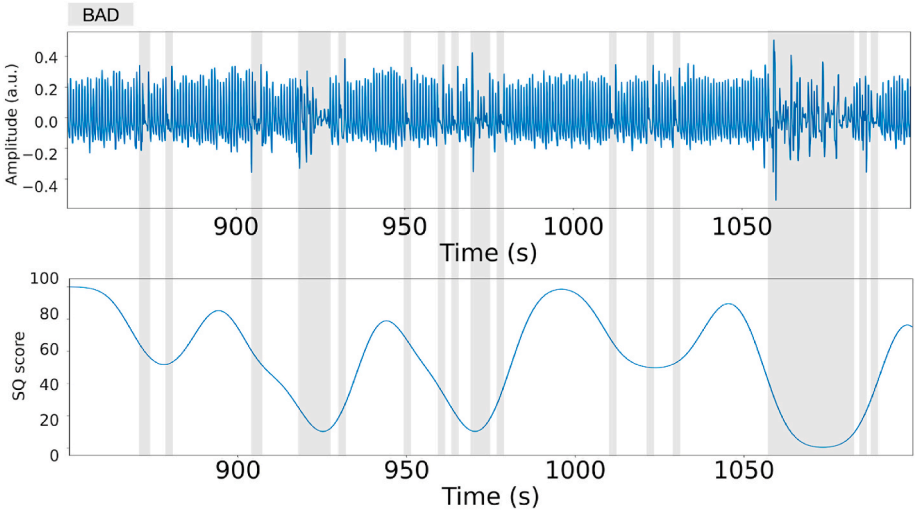

Fig. 4. An example of real-time signal quality assessment. Original photoplethysmogram waveform including noise (top) is shown. Change of signal quality score is calculated by multiplying the probability of 'Good' by 100 (bottom). Shade means 'Bad' interval manually annotated.

Table 4
Comparison of signal quality performance evaluation of this study with previous studies (AC: Accuracy; SE: Specificity; SP: Specificity; PPV: Positive Predictivity Value; AUROC: Area Under the Receiver Operating Characteristic Curve, MLP: Multi-layer perceptron, SVM: Support vector machine, CNN: Convolutional neural network, DCNN: Deep CNN).

Study	Method	Amount of sample	Analysis interval	Performance metrics					
				AC	SE	SP	PPV	F1 score	AUROC
Elgandi [18]	Feature based	106	60-s records	–	0.806	–	0.691	0.744	–
Vadrevu and Manikandan [19]	Decision rule based	38,620	5-s segments	0.978	0.993	0.953	–	–	–
Reddy [20]	Decision rule based	30,000	5-s segments	0.932	0.982	0.907	–	–	–
Alam [21]	Feature based	9,200	Pulse	0.965			–	–	–
Li and Clifford [22]	MLP	1,055	Pulse	0.952	0.990	0.806	0.952	–	–
Pereira [23]	SVM	(finger) 28,667 (radial) 12,819	30-s segments	(finger) 0.948 (radial) 0.959	–	–	–	–	–
Roh and Shin [25]	2D CNN with recurrence plot	49,561	Pulse	0.975	0.964	0.987	0.848	–	0.994
The Proposed	DCNN	1,594,003	5-s segments	0.978	0.948	0.993	0.985	0.969	0.980

the analysis interval by time interval, there is no annotation ambiguity in the analysis interval. Therefore, 'good' pulses and 'bad' pulses might be mixed within a 5-s interval. However, in the case of a single pulse, there is no such ambiguity. Thus, higher accuracy can be expected compared to when classification is performed based on the interval. Consequently, it is difficult to accurately compare the accuracy of signal quality assessment techniques unless a standard data set is used. From this point of view, it can be said that it is important to secure the stability of the technology by using a large amount of data set while maintaining a high level of accuracy of PPG signal quality assessment. Results of this study showed excellent performance and stability that were better than those of existing signal quality assessment studies. The accuracy presented in this study was 97.8%, which was equivalent to the result of a previous study [19] that showed the highest accuracy using a 5-s interval. However, the present study has an advantage in that it has been verified using 50 times more big data compared to the previous study, thus securing the stability and versatility of the algorithm. On the other hand, the AUC was somewhat lower than that in a previous study [25], which showed the best performance among studies that assessed signal quality by dividing the waveform into individual pulses. However, as described above, when evaluating signal quality with segment including multiple pulses, that the performance might be underestimated. Therefore, the method proposed in this study is the most stable one with a high-performance method among 5-s segment-based PPG signal quality assessment methods presented so far. Thus, the method proposed in this study could be considered as the most reasonable signal quality assessment method.

5. Conclusion

As PPG measurement becomes popular not only in hospitals, but also in everyday environments, it is becoming increasingly important to distinguish between PPG signals and noise, especially motion artifacts. In this respect, our study on the assessment of PPG signal quality showed the possibility of discriminating between noise and clean signals from measured PPG with a high accuracy. The deep CNN-based PPG waveform quality assessment technology developed and verified in this study classified the quality of a 5-s length PPG segment into 'Good' or 'Bad' and showed better performance than previous methods, with an accuracy of 97.8% and an AUC of 0.980 despite excluding complex preprocessing process such as peak detection. In addition, this result could be said to have high reliability as it was developed using PPG big data of about 30 GB, much larger than those in previous studies. The developed algorithm can be applied to a PPG measurement device. It could be used for pre-screening regardless whether data are available. It could be used before providing analyzed results to the user or to reduce false alarms caused by waveform distortion. Results of this study have been verified offline. They are expected to be applied to clinical medical devices or personal healthcare devices through online verification by clinical trials and/or computer simulations.

Declaration of competing interest

None declared.

Acknowledgment

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF- 2018R1D1A3B07046442), Republic of Korea, and supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, South Korea, (HI21C0011).

References

[1] D. Curtis, E. Shih, J. Waterman, J. Guttag, J. Bailey, T. Stair, R.A. Greenes, L. Ohno-Machado, Physiological signal monitoring in the waiting areas of an emergency room, in: Proceedings of the ICST 3rd International Conference on Body Area Networks, 2008, pp. 1–8.
[2] P. Hubner, A. Schober, F. Sterz, P. Stratil, C. Wallmueller, C. Testori, D. Grassmann, N. Lebl, I. Ohrenberger, H. Herkner, Surveillance of patients in the waiting area of the department of emergency medicine, Medicine 94 (2015).
[3] C.E. King, M. Sarrafzadeh, A survey of smartwatches in remote health monitoring, J. Healthcare Informatics Res. 2 (2018) 1–24.
[4] J. Allen, Photoplethysmography and its application in clinical physiological measurement, Physiol. Meas. 28 (2007) R1.
[5] E. Gil, M. Orini, R. Bailon, J.M. Vergara, L. Mainardi, P. Laguna, Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions, Physiol. Meas. 31 (2010) 1271.
[6] A. Schäfer, J. Vagedes, How accurate is pulse rate variability as an estimate of heart rate variability?: a review on studies comparing photoplethysmographic technology with an electrocardiogram, Int. J. Cardiol. 166 (2013) 15–29.
[7] S.C. Millasseau, R. Kelly, J. Ritter, P. Chowienczyk, Determination of age-related increases in large artery stiffness by digital pulse contour analysis, Clin. Sci. 103 (2002) 371–377.
[8] K. Takazawa, N. Tanaka, M. Fujita, O. Matsuoka, T. Saiki, M. Aikawa, S. Tamura, C. Ibukiyama, Assessment of vasoactive agents and vascular aging by the second derivative of photoplethysmogram waveform, Hypertension 32 (1998) 365–370.
[9] B.-M. Choi, J.Y. Yim, H. Shin, G. Noh, Novel analgesic index for postoperative pain assessment based on a photoplethysmographic spectrogram and convolutional neural network: observational study, J. Med. Internet Res. 23 (2021), e23920.
[10] M. Huiku, K. Uutela, M. Van Gils, I. Korhonen, M. Kymäläinen, P. Meriläinen, M. Paloheimo, M. Rantanen, P. Takala, H. Viertiö-Oja, Assessment of surgical stress during general anaesthesia, Br. J. Anaesth. 98 (2007) 447–455.
[11] Y.L. Yang, H.S. Seok, G.-J. Noh, B.-M. Choi, H. Shin, Postoperative pain assessment indices based on photoplethysmography waveform analysis, Front. Physiol. 9 (2018) 1199.
[12] C. El-Hajj, P.A. Kyriacou, Deep learning models for cuffless blood pressure monitoring from PPG signals using attention mechanism, Biomed. Signal Process Control 65 (2021) 102301.
[13] M. Hosanee, G. Chan, K. Welykholowa, R. Cooper, P.A. Kyriacou, D. Zheng, J. Allen, D. Abbott, C. Menon, N.H. Lovell, Cuffless single-site photoplethysmography for blood pressure monitoring, J. Clin. Med. 9 (2020) 723.
[14] S.G. Khalid, H. Liu, T. Zia, J. Zhang, F. Chen, D. Zheng, Cuffless blood pressure estimation using single channel photoplethysmography: a two-step method, IEEE Access 8 (2020) 58146–58154.
[15] G. Wang, M. Atef, Y. Lian, Towards a continuous non-invasive cuffless blood pressure monitoring system using PPG: systems and circuits review, IEEE Circ. Syst. Mag. 18 (2018) 6–26.
[16] K.K. Tremper, Pulse oximetry, Chest 95 (1989) 713–715.
[17] C. Orphanidou, T. Bonnici, P. Charlton, D. Clifton, D. Vallance, L. Tarassenko, Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring, IEEE J. Biomed. Health Informat. 19 (2014) 832–838.
[18] M. Elgendi, Optimal signal quality index for photoplethysmogram signals, Bioengineering 3 (2016) 21.
[19] S. Vadrevu, M.S. Manikandan, Real-time PPG signal quality assessment system for improving battery life and false alarms, IEEE Trans. Circuits Syst. II: Express Briefs 66 (2019) 1910–1914.
[20] G.N.K. Reddy, M.S. Manikandan, N.N. Murty, On-device integrated ppg quality assessment and sensor disconnection/saturation detection system for IoT health monitoring, IEEE Trans. Instrum. Meas. 69 (2020) 6351–6361.
[21] S. Alam, R. Gupta, K.D. Sharma, On-board signal quality assessment guided compression of photoplethysmogram for personal health monitoring, IEEE Trans. Instrum. Meas. 70 (2021) 1–9.
[22] Q. Li, G.D. Clifford, Dynamic time warping and machine learning for signal quality assessment of pulsatile signals, Physiol. Meas. 33 (2012) 1491.
[23] T. Pereira, K. Gadhoumi, M. Ma, X. Liu, R. Xiao, R.A. Colorado, K.J. Keenan, K. Meisel, X. Hu, A supervised approach to robust photoplethysmography quality assessment, IEEE J. Biomed. Health Informat. 24 (2019) 649–657.
[24] M.S. Roy, R. Gupta, K.D. Sharma, Photoplethysmogram Signal Quality Evaluation by Unsupervised Learning Approach, 2020, IEEE Applied Signal Processing Conference (ASPCON), IEEE, 2020, pp. 6–10.
[25] D. Roh, H. Shin, Recurrence plot and machine learning for signal quality assessment of photoplethysmogram in mobile environment, Sensors 21 (2021) 2188.
[26] H. Gao, X. Wu, C. Shi, Q. Gao, J. Geng, A LSTM-based realtime signal quality assessment for photoplethysmogram and remote photoplethysmogram, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3831–3840.
[27] K. O'Shea, R. Nash, An Introduction to Convolutional Neural Networks, 2015 arXiv preprint arXiv:1511.08458.
[28] H. Li, J. Chen, H. Lu, Z. Chi, CNN for saliency detection with low-level feature integration, Neurocomputing 226 (2017) 212–220.
[29] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[31] D.P. Kingma, J. Ba, Adam, A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.
[32] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. (2012) 25.