HHS Public Access

Author manuscript

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2022 October 11.

Published in final edited form as:

IEEE J Biomed Health Inform. 2020 March ; 24(3): 649–657. doi:10.1109/JBHI.2019.2909065.

A Supervised Approach to Robust Photoplethysmography Quality Assessment

Tania Pereira,
Department of Physiological Nursing and Center for Physiologic Research, University of California San Francisco, San Francisco, CA 94117-1080 USA

Kais Gadhoumi,
Department of Physiological Nursing and Center for Physiologic Research, University of California San Francisco, San Francisco, CA 94117-1080 USA

Mitchell Ma,
Department of Physiological Nursing and Center for Physiologic Research, University of California San Francisco, San Francisco, CA 94117-1080 USA

Xiuyun Liu,
Department of Physiological Nursing and Center for Physiologic Research, University of California San Francisco, San Francisco, CA 94117-1080 USA

Ran Xiao,
Department of Physiological Nursing and Center for Physiologic Research, University of California San Francisco, San Francisco, CA 94117-1080 USA

Rene A. Colorado,
Department of Neurology School of Medicine, University of California San Francisco, San Francisco, CA 94117-1080 USA

Kevin J. Keenan,
Department of Neurology School of Medicine, University of California San Francisco, San Francisco, CA 94117-1080 USA

Karl Meisel,
Department of Neurology School of Medicine, University of California San Francisco, San Francisco, CA 94117-1080 USA

Xiao Hu
Department of Physiological Nursing, Center for Physiologic Research, Department of Neurological Surgery, Affiliate Faculty of ICHS, University of California San Francisco, San Francisco, CA 94117-1080 USA

Abstract

Corresponding author: Tania Pereira. tania.pereira@ucsf.edu.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Early detection of Atrial Fibrillation (AFib) is crucial to prevent stroke recurrence. New tools for monitoring cardiac rhythm are important for risk stratification and stroke prevention. As many of new approaches to long-term AFib detection are now based on photoplethysmogram (PPG) recordings from wearable devices, ensuring high PPG signal-to-noise ratios is a fundamental requirement for a robust detection of AFib episodes. Traditionally, signal quality assessment is often based on the evaluation of similarity between pulses to derive signal quality indices. There are limitations to using this approach for accurate assessment of PPG quality in the presence of arrhythmia, as in the case of AFib, mainly due to substantial changes in pulse morphology. In this paper, we first tested the performance of algorithms selected from a body of studies on PPG quality assessment using a dataset of PPG recordings from patients with AFib. We then propose machine learning approaches for PPG quality assessment in 30-s segments of PPG recording from 13 stroke patients admitted to the University of California San Francisco (UCSF) neuro intensive care unit and another dataset of 3764 patients from one of the five UCSF general intensive care units. We used data acquired from two systems, fingertip PPG (fPPG) from a bedside monitor system, and radial PPG (rPPG) measured using a wearable commercial wristband. We compared various supervised machine learning techniques including k-nearest neighbors, decisions trees, and a two-class support vector machine (SVM). SVM provided the best performance. fPPG signals were used to build the model and achieved 0.9477 accuracy when tested on the data from the fPPG exclusive to the test set, and 0.9589 accuracy when tested on the rPPG data.

Keywords

Pulse Oximetry; Signal Quality; Supervised Machine Learning; Atrial Fibrillation; Stroke; Annotated Data

I. Introduction

ATRIAL fibrillation is the most common arrhythmia, with a prevalence of 3% among adults over 20 years of age [1] and of 30% among all stroke patients [2]. Patients with AFib face a five-fold increase risk of stroke [3]. Stroke patients with AFib have been shown to have a worse neurological outcome than stroke patients without AFib [4]. AFib can occur paroxysmally (intermittently) and briefly. The AFib diagnosis requires long term continuous electrocardiogram (ECG) and is defined as an irregularly irregular rhythm with no discernible P waves, during 30-sec or more [1]. AFib detection allows for effective prevention of secondary stroke. Therefore, efforts to develop novel approaches for AFib detection and long term monitoring is of great importance in primary and secondary stroke prevention [5]. Recent approaches for AFib detection are based on wearable and mobile solutions [5]–[8]. PPG is an optical method for non-invasively measuring changes in blood volume in the microvascular bed of well-perfused tissues such as the surface of the finger-tip, the wrist, the ear-lobe, and the forehead [9]. However, interference-free and clean PPG signals are difficult to acquire and pose a major challenge in real world applications. Signal integrity is crucial for identifying pathological abnormalities and avoiding false detections. The definition of good quality PPG is not straightforward and requires taking into consideration contextual factors, for wearables solutions the accelerometry have been

used to identify motion artifacts [10]. Synchronized ECG and PPG have been used in some studies to confidently distinguish physiological contaminated PPG segments [11].

Several studies have proposed algorithms dedicated to quality assessment of PPG and similar pulsatile signals. S. Asgari *et al.* [12] proposed a signal quality index (SQI) for the assessment of arterial blood pressure (ABP) signal quality based on singular value decomposition. The algorithm was validated in 1336 10-sec segments (18472 beats) and achieved a true positive rate of 99.06% and a false positive rate of 7.69%. A. Sukor *et al.* [10] proposed an SQI based on waveform morphology analysis and reported an accuracy of 83 ± 11% using 104 60-sec PPG segments. G. Clifford *et al.* [13] introduced dynamic time warping to stretch each beat to match a running template and combined with correlation and the percentage of the beat that appeared to be clipped, and then used these features in a multi-layer perceptron neural network to learn the relationships between the parameters in the presence of good and bad quality pulses. Using a database with 1055 6-sec segments of PPG they obtained an accuracy of 95.2%. W. Karlen *et al.* [11] estimated signal quality based on cross-correlation of consecutive pulse segments and reported a sensitivity of 96.21%. C. Orphanidou *et al.* [14] developed an SQI based on a sequence of several thresholds for heart rate, inter-beat intervals, and template matching correlation, which achieved a sensitivity of 91% and a specificity 95%. C. Liu *et al.* [15] presented a method for assessing the ABP signal quality based on Gaussian template matching (four typical pulse templates were generated and each template consisted of three positive Gaussian functions), which achieved 90.79% accuracy. These signal quality assessment approaches either compare the similarity between consecutive beats or use a static evaluator algorithm that relies on thresholds derived from “common-sense” physiology. These methods showed good accuracy in normal subjects and in some cases with arrhythmia, but were not tested on data from patients with AFib [11], [14], [16]. PPG recordings with AFib episodes are marked by apparent morphological discrepancies between consecutive pulses. This loss of pulse integrity due to AFib often hampers the performance of available quality assessment algorithms. An example of morphological variability in AFib pulses is shown in Fig. 1(c). Fig. 1(a) represents a good PPG for normal sinus rhythm (NSR); Fig. 1(b) shows a bad PPG signal for NSR, and Fig. 1(d) corresponds to a bad PPG signal for AFib. Co-recorded ECG signals provide a physiological context that confirms the observed variability in PPG is physiological and not artifactual. A delay between the two signals reflects the pulse arrival time.

The main objective of this work is to develop an algorithm for PPG quality assessment that ensures that irregular signals due to AFib are not misidentified as poor quality signals. Machine learning techniques allow new approaches to address this challenge. Prior studies used support vector machine to distinguish between good and bad PPG signals with high accuracy; however, these studies used only small datasets, and datasets that did not contain AFib episodes [16]–[20]. We propose supervised machine learning based techniques for robust PPG quality assessment using spectral and temporal features and compare these approaches to a set of classifiers based on K-nearest neighbors, decisions trees, and two-class SVM. Signals from two different groups of patients were used: stroke patients and general intensive care unit (ICU) patients from the UCSF Medical Center. Fingertip PPG (fPPG) and ECG signals from the bedside monitor and simultaneous radial PPG (rPPG) and

accelerometer signals from the E4 system (Empatica, Italy) were acquired from 13 stroke patients. For the 3764 general ICU patients, fPPG and ECG from their bedside monitor were acquired. The data from the fPPG were used to train the model for the PPG quality assessment, and the model was tested in an independent test set from fPPG and data from the rPPG watch. This test allowed us to evaluate the performance of the model using PPG measured by different wearable systems in different areas of the body. This study design tested whether the model is independent of the measurement device and if it can be applied to the other wearables [21].

II. Materials and Methods

A. Data Collection and Study Population

PPG waveform data from a wearable device and pulse oximeter were collected prospectively or retrospectively in inpatients admitted to UCSF Medical Center. In a first cohort of patients, PPG waveform recordings from fingertip pulse oximeters and from a wrist wearable device (E4, Empatica Inc) were prospectively collected from 13 stroke patients (age range 19 to 91 years, median = 73.5) admitted to the Neuro ICU between October 2016 and January 2018. These patients were diagnosed with acute ischemic stroke, were at least 18 years old, and spoke English. Patients with significant problems related to their attention, or alertness, cognitive function, or who had an inability to communicate were excluded unless a legally authorized representative could consent on their behalf. All enrolled patients provided written informed consent to protocols approved by UCSF's Institutional Review Board. From these 13 patients, we created two dataset: group A1, with PPG data from the bedside monitor devices (fPPG); group A2, with PPG data acquired in the radial (rPPG) from the E4 wrist band (Fig. 2). Between 3 h–22 h of continuous PPG recordings (median = 10.5 h) were collected in stroke patients (group A1 and A2). Eight of 13 patients had AFib episodes as documented by the clinicians at the time of recording.

In group B, four randomly selected 30-sec segments were extracted per each patient from 3764 patients admitted to one of the five general intensive care units between March 2013 and December 2016 as described in a previous study [22]. To determine the number of patients under AFib from this group, we selected the patients assigned with ICD9 or ICD10 codes for AFib (ICD9 Diagnosis Code 427.31; ICD10 Diagnosis Code I48.91 Unspecified atrial fibrillation, I48.0 Paroxysmal atrial fibrillation, I48.1 Persistent atrial fibrillation and I48.2 Chronic atrial fibrillation). In total, we identified 1072 patients that correspond to 28% of the 3764 ICU patients. Due to low accuracy of the assignment ICD9 and ICD10 codes, we selected from the previous list the patients who were medicated with anticoagulants indicated for atrial fibrillation treatment: warfarin, apixaban, rivaroxaban, and dabigatran. 500 patients were final identified and they represent 13% of the 3764 patients from the general ICU. The number of AFib cases identified in the general ICU is in line with values reported in the literature reported with AFib cases in a range between 6% to 26% in adult medical ICUs [23].

ECG recordings available from BedMasterEx (Excel Medical Inc, USA) were collected for all patients. ECG waveforms and fPPG waveform recordings were sampled at 240 Hz while rPPG recordings from E4 devices were sampled at 64 Hz. Each recording was split into 30

s-long non-overlapping strip segments for analysis. Data were normalized between zero and one across both modalities to adjust for difference of gain between PPG modalities. rPPG signals were upsampled at 240 Hz to match the sampling frequency of fPPG signals.

B. Annotation Process

In order to create a gold standard of signal quality assessment, we used a web application developed in-house to simplify the human annotation task and reliably store labels. Two annotation projects were used, one for the evaluation of PPG signal quality based on visual assessment by a group of five biomedical signal experts, and another project for AFib annotation by seven clinicians.

1) PPG Signal Quality Assessment: Each PPG segment was labelled as: *Good*, *Bad* or *Not Sure*. Taking into account the physiological context, PPG segment labelling was based on a set of heuristic rules. In order to be labelled as *Good*, a segment had to: 1) reflect the response of blood volume to the underlying pathophysiological characteristics of the cardiovascular system, irrespective of the particular shape of the pulse; 2) show a consistent number of inflection points; 3) be artifact-free and 4) be free of irregular shapes that cannot be explained by ECG changes.

The ECG signal was used for visual inspection during the annotation process to help the annotators to understand if the changes in the PPG waveform were originated during the cardiac cycle. For this reason, the synchronization of the two signals was not a critical point, however, we can see in the Fig. 1 delay between the two signals (PPG and ECG) that is two seconds and it is constant during the acquisition. The synchronization of the acquisition of bedside monitor and E4 was done using a camera that records the time that acquisition with E4 device started, and this time reference was used to select the correspondent start point time of bedside monitor acquisition. This synchronization allows to use the ECG information for E4 data annotation – group A2, and for this group the accelerometer waveforms were also displayed as another source to help the annotation decision. The segments were randomly assigned to the annotators to avoid the bias of evaluating a block of signals from the same patient that would have included some level of similarity among each other. Cohen's kappa was used to assess inter-rater variability using a subset of 100 30-sec segments from collected PPG segments in both groups of patients. Remaining samples were annotated without overlapped entries among annotators, with the exception of rPPG files corresponding to cases labelled as AFib by the clinicians, which were annotated for quality assessment by all five annotators. AFib represents a challenge for quality assessment of rPPG, mainly due to irregularities in pulse waveform and the inherently high number of poor-quality rPPG segments that arise from motion artifacts. In this subset of segments, majority voting was applied to decide the ultimate assessment label.

2) AFib Annotation: Seven clinicians labelled each 30-sec segments using three labels: *AFib*, *Not AFib*, or *Not Sure* based on the guidelines identification [1]. Cohen's kappa was determined for the group of seven clinicians involved in AFib annotation using a subset of 100 30-sec ECG segments.

C. Feature Extraction

Forty-two temporal-domain and spectral-domain features were extracted from each of the 30-sec segments. Time domain statistics obtained were as follows: mean, median, standard deviation, variance, interquartile range, skewness, kurtosis, root mean square, Shannon entropy, and mean and standard deviation of first derivative. Frequency domain statistics obtained were as follows: first- to fourth-order moments in the frequency domain, median frequency, spectral entropy, total spectral power, and peak amplitude in frequency band between 0 to 10 Hz [24]. We analyzed the spectral content of the signal using the periodogram spectral estimation technique. Non-linear features derived by the Poincare plot were used: SD1, standard deviation of the short-term beat-to-beat interval variability; the major axis SD2, the standard deviation of the long-term beat-to-beat interval variability; and the SD1/SD2 ratio. Beat-to-beat analysis was used with four templates based on Gaussian waves to test the cross-correlation with each beat from the 30-sec segment, and the mean, standard deviation and range of the maxima list for cross-correlation results were determined. Beat-to-beat differences were also used and were determined by the interquartile range for the differences of time domain statistics applied to each beat: mean, median, standard deviation, variance, interquartile range, range, skewness, kurtosis, root mean square, sample entropy, Shannon entropy and mean and standard deviation of first derivative of the signal. In beat-to-beat analysis, the mean of area under curve was determined; and the minimum and maximum period of a beat in the segment were used. Due to the large difference in characteristics (amplitude and variation) of the feature components, a normalization procedure was performed by subtracting the mean over all training values and dividing by the corresponding standard deviation [25].

D. Classification

Three machine learning algorithms were trained and tested to classify 30 s-segments into one of two class labels (Good, Bad): Support Vector Machine, K-nearest neighbors and Decision tree. The choice of these algorithms was motivated by their robustness and generalization power in high-dimensional classification problems [16], [26]–[29]. All the algorithms were implemented in Matlab 2017a using the Statistics and Machine Learning Toolbox (Mathworks Inc, USA).

1) Baseline Algorithms: We compared the performance of the proposed approach to existing (baseline) methods. Some of these methods were adapted to the length of the segments in our dataset (30 s): Method 1 (based on A. Sukor's method [10]); Method 2 (based on W. Karlen's method [11]); Method 3 (based on S. Asgari's method [12]); Method 4 (based on G Clifford's method [13]); Method 5 (based on C. Orphanidou's method [14]) and Method 6 (based on C. Liu's method [15]). The performance of baseline algorithms was assessed using data measured by both systems, fPPG and rPPG, and was tested in the entire dataset from stroke patients and in a subset of AFib cases. These tests allow for comparison of performance results with the novel algorithm proposed in this work.

E. Experimental Design

This section consists of the description of several tests implemented in this work to evaluate the classifier performance. We designed three different experiments, each one with specific objective.

The objective in the first experiment (Exp 1 in the Fig. 2) was to build a model using fPPG data from a subset of the stroke patients (9 patients that correspond approximately to 75% of fPPG segments from stroke patients), and show the performance of this model in the unseen patients (fPPG segments from the excluded 4 stroke patients); and on data from another wearable device (rPPG segments from same 4 excluded stroke patients) and on all data from group B. The three classifiers (SVM, KNN, DT) were applied on this experiment to select the best one.

On experiment 2 (represented as Exp 2 in Fig. 2), the objective was to verify if the performance of the model increases by adding fPPG segments from a much larger patient cohort (group B) to the previous training dataset composed by the selected fPPG data from the 9 stroke patients.

The experiment 3 (Exp 3 in the Fig. 2) was dedicated to test the performance using data that were annotated regarding AFib, and in this experiment, we selected a test set of fPPG and rPPG segments labeled as AFib (data from 5 stroke patients were annotated with AFib), we used the fPPG data of the rest of 8 patients to train the model.

To automated hyperparameter tuning, we used a Bayesian optimization to conduct a guided search for the best hyperparameters for each classifier [30]. The selection of the optimal hyperparameters was done by minimizing a ten-fold cross-validation loss, this performance measure is an average of the test error over the 10 trials and gives an estimate of the expected generalization error [31]. For KNN, two hyperparameters were optimized: the number of neighbors and the distance metric. In the DT, we optimize the minimum number of leaf node observations. For SVM, we used the RBF kernel (defined in our previous work [32]) and we optimized the parameters C and sigma [33]. Sigma the scaling factor in the gaussian radial basis function kernel. C is the marginal factor parameter that is a regularization factor between the width of the margin and the total distance of each error from the margin. Various pairs of (C, sigma) values were tried, and for both parameters an exponentially growing sequences between the range [1e-5,1e5] were used. The optimal values for the hyperparameters are provided in a Table I from the Supplemental Material. For test, the performance analysis was conducted considering the accuracy (Accu), F1-score (F1), sensitivity (Sen), specificity (Spe), receiver operating characteristic (ROC) and the area under the ROC curve (AUC).

1) Experiment 1. Training Model Using Data in Group A1: In experiment 1 we applied three classifiers: SVM, KNN and DT. We trained each classifier using dataset data from nine patients randomly selected from the 13 strokes patients (Fig. 3). The data from the four remaining stroke patients were used to test the model. In order to avoid overfitting and to test each model in a prospective setting, approximately 25% (15 to 35%) of the data samples were chosen as a test set, and these samples were never involved in the training

phase. The remaining 75% (65 to 85%) of the data was used for learning the best model and determining the best parameters through 10-fold cross-validation, after being normalized to avoid within-subject differences in amplitude and variation among features. Training samples averaged values for each feature, and corresponding standard deviation values were stored to normalize test feature sets, enabling us to map novel values into the training model features space.

2) Group A2 Test: The model trained with fPPG data group A1 was applied to the rPPG data group A2, and was ensured that rPPG data selected for testing was not from patients used for training (Fig. 3). The rPPG subset used for testing was normalized by the mean and standard deviation of training data from the fPPG group A1.

3) Group B Test: The model was tested with the dataset from general ICU patients (group B), which represents a completely novel dataset and contains an extensive number of different subjects, to evaluate the generalization of model performance.

The train and test datasets for experiment 1 are described in the following scheme (Fig. 3). This experiment allows to selection of the best classifier. Only the classifier selected on this experiment was apply in the following experiments.

4) Experiment 2. Training Model Using Data From Group A1 and B: To estimate the training set size, which is important in developing predictors that operate near their respective plateaus, we examined how the model performance characteristics improved as the training dataset size increased. In order to increase the variability of data used to train the model, we added the fPPG data from group B to the training dataset. We selected a variable percentage (between 10 to 100%) of group B data that was used to train the model. The model performance was evaluated using group A1 and A2 data from the excluded stroke patients. The described tests are depicted in the Fig. 4.

5) Experiment 3. Training Model Using Data From Group A1 and Test in AFib Labelled Cases: The AFib annotation project allowed for identification of AFib episodes from five stroke patients. The remaining eight patients have not yet had their data annotated at the time of this report. However, among these remaining eight patients with unannotated data we know that there were at least three patients with AFib episodes during their rPPG and fPPG recordings (group A1 and A2) based on clinical data from their hospitalization. The eight patients with unannotated data were selected for training and the five patients with annotated data as AFib episodes were used for testing (Fig. 5). This approach ensured that there were enough AFib cases in the training data from segments not yet annotated.

III. Results

A. Agreement Among Annotators

Inter-rater agreement was high among the five annotators of signal quality, with a kappa coefficient of 0.87 for fPPG data (group A1) and 0.83 for rPPG data (group A2). The agreement assessment for the majority voting project, with AFib cases in rPPG data, resulted in a lower kappa coefficient of 0.70, due to the great difficulty in classifying this subset. The

AFib classification project yielded a kappa of 0.64 indicating a moderate agreement among the seven clinicians involved.

B. Signal Quality Annotations

For signal quality assessment, annotators labelled 15824 30-sec PPG segments measured with fPPG from stroke patients – group A1, and 12819 segments from stroke patients measured with rPPG (group A2). Table I summarizes the distribution of the annotations per stroke patient for group A1 and A2. These results show the great difference in signal quality between the two systems, with a large part of good signals from fPPG acquisitions: 10341 segments were labelled as good (65.4%), and 5483 were labelled as bad (34.6%). In the rPPG dataset, 9292 (72.5%) segments were labelled as bad, and 3527 (27.5%) segments were labelled as good. Annotated data was used as the gold standard for the supervised machine learning implemented in this work. In group B, 4 randomly selected 30-sec segments were extracted per each patient from 3764 patients, after excluded the flat data we got a final set of 12843 segments. 8728 (68.0%) segments were labelled as good and 4115 (32.0%) were labelled as bad PPG segments.

C. Performance of Baseline Algorithms

The baseline algorithms were applied using all data from stroke patients and from both systems. The results obtained for the six baseline methods are depicted in Table II.

D. Experiment 1: Classifiers Comparison

Fig. 6 shows the ROC curve, AUC, and accuracy for three classifiers in experiment 1. SVM showed better accuracy in all tests with data from stroke patients (group A1 and A2) and from the general ICU patients (group B). Based on these results, the SVM classifier was selected and applied on the next experiments. Table III contains the results plotted on Fig. 6 for SVM classifier, and shows the nine patients used for training the model from the total dataset of group A1. Results shown in Table III demonstrate consistently good performance (accuracy >0.90 across different groups) independently of patient selection.

E. Experiment 2: Test Performance for Models Trained With Data From Group A1 and B

We gradually increased the number of segments from group B used to train the model and evaluated the model performance for each new training set. However, the learning curve did not show any increase of the accuracy, instead showed a stable accuracy values for the increase in group B data used for training.

F. Experiment 3: Performance for AFib Cases

Using a selected subset of AFib cases (1617 30-sec segments), the performance of the six baseline cases and our model are represented in the Table IV. A clear decrease in the performance results were verified for the six baseline methods, with a specific subset of the AFib episodes from five stroke patients.

IV. Discussion

In this work, we proposed a supervised machine technique for PPG quality assessment. We built a database that consisted of 12843 segments of 30-second finger PPG signals from 3764 patients in general ICU, 15824 segments of finger PPG and 12819 segments of radial PPG signals from 13 stroke patients, whose signal quality were annotated with high inter-rater agreement.

Considering the list of initial classifiers tested in this work (SVM, DT and KNN) all showed good performances with an average accuracy higher than 0.85, likely due to the good feature engineering that captures the important characteristics from the segments relative to their quality. Feature engineering has a great impact in the classifier performance. Despite the good overall results, the SVM classifier showed better results for the three datasets (groups A1, A2, and B) with an accuracy that was higher than 0.90 for group B and higher than 0.93 for group A1 and A2. SVM has been applied in physiological data due to superior performance of SVM in classifying high-dimensional data.

The high inter-rater agreement of finger PPG signal quality annotations was achieved even for cases with mixed signal distortions due to movement artifacts and/or AFib because simultaneously recorded multi-lead ECG signals were presented to the annotators as well. We believe that the higher performance of our SVM-based classifier as compared to existing empirically designed approaches is due to the classifier training with this high quality database. Results observed for baseline algorithms showed that their performance in our dataset is worse than what was reported in the original studies [10]–[15]. We believe the reduction in the performance of these algorithms was caused by having a large number of signal entries with AFib in the test dataset in the present study. These algorithms were evaluated in their original studies with datasets that did not consider the impact of AFib, in particular on the shape of PPG pulses. In addition, a subset of 1617 signal segments were annotated as AFib cases. Using this database, we showed that SVM-based classifier showed better performance than six existing approaches of PPG signal quality assessment when they were applied to process fPPG signals. The degree of inter-rater agreement dropped when the same approach of annotation was applied to rPPG signals.

One of the great challenges of this work was to build a model using fPPG data that could be applied to infer PPG signal quality in wearables. Using data from one single device is a limitation of many of the prior studies in this field since the classification may be limited to signal characteristics of that device [34]. Here, a new approach to leverage the abundance of large in-hospital bedside monitor PPG datasets to build robust models was used. The model trained with bedside monitor data showed good accuracy when tested in data from both systems: fPPG and rPPG.

Using mixed datasets from stroke and general ICU patients did not improve the model performance, suggesting no relevant information gain is obtained by increasing the number of training samples [35]. An alternative explanation is that we only sampled four 30-second segments from each of a large number of ICU patients, hence the resultant 12843 segments not only under-represented richness in this dataset as compared to using a much larger

sample of records from the 13 patients with stroke. On the other hand, the number of segments originally used in the model from stroke patients is approximately 75% of the total segments from stroke patients that is approximately 11868 segments, corresponding to the nine stroke patients selected for the training. If we use all data extracted from general ICU for training, this contribution will be 12843 segments. Ideally, we would have a data set with a size with order of magnitude to add to the training dataset to affect the model performance. Another possible reason for not achieving improved performance is the limitation of the classical machine learning techniques as in the case of SVM. More recent deep learning approaches would gain more substantial performance gain with more data than classical machine learning algorithms do. Therefore, it remains interesting to verify if a much larger sample from the ICU patients, to increase the richness and size of the database of annotated PPG signal, could result in further improvement in classification accuracy by using deep learning techniques.

1) Limitations

We did not have AFib annotation for all records, i.e., we did not know confidently whether signal segments from all the remaining patients had AFib or not. However, eight patients with stroke had clinical documentations of AFib. Therefore, it is very likely AFib was presented in the signal entries of these patients. Furthermore, the number of annotated AFib cases provided enough episodes to prove that our approaches have a good performance even in these most difficult cases.

V. Conclusion

The objective of this work was to develop a method to classify PPG signal quality that can overcome the limitations of previous methods when AFib is present. Here, we presented a set of experimental algorithms to efficiently evaluate PPG signal quality in short 30-sec segments. Two-class SVM with RBF kernel approach demonstrated good and robust performance that could be applied to assess the quality of PPG signal acquired from different devices. The proposed approach is particularly robust to physiological signal irregularities induced by AFib. Our model showed a higher performance even in the particular dataset composed of AFib segments measured with bedside monitor or with a wearable device. The proposed model also demonstrated great generalization with good testing performance on a large dataset composed of a large number of patients from ICUs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the U NIH funds R01NHLBI128679, R18HS022860, and R01GM111378, in part by UCSF Middle-Career Scientist Award, and in part by UCSF Institute of Computational Health Sciences under Grant RAS A127552.

References

[1]. Kirchhof P et al. , "2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS," Eur. Heart J, vol. 37, no. 38, pp. 2893–2962, 2016. [PubMed: 27567408]

[2]. Glotzer TV and Ziegler PD, "Cryptogenic stroke: Is silent atrial fibrillation the culprit?," Heart Rhythm, vol. 12, no. 1, pp. 234–241, 2015. [PubMed: 25285649]

[3]. Kamel H, Okin PM, Elkind MSV, and Iadecola C, "Atrial fibrillation and mechanisms of stroke: Time for a new model," Stroke, vol. 47, no. 3, pp. 895–900, 2016. [PubMed: 26786114]

[4]. Christina Steger MA, Pratter A, Martinek-Bregel M, Andreas Valentin CS, and Slany J, "Stroke patients with atrial fibrillation have a worse prognosis than patients without: data from the Austrian stroke registry," Eur. Heart J, vol. 25, no. 19, pp. 1734–1740, 2004. [PubMed: 15451152]

[5]. Steinhubl SR et al. , "Rationale and design of a home-based trial using wearable sensors to detect asymptomatic atrial fibrillation in a targeted population: The mHealth Screening to Prevent Strokes (mSToPS) trial," Amer. Heart J, vol. 175, pp. 77–85, 2016. [PubMed: 27179726]

[6]. Kim S, Im S, and Park T, "Characterization of quadratic nonlinearity between motion artifact and acceleration data and its application to heartbeat rate estimation," Sensors, vol. 17, no. 8, p. 1872, 2017.

[7]. McConnell MV, Turakhia MP, Harrington RA, King AC, and Ashley EA, "Mobile health advances in physical activity, fitness, and atrial fibrillation: Moving hearts," J. Amer College Cardiol, vol. 71, no. 23, pp. 2691–2701, 2018.

[8]. Goceri E, and Gunay M, "Future healthcare: Will digital data lead to better care?," in Proc. New Trends Issues Adv. Pure Appl. Sci, 2017, pp. 7–11.

[9]. Kamshilin AA et al. , "A new look at the essence of the imaging photoplethysmography," Sci. Rep, vol. 5, pp. 1–9, 2015.

[10]. Sukor JA, Redmond SJ, and Lovell NH, "Signal quality measures for pulse oximetry through waveform morphology analysis," Physiol. Meas, vol. 32, no. 3, pp. 369–384, 2011. [PubMed: 21330696]

[11]. Karlen W, Kobayashi K, Ansermino JM, and Dumont GA, "Photoplethysmogram signal quality estimation using repeated Gaussian filters and cross-correlation," Physiol. Meas, vol. 33, no. 10, pp. 1617–1629, 2012. [PubMed: 22986287]

[12]. Asgari S, Bergsneider M, and Hu X, "A robust approach toward recognizing valid arterial-blood-pressure pulses," IEEE Trans. Inf. Technol. Biomed, vol. 14, no. 1, pp. 166–172, Jan. 2010. [PubMed: 19884099]

[13]. Li Q and Clifford GD, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," Physiol. Meas, vol. 33, no. 9, pp. 1491–1501, 2012. [PubMed: 22902950]

[14]. Orphanidou C, Bonnici T, Charlton P, Clifton D, Vallance D, and Tarassenko L, "Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring," IEEE J. Biomed. Health Inform, vol. 19, no. 3, pp. 832–838, May 2015. [PubMed: 25069129]

[15]. Liu C, Li Q, and Clifford GD, "Evaluation of the accuracy and noise response of an open-source pulse onset detection algorithm on pulsatile waveform databases," in Proc. Comput. Cardiol, 2016, vol. 43, pp. 913–916.

[16]. Elgendi M, "Optimal signal quality index for photoplethysmogram signals," Bioengineering, vol. 3, no. 4, p. 21, 2016.

[17]. Chong JW et al. , "Photoplethysmograph signal reconstruction based on a novel hybrid motion artifact detection–reduction approach. Part I: Motion and noise artifact detection," Ann. Biomed. Eng, vol. 42, no. 11, pp. 2238–2250, 2014. [PubMed: 25092422]

[18]. Zhang Y and Pan J, "Assessment of photoplethysmogram signal quality based on frequency domain and time series parameters," in Proc. 10th Int. Congr. Image Signal Process., BioMed. Eng. Inform., 2017, pp. 1–5.

[19]. Couceiro R, Carvalho P, Paiva RP, Henriques J, and Muehlsteff J, “Detection of motion artifacts in photoplethysmographic signals based on time and period domain analysis,” in Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., 2012, pp. 2603–2606.

[20]. Couceiro R, Carvalho P, Paiva RP, Henriques J, and Muehlsteff J, “Detection of motion artifact patterns in photoplethysmographic signals based on time and period domain analysis,” Physiol. Meas, vol. 35, no. 12, pp. 2369–2388, 2014. [PubMed: 25390186]

[21]. Kamaleswaran R, Mahajan R, and Akbilgic O, “A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using varying length single lead electrocardiogram,” Physiol. Meas, vol. 39, 2018, Art. no. 035006.

[22]. Drew BJ et al. , “Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients,” PLoS One, vol. 9, no. 10, 2014, Art. no. 0110274.

[23]. Tseng Y-H, Ko H-K, Tseng Y-C, Lin Y-H, and Kou YR, “Atrial fibrillation on intensive care unit admission independently increases the risk of weaning failure in nonheart failure mechanically ventilated patients in a medical intensive care unit,” Medicine, vol. 95, no. 20, pp. 1–9, 2016.

[24]. Álvarez D, Hornero R, Marcos JV, and Del Campo F, “Feature selection from nocturnal oximetry using genetic algorithms to assist in obstructive sleep apnoea diagnosis,” Med. Eng. Phys, vol. 34, no. 8, pp. 1049–1057, Oct. 2012. [PubMed: 22154238]

[25]. Nguyen MH and de la Torre F, “Optimal feature selection for support vector machines,” Pattern Recognit., vol. 43, no. 3, pp. 584–591, 2010.

[26]. Gargiulo F, Fratini A, Sansone M, and Sansone C, “Subject identification via ECG fiducial-based systems: Influence of the type of QT interval correction,” Comput. Methods Programs Biomed, vol. 121, no. 3, pp. 127–136, 2015. [PubMed: 26143963]

[27]. Wen T and Zhang Z, “Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic EEG multiclassification,” Medicine, vol. 96, no. 19, pp. 1–17, 2017.

[28]. Pereira T, Paiva JS, Correia C, and Cardoso J, “An automatic method for arterial pulse waveform recognition using KNN and SVM classifiers,” Med. Biol. Eng. Comput, vol. 54, pp. 1049–1059, 2015. [PubMed: 26403299]

[29]. Paiva JS, Cardoso J, and Pereira T, “Supervised learning methods for pathological arterial pulse wave differentiation: A SVM and neural networks approach,” Int. J. Med. Inform, vol. 109, pp. 30–38, 2018. [PubMed: 29195703]

[30]. Snoek J, Larochelle H, and Adams RP, “Practical Bayesian optimization of machine learning algorithms,” Proc. NIPS’12 25th Int. Conf. Neural Inform. Process. Syst., 2016, vol. 2, pp. 2951–2959.

[31]. Duan K, Keerthi S, and Poo A, “Evaluation of simple performance measures for tuning SVM hyper parameters,” Neurocomputing, vol. 51, pp. 41–59, 2001.

[32]. Pereira T, Gadhoumi K, Ma M, Colorado R, Keenan K, and Meisel K, “Robust assessment of photoplethysmogram signal quality in the presence of atrial fibrillation,” in Proc. Comput. Cardiol, 2018.

[33]. Hsu C-W, Chang C-C, and Lin C-J, “A practical guide to support vector classification,” BJU Int, vol. 101, no. 1, pp. 1396–1400, 2008. [PubMed: 18190633]

[34]. Kamaleswaran R, Mahajan R, and Akbilgic O, “A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using varying length single lead electrocardiogram,” Physiol. Meas, vol. 39, 2018, Art. no. 035006.

[35]. Mooney RJ, “Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning,” in Proc. Conf. Empir. Methods iNatural Lang. Process., 1996, pp. 82–91.

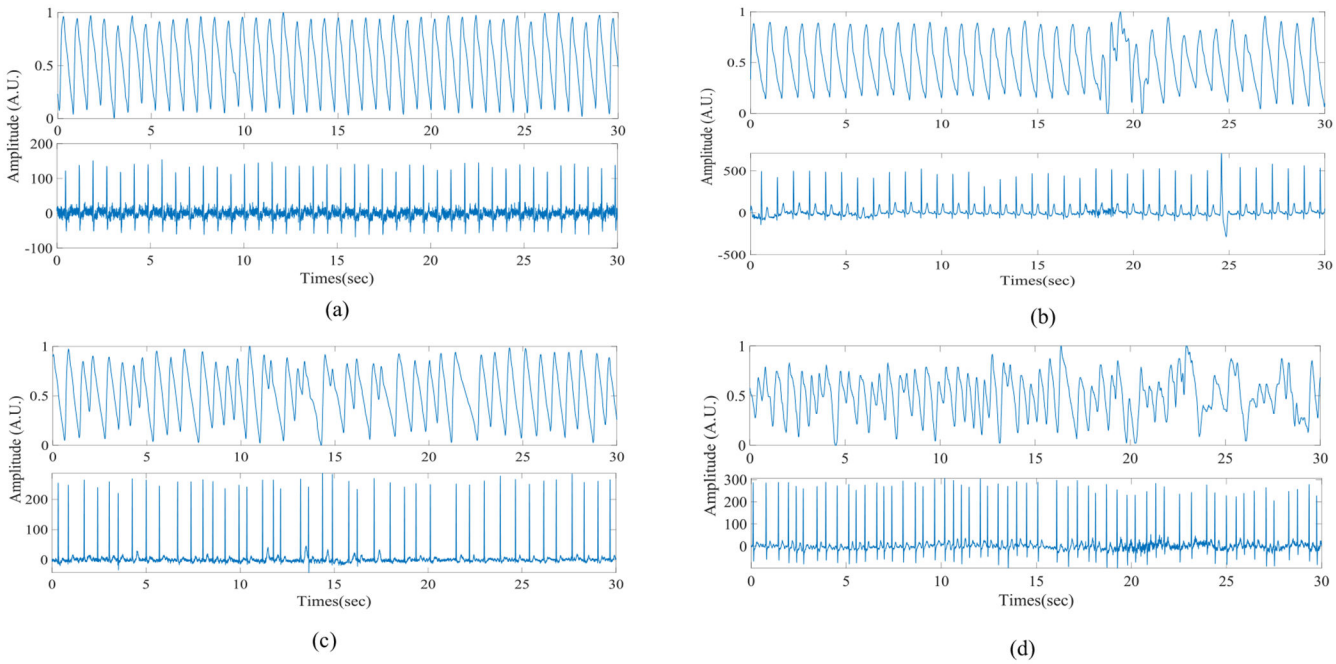

Fig. 1.

Examples of 30-sec segments of PPG co-recorded with ECG showing the cardiac rhythm. (a) Good PPG signal for normal sinus rhythm (NSR); (b) Bad PPG signal for NSR; (c) Good PPG signal for AFib case; (d) Bad PPG signal for AFib.

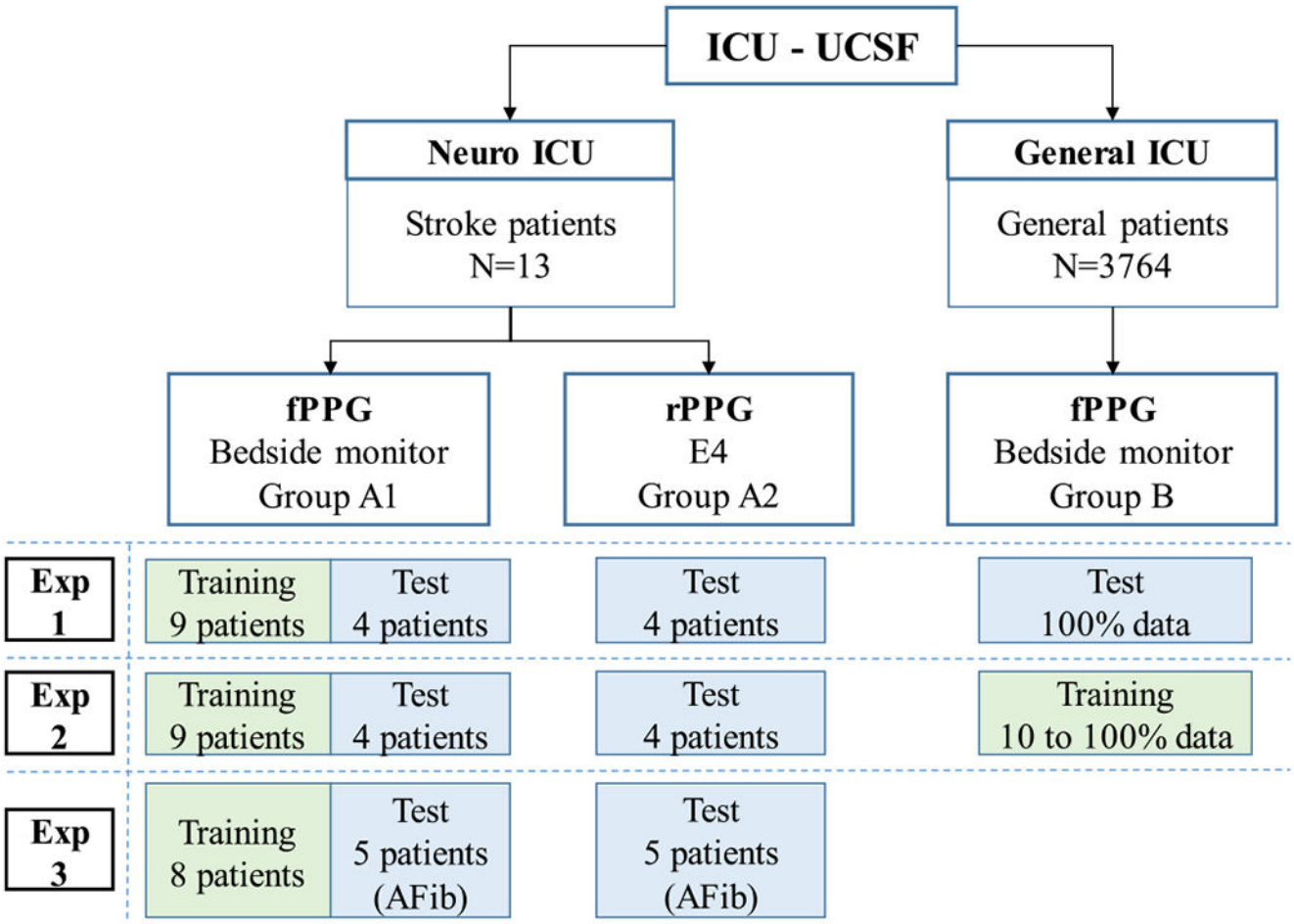

Fig. 2.

Block diagram illustrating the split of data in the different groups. Separation by the clinical features: stroke patients and general group of patients, and by the device used to acquire the PPG signal. Group A1 – data from fPPG of stroke patients; group A2 – data from rPPG of stroke patients; group B – data from fPPG of general patients. The training (green boxes) and test (blue boxes) represent the split data by the three experiments (Exp) described in the next sections.

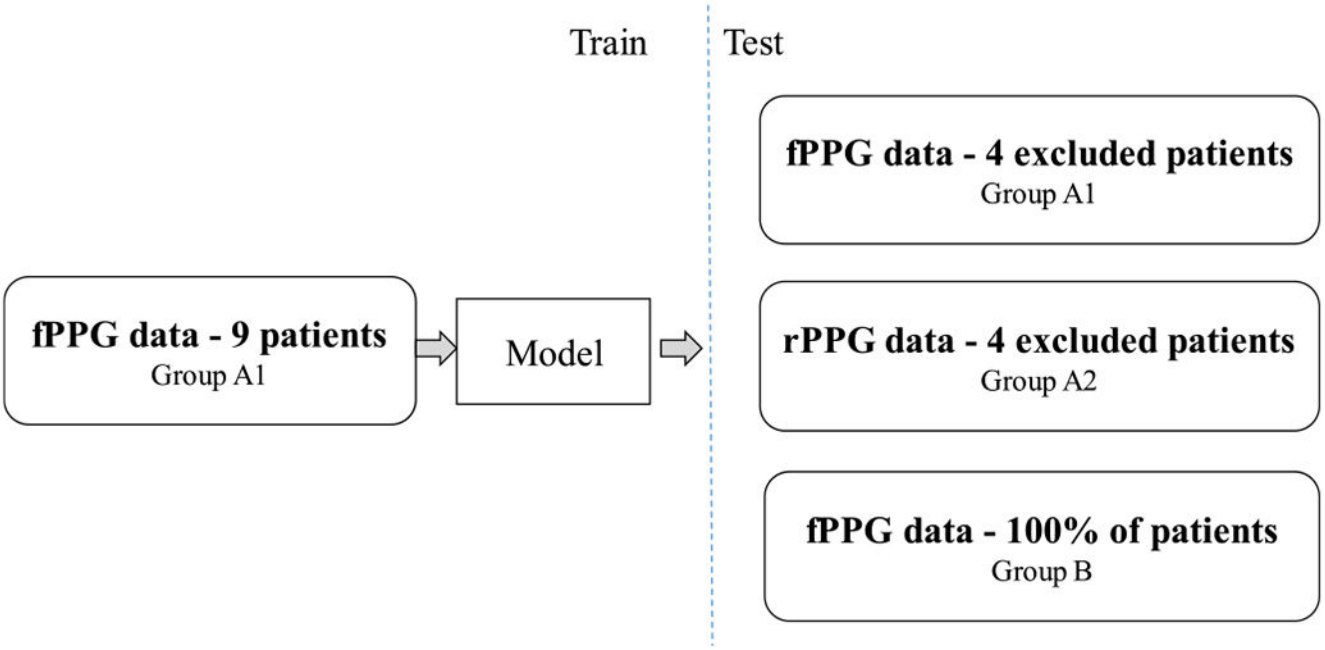

Fig. 3.

Block diagram of the tests of model performance for experiment 1. The classifier was trained with partial fPPG data and tested in the excluded part of fPPG data (4 patients excluded); tested with rPPG data from the same 4 patients excluded of the training data set, and tested with data from general ICU patients.

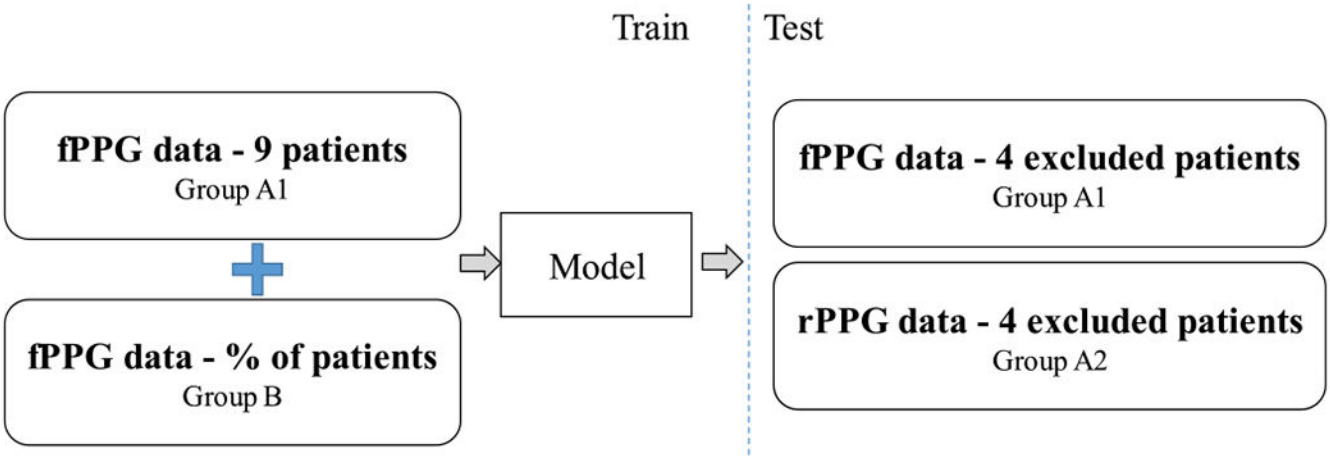

Fig. 4.

Block diagram of tests of model performance for experiment 2. The classifier was trained with partial fPPG data from 9 stroke patients and a part of fPPG from ICU patients: and tested in the fPPG data and rPPG data from the 4 patients excluded from the training data set.

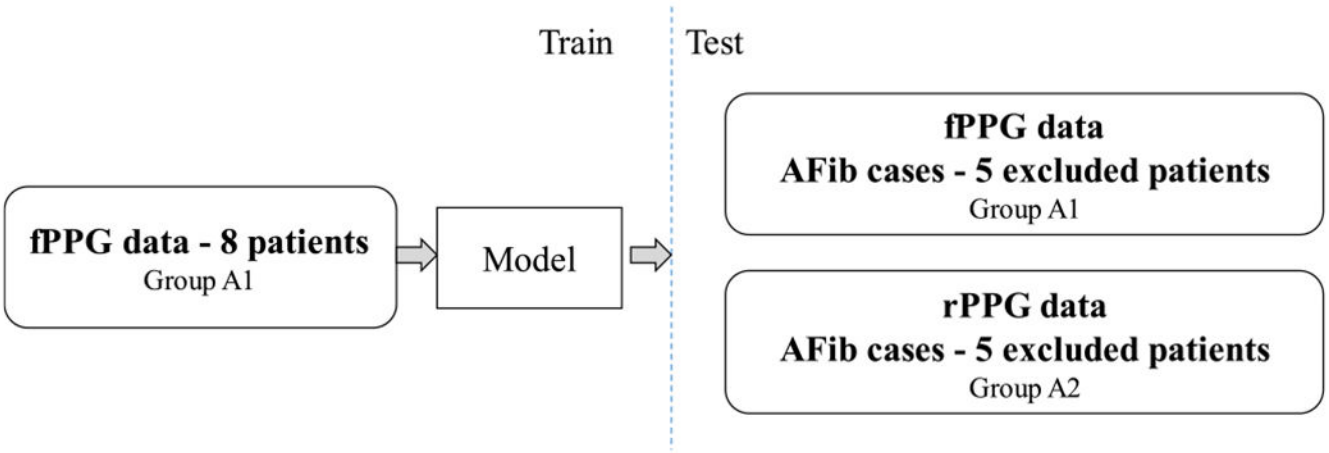

Fig. 5.

Block diagram of performance assessment for an experiment 3. The classifier trained with fPPG data from 8 patients and tested in the excluded part of fPPG and rPPG data from the 5 patients identified as AFib cases.

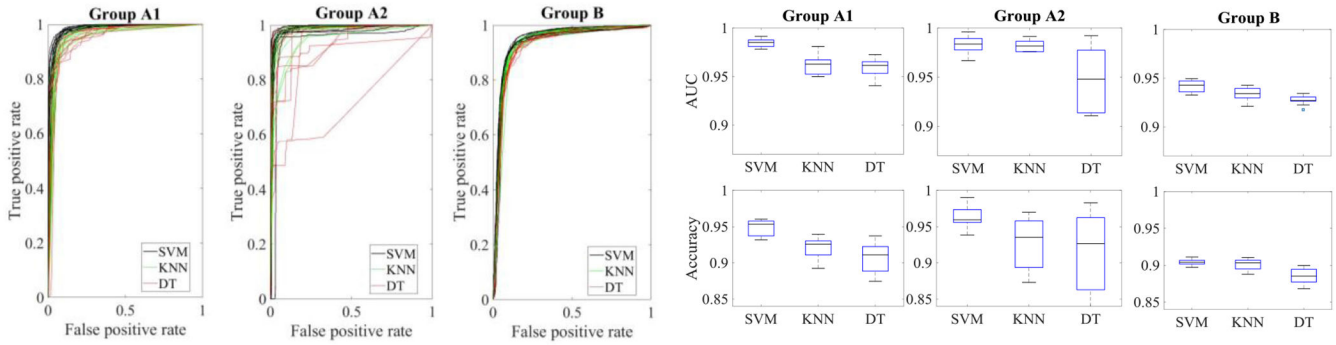

Fig. 6.

ROC curve, AUC and accuracy for different classifiers (SVM, KNN and DT) trained by fPPG from 9 patients from group A1 and tested with the data from the excluded 4 patients for group A1 and A2 and for all patients from group B (experiment 1).

TABLE I

DISTRIBUTION OF fPPG AND rPPG 30-SEC SEGMENTS FROM 13 STROKE PATIENTS, LABELED AS GOOD OR BAD QUALITY

	fPPG segments Group A1			rPPG segments Group A2		
Patient	**Total (n)**	**Good (%)**	**Bad (%)**	**Total (n)**	**Good (%)**	**Bad (%)**
1	365	35.3	64.7	319	0.3	99.7
2	354	97.7	2.3	294	42.5	57.5
3	301	75.1	24.9	235	34.0	66.0
4	286	65.7	34.3	244	-	100.0
5	1172	56.7	43.3	1019	2.6	97.4
6	960	85.4	14.6	634	62.6	37.4
7	2187	58.0	42.0	1648	14.2	85.8
8	399	34.1	65.9	1447	48.8	51.2
9	1845	60.4	39.6	1209	24.4	75.6
10	2310	88.8	11.2	1484	24.9	75.1
11	2298	45.3	54.7	1966	6.8	93.2
12	611	5.7	94.3	468	-	100.0
13	2736	84.8	15.2	1852	62.5	37.5
Total	15824	65.4	34.6	12819	27.5	72.5

TABLE II

PERFORMANCE RESULTS FOR SIX BASELINE METHODS USING DATA FPPG AND RPPG 30-SEC SEGMENTS FROM 13 STROKE PATIENTS

	fPPG data - Group A1				rPPG data - Group A2			
Patient	**Accu**	**F1**	**Sen**	**Sp**	**Accu**	**F1**	**Sen**	**Sp**
Method 1 [10]	0.8475	0.8923	0.9668	0.6225	0.8503	0.7749	0.7900	0.8795
Method 2 [11]	0.9024	0.9250	0.9206	0.8681	0.8808	0.8122	0.7901	0.9246
Method 3 [12]	0.7013	0.8127	0.9916	0.1537	0.6293	0.1845	0.1286	0.8717
Method 4 [13]	0.7959	0.8641	0.9933	0.4235	0.7937	0.7561	0.9803	0.7033
Method 5 [14]	0.8179	0.8466	0.7688	0.9106	0.7240	0.2957	0.1777	0.9885
Method 6 [15]	0.8265	0.8821	0.9938	0.5109	0.8753	0.8147	0.9960	0.8295

TABLE III

PERFORMANCE RESULTS FOR EXPERIMENT 1 USING SVM CLASSIFIER

Training data	**Test fPPG data – Group A1**				**Test rPPG data – Group A2**				**Test fPPG data – Group B**			
Patients for training	**Accu**	**F1**	**Sen**	**Sp**	**Accu**	**F1**	**Sen**	**Sp**	**Accu**	**F1**	**Sen**	**Sp**
1,2,3,6,7,9,11,12,13	0.9513	0.9668	0.9714	0.8972	0.9607	0.9272	0.9519	0.9638	0.9050	0.9287	0.9103	0.8938
3,5,6,7,8,9,10,12,13	0.9373	0.9415	0.9783	0.8937	0.9901	0.9459	0.9423	0.9949	0.9067	0.9304	0.9169	0.8851
1,3,4,5,6,9,10,11,13	0.9578	0.9577	0.9513	0.9643	0.9583	0.9266	0.9540	0.9599	0.9031	0.9271	0.9065	0.8960
1,3,4,6,7,9,11,12,13	0.9561	0.9710	0.9744	0.8998	0.9574	0.9279	0.9487	0.9609	0.8981	0.9230	0.8995	0.8950
1,2,3,5,6,7,9,10,13	0.9324	0.9166	0.9536	0.9189	0.9736	0.9351	0.9345	0.9836	0.9023	0.9266	0.9078	0.8906
1,2,3,5,7,8,9,10,11	0.9558	0.9704	0.9875	0.8690	0.9384	0.9384	0.9646	0.9136	0.9112	0.9342	0.9276	0.8763
3,4,5,6,7,8,11,12,13	0.9604	0.9735	0.9747	0.9181	0.9459	0.8936	0.9507	0.9443	0.9146	0.9367	0.9307	0.8804
2,3,4,5,7,8,9,11,13	0.9597	0.9721	0.9802	0.9083	0.9559	0.9181	0.9336	0.9640	0.9018	0.9262	0.9060	0.8928
1,3,6,7,8,9,10,12,13	0.9319	0.9390	0.9629	0.8947	0.9872	0.9212	0.9196	0.9932	0.9054	0.9293	0.9153	0.8843
2,3,5,6,9,10,11,12,13	0.9472	0.9498	0.9402	0.9551	0.9634	0.9308	0.9575	0.9654	0.8973	0.9222	0.8952	0.9018
Mean	**0.9514**	**0.9587**	**0.9714**	**0.9087**	**0.9600**	**0.9266**	**0.9475**	**0.9606**	**0.9054**	**0.9291**	**0.9132**	**0.8888**
Std	**0.0100**	**0.0188**	**0.0119**	**0.0257**	**0.0149**	**0.0148**	**0.0097**	**0.0229**	**0.0050**	**0.0042**	**0.0103**	**0.0068**

TABLE IV

PERFORMANCE OF BASELINE ALGORITHMS AND SVM CLASSIFIER TESTED WITH THE DATA LABELED AS AFib

	fPPG data - Group A1				rPPG data - Group A2			
Method	**Accu**	**F1**	**Sen**	**sp**	**Accu**	**F1**	**Sen**	**sp**
Method 1 [10]	0.5778	0.7150	0.7698	0.1540	0.9036	0.5745	0.6279	0.9355
Method 2 [11]	0.4781	0.5965	0.5604	0.2966	0.9313	0.6545	0.6279	0.9664
Method 3 [12]	0.6674	0.7965	0.9458	0.0529	0.8651	0.1250	0.0930	0.9543
Method 4 [13]	0.6057	0.7439	0.8323	0.1057	0.7349	0.4330	0.9767	0.7070
Method 5 [14]	0.3384	0.3595	0.2698	0.4897	0.9048	0.2178	0.1279	0.9946
Method 6 [15]	0.6194	0.7545	0.8500	0.1103	0.8940	0.6589	0.9884	0.8831
Our Model	**0.9312**	**0.9468**	**0.9368**	**0.9205**	**0.9075**	**0.7557**	**0.7514**	**0.9442**