Optimized Signal Quality Assessment for Photoplethysmogram Signals Using Feature Selection

Fahimeh Mohagheghian, Dong Han, *Student Member, IEEE*,
Andrew Peitzsch, *Student Member, IEEE*, Nishat Nishita, Eric Ding, Emily L. Dickson, Danielle DiMezza,
Edith M. Otabil, Kamran Noorishirazi, Jessica Scott, Darleen Lessard, Ziyue Wang, Cody Whitcomb,
Khanh-Van Tran, Timothy P. Fitzgibbons, David D. McManus, and Ki. H. Chon, *Fellow, IEEE*

Abstract*—*Objective: **With the increasing use of wearable healthcare devices for remote patient monitoring, reliable signal quality assessment (SQA) is required to ensure the high accuracy of interpretation and diagnosis on the recorded data from patients. Photoplethysmographic (PPG) signals non-invasively measured by wearable devices are extensively used to provide information about the cardiovascular system and its associated diseases. In this study, we propose an approach to optimize the quality assessment of the PPG signals. *Methods:* We used an ensemble-based feature selection scheme to enhance the prediction performance of the classification model to assess the quality of the PPG signals. Our approach for feature and subset size selection yielded the best-suited feature subset, which was optimized to differentiate between the clean and artifact corrupted PPG segments. *Conclusion:* A high discriminatory power was achieved between two classes on the test data by the proposed feature selection approach, which led to strong performance on all dependent and independent test datasets. We achieved accuracy, sensitivity, and specificity rates of higher than 0.93, 0.89, and 0.97, respectively, for dependent test datasets, independent of heartbeat type, i.e., atrial fibrillation (AF) or non-AF data including normal sinus rhythm (NSR), premature atrial contraction (PAC), and premature ventricular contraction (PVC). For independent test datasets, accuracy, sensitivity, and specificity rates were greater than 0.93, 0.89, and 0.97, respectively, on PPG data recorded from AF and non-AF subjects. These results were found to be more accurate than those of all of the contemporary methods cited in this work. *Significance:* As the results illustrate, the advantage of our proposed scheme is its robustness against dynamic variations in the PPG signal during long-term 14-day recordings accompanied with different types of physical activities and a diverse range of fluctuations and waveforms caused by different individual hemodynamic characteristics, and various types of recording devices. This robustness instills confidence in the application of the algorithm to various kinds of wearable devices as a reliable PPG signal quality assessment approach.**

***Index Terms*—Biomedical signal processing, feature extraction, machine learning, photoplethysmography.**

I. Introduction

In recent years, the use of modern wearable devices such as smartwatches, fitness and health trackers/bands, and health patches has been growing for monitoring of human vital signs. PPG sensors are common in wrist-worn devices and are often accompanied by accelerometers to measure body movement. PPG is a non-invasive sensing technique to record tissue and blood volume alterations through optical absorption and scattering that enable monitoring of heart rates (HR), heart rhythms, and hemoglobin oxygen saturation (SpO2).

The reliability of the estimated HR is highly correlated to the quality of the underlying recorded PPG signals, which are susceptible to different types of noise and artifact, particularly motion artifacts (MAs). These artifacts can be in the same frequency range as the HR signal and thus, motion artifact reduction for these devices is challenging. In many applications for PPG signals, quality assessment algorithms are used to recognize and reject the noisy PPG segments. PPG quality assessment becomes more challenging when ectopic heartbeats, e.g., PAC, PVC, and AF are present. The waveform characteristics of the PPG signals during these ectopic rhythms can resemble the artifact-contaminated PPG segments.

Increased frequency of PACs increases the risk of mortality attributable to myocardial infarction, heart failure, and sudden

Manuscript received January 21, 2022; revised February 28, 2022; accepted March 8, 2022. This work was supported in part by NIH under Grant 1R01 HL137734 and in part by NSF under Grant 1522087. *(Corresponding author: Ki. H. Chon.)*

Ki. H. Chon is with the Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06084 USA (e-mail: ki.chon@uconn.edu).

Fahimeh Mohagheghian, Dong Han, and Andrew Peitzsch are with the Department of Biomedical Engineering, University of Connecticut, USA.

Nishat Nishita is with the Department of Public Health Sciences, University of Connecticut Health, USA.

Eric Ding, Danielle DiMezza, Edith M. Otabil, Kamran Noorishirazi, Jessica Scott, Darleen Lessard, Ziyue Wang, Khanh-Van Tran, Timothy P. Fitzgibbons, and David D. McManus are with the Division of Cardiology, University of Massachusetts Medical School, USA.

Emily L. Dickson is with the College of Osteopathic Medicine, Des Moines University, USA.

Cody Whitcomb is with the School of Medicine, Tufts University, USA.

This article has supplementary downloadable material available at https://doi.org/10.1109/TBME.2022.3158582, provided by the authors.

Digital Object Identifier 10.1109/TBME.2022.3158582

cardiac death [1]. Further, frequent PVC is associated with heart failure as well as serious heart arrhythmias such as ventricular fibrillation (VF) and AF [2]. Thus, it is crucial to differentiate between clean and corrupted PPG signals in the presence of these ectopic heartbeats.

Several computational approaches such as machine leaning (ML), deep neural network (DNN), and heuristic rules-based frameworks have been proposed to detect the artifact parts of pulsatile physiological signals. Sukor *et al.* [3] employed a simple decision-tree classifier using waveform morphological features of the PPG signals. The performance of their algorithm was validated on 104 signals included 7669 beats. They achieved a mean Cohen's kappa coefficient (κ) of 0.64 and the mean sensitivity, specificity, and accuracy were 89%, 77%, and 83%, respectively, based on the definition of a positive being a clean pulse. In [4], dynamic time warping was applied to nonlinearly stretch each beat to fit a dynamic beat template and combine it with other related features. Then, a multi-layer perceptron neural network was used to determine the signal quality index (SQI) using an expert-labeled database of 1,055 6-sec PPG segments, during both normal and arrhythmic events. The authors in [5] proposed a combination of morphological characteristics and temporal variability information in the PPG signals to yield an adaptive SQA approach.

In [6], the authors developed an algorithm to segment pulse oximetry signals into pulses and estimate the signal quality in real time. Cross-correlation of consecutive pulse segments was used to estimate an SQI, which was significantly lower in the presence of artifacts compared to SQI values of clean signals in the test dataset. The authors in [7] proposed an SQI based on adaptive template matching between the average PPG-pulse waves and each individual PPG-pulse to assess PPG signal quality for reliable heart rate detection using wearable sensors. The authors in [8] developed and tested eight SQIs based on eight features for 106 annotated 60-sec recordings of PPG data. To identify the best feature, all indices were evaluated using four classifiers. The author showed that skewness outperformed the other features with overall F1 scores of 86.0%, 87.2%, and 79.1%, to discriminate between excellent PPG and acceptable, acceptable combined with noisy, and noisy recordings, respectively, when clean PPGs are positives. Dao *et al.* [9] proposed an approach called TifMA, using the signal time–frequency (TF) spectrum developed based on a TF technique named variable frequency complex demodulation (VFCDM) to detect the motion artifact-corrupted PPGs. In [10] a real-time automatic SQA algorithm for PPG was suggested based on the hierarchical decision rules in combination with simple features. The algorithm achieved an average of 99.29%, 95.31%, and 97.76% for sensitivity, specificity, and accuracy, respectively, when positives are acceptable PPG segments. In [11], six morphological features were proposed using beat-scale SQA for PPGs using machine learning approach. Forty-six 30-min annotated PPG segments from patients with atrial fibrillation, hypoxia, acute heart failure, pneumonia, acute respiratory distress syndrome (ARDS), and pulmonary embolism were tested. The authors showed the high performance of their constructed support vector machine (SVM) model in terms of sensitivity and positive predictive value (PPV) on their test data. In [12], temporal and spectral features were extracted from each PPG segment recorded from patients with atrial fibrillation. The authors achieved accuracy of 0.9477 and 0.9589 from fingertip PPG and radial PPG, respectively, using an SVM classifier.

In this study, our main objective was to identify the best feature subset to ensure accurate noise detection and quality assessments for PPG signals with a diverse range of morphologies, for both non-AF and AF data, as the latter can be mis-detected as noisy non-AF PPG signals.

II. Method

A. Dataset Description

This section consists of descriptions of the datasets used in this study, including the data collected in our current study (Pulsewatch) and our previous study (UMMC Simband), and publicly available datasets (Stanford University's PPG dataset and MIMIC III).

1) Data Collection—Pulsewatch Dataset: The PPG data were collected in a multi-phase study called Pulsewatch. Details of the study phases can be found in [13]. The study consisted of two parts. Design and development of the Pulsewatch system (app and watch algorithms) were completed in Part I. Part II included data collection in clinical and AF trials For the clinical trial, participants with a prior history of stroke/transient ischemia (TI) (n=90) were asked to use the gold-standard Cardiac Insight cardiac patch monitor device, smartwatch, and a Samsung smartphone that had the Pulsewatch study apps downloaded on it. For the AF trial, the patients with confirmed persistent AF were recruited for a short duration experiment (about 20 min) (UIDs #301-329, see Appendix I) or 7 days data collection (UID #400, see Appendix I). Formal ethical approval for this study has been obtained from the University of Massachusetts Medical School Institutional Review Board (approval number H00016067). Written informed consent was collected from all patient participants. The reference ECG and smartwatch data were simultaneously measured from the chest and wrist using a 2-lead rhythm patch device (Cardea SOLO, Cardiac Insight Inc., Bellevue, WA, USA) and, Samsung Gear S3, or Galaxy Watch 3 (Samsung, San Jose, CA, USA), respectively. The patch ECG data, which were used as the reference, consisted of one-channel signals sampled at 250 Hz. The smartwatch data consisted of a one-channel PPG signal and a one-channel magnitude of the accelerometer (ACC) signal. Smartwatch signals were all sampled at 50 Hz and were automatically segmented into 30-sec lengths. The enrolled patients wore the smartwatch and ECG patch 24 hours a day with no restriction on their regular daily activities, for 14 consecutive days. Due to the 7-day battery limitation, patients switched to a second new ECG patch on the 7th day of the trial. Smartwatches were charged daily for 1 h.

2) UMMC Simband Dataset: 37 patients (28 male and 9 female), aged 50-91 years old participated in the smartwatch study at the ambulatory cardiovascular clinic at University of Massachusetts Medical Center (UMMC). Their recorded signals contain AF and non-AF data including cardiac arrhythmias, such as PAC and PVC. Details of subject characteristics, monitoring duration, and arrhythmia burden are provided in [14]. Reference ECG and smartwatch data were simultaneously measured from the chest and wrist using a 7-lead Holter monitor (Rozinn RZ153+ Series, Rozinn Electronics Inc., Glendale, NY, USA) and Simband 2 (Samsung Digital Health, San Jose, CA, USA), respectively. ECG data were composed of 3-channel signals, each sampled at 180 Hz. Simband data were comprised of 8-channel PPG signals (sampled at 128 Hz), three-axis accelerometers and a one-lead ECG. Only the 5th PPG channel (green LED color, wavelength 520–535 nm) was used for data analysis since it consistently provided the best signal quality. The alignment of the Simband and Holter ECG signals was performed by estimating the cross-correlation between them. In this study, PPG and ACC data were segmented into 30-sec length segments with no overlap and down-sampled to 50 Hz and 20 Hz, respectively. This dataset was created as part of a preliminary study conducted previously at University of Connecticut (UConn). The dataset is available for download on our lab's website listed in the Supplementary Materials section.

3) Stanford University's PPG Dataset: An open access database has been provided by Stanford University, which collected the data from participants undergoing elective cardioversions (CV) or elective stress tests to develop a convolution neural network (CNN)-based AF event detection model called DeepBeat [15]. Data were extracted from a wrist-based PPG wearable device (Simband), sampled at 128 Hz, and partitioned into 25-sec segments. Average monitoring time was about 20 min post and 20 min prior to the CV procedure for 132 participants with confirmed AF diagnosis undergoing direct current cardioversion for the treatment of AF. Average monitoring time was about 45 min for the 42 participants in the elective exercise stress test.

4) MIMIC III Dataset: The publicly available Medical Information Mart for Intensive Care (MIMIC III) database provides continuous ECG and pulse oximetry waveforms (PLETH) from patients in critical care at a large tertiary care hospital [16]. All signals were originally sampled at 125 Hz.

In this study, we used data that had been prepared for a previous AF study [17], in which four batches of 50 ECG recordings from patients hospitalized with sepsis were randomly selected. The ECG signals were annotated by board-certified physicians specializing in AF management. Then, one batch was used for training, which contained 25 AF subjects. Since each subject's recording contained hundreds of hours of data, a subset of 5 AF subjects' corresponding ECG and PPG data were randomly selected and annotated. In this study, we used the PPG data from those 5 AF subjects to have a comparable number of AF and non-AF segments for testing. The data were down-sampled to 50 Hz and partitioned into 30-sec lengths with no overlap. The MIMIC III data used in this study is available for download on our lab's website https://biosignal.uconn.edu/resources/, listed in the Supplementary Materials.

B. Signal Annotation

PPG signal annotation is known to be a complicated and subjective procedure. Hence, to have consistent annotation, the heart rate extracted from a PPG signal was compared to the aligned ECG HR as the reference for clinical and AF trials, Simband, and MIMIC III datasets. Two people with significant experience with PPG signals reviewed all segments manually and performed the annotations. The final annotation was based on the consensus of the two experts' adjudications. When a segment adjudication by the two experts was in disagreement, a third expert reviewer's opinion was sought and the final decision was based on the view of the majority.

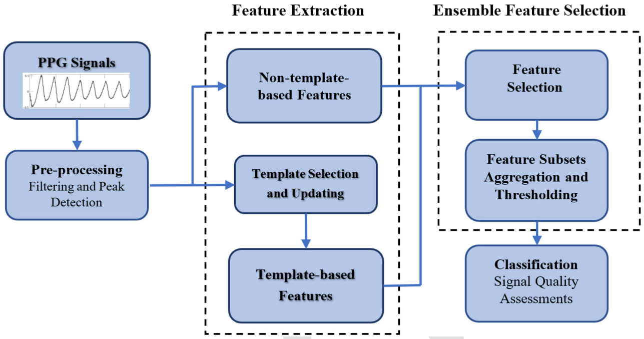
Feature Extraction
Ensemble Feature Selection
PPG Signals
Pre-processing
Filtering and Peak Detection
Non-template-based Features
Template Selection and Updating
Template-based Features
Feature Selection
Feature Subsets Aggregation and Thresholding
Classification
Signal Quality Assessments

Fig. 1. The block diagram of the proposed approach for PPG signal quality assessment.

The experts performed adjudication by observing the PPG pulse waveform and comparing the heart rates extracted from PPG and the corresponding ECG. A PPG segment was annotated as being noisy if the HRs calculated from the PPG segment deviated more than 5 seconds from clean ECG heart rates or the PPG waveforms were corrupted for more than 5 seconds, otherwise, it was annotated as clean. We chose the 5 s limit on the corrupted signal as our previously AF detection study has shown that in a 30-sec data segment the AF detection algorithm's accuracy is not affected with less than 5 seconds of noisy data [18].

In addition, annotation of the PPG signal rhythm was performed by three experts. Each PPG segment was annotated using two labels: AF or non-AF (including NSR, PAC, and PVC). The aligned ECG signal was used as the reference for rhythm annotation of the PPG.

C. Preprocessing

Fig. 1 illustrates the overall block diagram of our proposed approach for SQA of the PPG segments. In the preprocessing, all PPG segments were first filtered by Butterworth high-pass and low-pass filters with cut off frequencies of 0.5 and 20 Hz, respectively. The filters removed baseline wander and other types of noise such as ambient light noise. Subsequently, PPG signal peak detection was performed using the waveform envelope peak detection (WEPD) algorithm [19]. In the WEPD algorithm, a waveform envelope is used to remove excessive beats caused by the dicrotic notch in the normal sinus rhythm (NSR) data, while still retaining sensitivity to irregular heartbeats in AF data.

TABLE I
FEATURES EXTRACTED FROM PPG SEGMENTS

Feature type	Domain	Feature index	Description	Analysis scale
Temporal	Time Domain	RMSSD	Root mean square of successive differences [12]	HR
		$\overline{\Delta p}$	Normalized pulse duration [11]	Beat
Statistical	Time Domain	Skew	Skewness [12]	Segment
		Kurt	Kurtosis [12], [20]	Segment
		pNN40	Percentage of successive beat intervals that differ by more than 40 msec	IBI
		pNN70	Percentage of successive beat intervals that differ by more than 70 msec	IBI
		SampEn	Sample entropy [12], [20]	Segment
		$D(T^-\|\|B^-)$	Dissimilarity measure of negative-peaked beats [11]	Beat
	Time-Freq. Domain Wavelet coefficients (cA4, cD2, cD3, cD4)*	W-STD	Standard deviation of wavelet transform	Beat
		W-Skew	Skewness of wavelet transform	Beat
		W-Kurt	Kurtosis of wavelet transform	Beat
		W-MSubEn	Sub-band energy of wavelet transform [8]	Beat
Morphological	Time Domain	$\overline{\Delta P^-}$	Normalized negative-to-negative peak jump [11]	Beat
		$\overline{\Delta P}$	Normalized beat amplitude jump [11]	Beat

*cA4, cD2, cD3, and cD4 represent the approximation coefficients scale 4, detail coefficients scale 2, detail coefficients scale 3, and detail coefficients scale 4, respectively.

D. Features

A number of features for classifying PPG signals as either clean or corrupted have been introduced in previous studies. We categorized the features extracted in this study into temporal, morphological, and statistical features in time and time-frequency domains (see Table I).

1) Non-Template-Based Features:

- RMSSD: The root mean square of successive differences (RMSSD) was extracted from the intervals between consecutive heartbeats, also known as interbeat intervals (IBI).
- Skew: As a measure of the probability distribution symmetry, skewness was calculated for each PPG segment.
- Kurt: Kurtosis, which is a statistical measure to describe the distribution of observed data around the mean, was calculated for each PPG segment.
- pNN40/pNN70: Percentage of successive IBI that differ by more than 40/70 msec was calculated.
- SampEn: Sample entropy of each PPG segment was calculated. Entropy quantifies how much the probability density function (PDF) of the signal differs from a uniform distribution and thus provides a quantitative measure of the uncertainty present in the signal.
- W-STD/W-Skew/W-Kurt/W-MSubEn: By transforming the signals from the original time domain to the time-frequency domain, it is possible to observe the variability in the spectral power of the different frequencies over time. We applied a wavelet transform for each PPG beat and then, extracted the following measures from approximation and detail coefficients: standard deviation, skewness, kurtosis, and average of sub-band energy.

2) Template-Based Features: Six features suggested in [11] were adopted in this study, however, a different strategy was used to select and update the template segment. Time intervals and morphological features were extracted from each PPG beat based on a distance from the baseline values obtained from the template segment (see next section).

E. Feature Extraction

To extract the template-based features from each PPG segment, we selected a clean congruous PPG segment as the template segment, which was used to extract the template beat and baseline values. We developed an adaptive framework to update the template segment in order to extract the characteristics of varied waveforms that arose from various arrhythmias or individuals' activities. To this aim, the first recognized clean PPG segment from each subject's data was considered as the template segment. According to the non-stationary characteristics of the PPG signals, the template segment was updated for PPG segments, which showed different dynamic characteristics over time. Thus, we updated the template segment for a predefined number of segments (which was 10 in this study).

1) Selecting and Updating the Template Segment: To select an initial template segment, a clean PPG segment was recognized based on the specific criteria below:

1) Number of detected peaks, which estimates the number of pulses, should be more than 80% of length of the PPG segment. Therefore, the number of peaks should be more than 24 in a 30-sec PPG segment (25 pulses or more in 30 seconds).
2) It is known that amplitudes of peak points are almost constant in a clean PPG segment [3], [6]. Hence, peak amplitude dispersion is a quantitative indicator for morphological variability of the waveform. We used normalized peak amplitude dispersion as an index to identify the clean signals:

$$D = \frac{s}{\mu} \tag{1}$$

where s and μ are standard deviation and mean value of the peak amplitudes within the PPG segment, respectively, and D is the Coefficient of Variation (CV) [21]. D, also known as the relative standard deviation (RSD), shows the extent of variability in relation to the mean of the peak amplitudes. To combine CV calculated for both positive and negative peaks, the measure D_{comb} is computed as:

$$D_{comb} = e^{-(|D_1|+|D_2|)} \tag{2}$$

where D_1 and D_2 are CV values for positive and negative peaks. By applying a threshold (Thr) on D_{comb}, the PPG segment is clean when: $D_{comb} > Thr$.

As the first PPG segment in the data set might not satisfy the clean segment criteria to be the template segment, we searched all PPG segments to find the first segment that met the criteria. Having the initial clean segment as the template, the features were extracted from the previous and subsequent segments. The template was updated to extract the features from subsequent segments, when a PPG segment was recognized to satisfy the criteria.

To extract the features, first, the baseline values were generated from a template segment. Then, the features were extracted from each PPG beat based on the (dis)similarity between each PPG beat and baseline values or template beat as proposed in [11].

To identify the best feature subset, a comprehensive feature set is desired. The likelihood of selecting the optimum feature subset is higher when there is a large number of features in the feature set. Thus, we extracted different types of features to constitute the initial feature set (134 features).

As the features $\overline{\Delta p}$, $\overline{\Delta P^{-}}$, ΔP, W-STD, W-Skew, W-Kurt, and W-MSubEn were extracted for each PPG beat, the average, standard deviation, skewness, and kurtosis were calculated for each PPG segment (see Table I).

2) Feature Selection Procedure: Once all the initial features were extracted from the PPG signals, feature selection was performed to specify which features are important for PPG noise detection. Among a number of approaches, a filter-wrapper feature selection method based on the IWSSr algorithm [22] was used in this study.

3) Improved IWSSr Algorithm: IWSSr uses symmetrical uncertainty (SU) to rank the features based on their relevance to the class labels [23]. Then, the optimal subset of features is selected using an incremental procedure, in which one feature at a time from unexplored features is added to the selected subset based on the performance of the selected subset on a minimum number of folds and average of the performance over all folds as the significance testing. Adding the features to the subset was accomplished repeatedly until no improvement on the subset performance occurred.

In this study, we improved the IWSSr algorithm by applying the backward search strategy to the feature subset as a revising step to the classic IWSSr algorithm (Algorithm 1). Using backward search strategy, higher computational efficiency was achieved during training the model, and model generalization error and feature redundancy were reduced by eliminating the irrelevant features. In our approach, the Minimum Redundancy Maximum Relevance (MRMR) method [24] was used to rank the features and then, a selected subset of features was created as in the IWSSr algorithm. In the second phase, the wrapper-based backward search was executed on the selected subset to remove redundant features by evaluating the obtained subset. Backward steps were accomplished as long as the evaluated performance improved, reducing the size of the subset by one feature.

Algorithm 1: Pseudo-Code for Improved IWSSr Algorithm.

```
Input: Data $D$, feature set $F$, class label $C$, and
  minimum number of folds with specific accuracy $nf$
Output: Selected feature subset $S$
Initialization: Rank the features using a filter method //
  We used MRMR method;
$S = R_1$ //The first feature is selected
accuracy=evaluate($DC, D^{\downarrow S\cup\{C\}}$); // DC: Discriminant
  classifier
for $i = 2$ to $n$ do
  bestOp=null;
  // Replacement
  for $j = 1$ to $length(S)$ do
    $S_{new}$=update(copy($S$),swap($S_j, R_i$))
    [$accuracy_{new}, num$]=evaluate(DC,$D^{\downarrow S_{new}\cup\{C\}}$);
    if $accuracy_{new} > accuracy$ && $num \geq nf$
    then
      bestOp=swap($S_j, R_i$);
      $accuracy = accuracy_{new}$;
    end if
  end for
  // Addition
  $S_{new}$=update(copy($S$),add($R_i$));
  [$accuracy_{new}, num$]=evaluate(DC,$D^{\downarrow S_{new}\cup\{C\}}$);
  if $accuracy_{new} > accuracy$ && $num \geq nf$ then
    bestOp=add($R_i$);
    $accuracy = accuracy_{new}$;
  end if
  //Replacement or addition
  if bestOp!=null then
    update($S$,bestOp);
  end if
end for
//Removing (backward search)
$accuracy = 0$;
while $accuracy_{new} > accuracy$ && $num \geq nf$ do
  for $j = 1$ to length($S$) do
    $S_{new}$=update(copy($S$),remove($R_j$));
    [$accuracy_{new}, num$]=evaluate(DC,$D^{\downarrow S_{new}\cup\{C\}}$);
    if $accuracy_{new} > accuracy$ && $num \geq nf$
    then
      bestOp=remove($R_j$);
      $accuracy = accuracy_{new}$;
    end if
    if $bestOp! = S$ then
      update($S$,bestOp);
    end if
  end for
end while
return $S$
```

4) Ensemble Feature Selection: The aim of the ensemble feature selection is to generate an ensemble of feature subsets and then aggregate them into a single feature subset under the assumption that the aggregated feature subset is more stable than each of the single results; by combining multiple feature subsets we reduce the probability of choosing an unstable subset [25].

Ensemble feature selection approaches have shown superior potential to remove less important features. This improves the

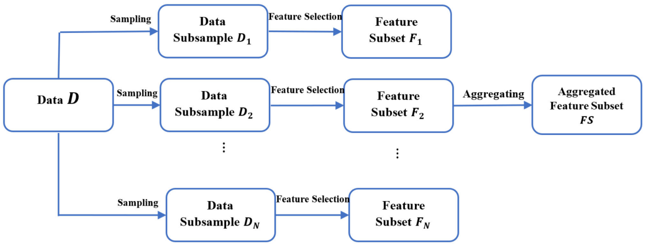

Fig. 2. Diagram of the ensemble feature selection approach.

robustness and yields more efficient results compared to the standard feature selection algorithms. Feature subsets selected by standard feature selection techniques are more likely locally optimal, while the ensemble feature selection approaches show more capability to achieve a better approximation to the optimum feature subset by averaging different hypotheses [26]. The strategy that we used in this study is termed functionally homogeneous ensemble selection, in which the data are partitioned by samples and a single feature selection method is applied to all partitions of the original data (analogous to data perturbation in the field of ensemble learning). Our employed scheme for feature selection can be summarized in three steps (see Fig. 2):

1) Divide the training dataset D into several partitions or subsets by randomly drawing observations containing 80% of D,
2) Apply a single feature selection algorithm to the subsets,
3) Combine the selected feature sets into a single feature set.

5) Aggregation of Feature Subsets: Different ranked feature lists extracted via the ensemble selection strategy, should be combined into a single list, which is the final ranked feature subset. Thus, an appropriate aggregation function (also, called combination function) is required to assign a score to each feature as the feature's score across all feature products. As one of the most commonly used approaches in classification, majority voting, which is based on the most-agreed upon class label, has been adopted for ensemble feature selection [27]. In this approach, which was also used in this study, the decision for each component i of the ensemble can be shown in a Boolean vector DM_i with the size of M, where M is the total number of features. Then, the decision for the ensemble is represented by an $N \times M$ matrix DM, where N is the number of ensemble components. In this representation, the binary cell value DM_{ij} indicates whether $f_j \in F_i$, where f_j is the jth feature among total features and F_i is the feature subset resulting from data partition D_i. Then, the ensemble vote (agreement) v_j is calculated for each feature f_j based on the ensemble decision matrix DM by: $v_j = \frac{\sum_i DM_{ij}}{N}$. The threshold Th $(0 < Th \leq 1)$ for ensemble votes can be applied to control the number of features being included in the final feature subset FS comprised of features with $v_j > Th$.

6) Ensemble Vote Threshold: In order to determine the optimal threshold of votes (v), [28] proposed to find the value, which minimizes the fitness criterion $f(v)$ based on the training classification error (E) and percentage of retained features (P).

$$f(v) = \alpha E(v) + (1 - \alpha) P(v) \tag{3}$$

where α is a parameter with a value in the interval $[0, 1]$ that measures the relative relevance of both values. The main disadvantage of this approach is that by involving a classifier to calculate the training classification error, the obtained threshold is dependent on the selected classification method. Another approach as proposed in [29], is to use problem complexity (or difficulty) measures to involve the features which reduce the complexity of the data. Complexity measures can examine the capability of a single feature to discriminate between classes. A well-known measure called Fisher's discriminant ratio (FR) calculates how separated two classes are according to a specific feature [30]. The generalized Fisher's ratio for a binary or multiclass problem is defined as:

$$FR = \frac{\sum_{k=1}^{C} n_k \cdot \delta(\mu, \mu_k)}{\sum_{k=1}^{C} \sum_{j=1}^{n_k} \delta(x_j^k, \mu_k)} \tag{4}$$

where n_k denotes the number of samples in class k, δ is a metric, μ is the overall mean, μ_k is the mean of class k, and x_j^k represents the sample j belonging to class k.

As the problem difficulty is inversely proportional to the Fisher's discriminant ratio, we proposed an efficient criterion based on the cumulative relevance of FR values as the complexity measure used in the fitness criterion below:

$$E(v) = \alpha CM(v) + (1 - \alpha) R(v) \tag{5}$$

where R is the retained feature ratio and CM is the complexity measure calculated by: $\frac{1}{\sum_j FR_j}$ for the features with $v_j > Th$.

As a high FR value represents high discriminability of the input feature, a low CM value is desirable. By increasing the number of features the CM value reduces, however, the R value increases. Thus, there is a trade-off to reducing CM and R values. By minimizing the fitness function $E(v)$, the optimum threshold of votes and feature number are achieved. The pseudocode for the proposed algorithm is presented in Algorithm 2.

F. Experimental Design

1) Training Datasets: In this study, 3432 PPG segments were selected as the training dataset from 53 subjects from three datasets: the clinical trial, the AF trial, and Stanford University's database. Table II shows the number of selected segments from the three datasets for training the classification model. Training data from the clinical trial were comprised of AF and non-AF PPG segments. To select the non-AF training data from the clinical trial dataset, we divided the data into hourly blocks. One hundred 30-sec segments were extracted from the hourly blocks of the two first 24-hour periods of data recording. The blocks were randomly selected for each subject. AF training data segments from the clinical trial included 806 30-sec segments extracted from randomly selected blocks from one subject, who demonstrated AF during recording. However, a few segments were excluded from the clinical trial training data due to data recording issues such as recording when the watch was not being worn by the subject. In total, 2793 30-sec PPG segments were selected from the clinical trial. Fig. 3 shows the distribution of the non-AF training data during the two first 24-hour periods

Algorithm 2: Pseudo-Code for the Proposed Ensemble Feature Selection Scheme.

```
Input: Data: $D_{(N\times M)}$= training dataset with $N$ samples
  and $M$ features
$X \leftarrow$ Set of features, $X = \{f_1, \ldots, f_M\}$
$s \leftarrow$ Number of subsamples of $D$
$DM_i \leftarrow$ Decision matrix for each subsample $i$ of $D$,
  $|DM_i| = M$
$\alpha \leftarrow$ Relative relevance of complexity measure and
  selected feature ratio
Output: Final feature subset $FS$ ($FS \subset X$)
//Obtaining a decision matrix to show selected features in
  each data subsample
//Initialize $DM_i$
for $i = 1$ to $s$ do
    $D_i \leftarrow$ subsample of $D$, maintaining the class
      distribution
    Apply the feature selection algorithm on $D_i$
    $F_i \leftarrow$ features selected by the feature selection
      algorithm
    for $j = 1$ to $M$ do
        $DM_{ij} \leftarrow$ 1 if jth feature is in $F_i$, otherwise $DM_{ij} \leftarrow$
          0
    end for
end for
  //Obtaining a threshold, $Th$, to select the feature subset
  $v_j \leftarrow \sum_i DM_{ij}$
for $Th = min(v)$ to $max(v)$ do
    $F_{Th} \leftarrow$ subset of features with $v_j > Th$
    $FR \leftarrow$ calculate Fisher ratio for each feature within
      $F_{Th}$
    $CM \leftarrow \frac{1}{\sum_j FR_j}$
    $R \leftarrow$ Ratio of retained features
    $E(Th) \leftarrow \alpha \times CM + (1 - \alpha) \times R$
end for
$Th \leftarrow min(E)$, $Th$ is the value, which minimizes the
  function $E$
$FS \leftarrow$ subset of features including the features with
  $v_i > Th$
return $FS$
```

TABLE II
DATASETS AND NUMBER OF SELECTED SUBJECTS AND SEGMENTS USED FOR TRAINING

Training Dataset	Total No. of Sub.	No. of AF Sub.	No. of non-AF Sub.	Total No. of Seg.	No. of AF Seg.	No. of non-AF Seg.
Clinical Trial	20*	1	20	2793	806	1987
AF Trial	18	17	1	439	394	45
DeepBeat	15	15	0	200	200	0

*One subject has both AF and non-AF segments. (Sub.=Subject. Seg.=Segment.)

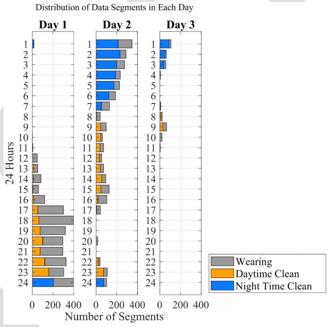

Fig. 3. Distribution of the non-AF training data from the two first 24-hour periods of 14 days clinical trial data collection.

TABLE III
NUMBER OF SUBJECTS AND PPG SEGMENTS UTILIZED AS DEPENDENT AND INDEPENDENT TEST DATA

Test Dataset	Compared Methods	Total No. of Sub. (AF/non-AF)	Total No. of Seg. (AF/non-AF)
Clinical Trial (Dependent Sub.)	Our approach, Method II [9], Method I [18]	16* (1/16)	4013 (373/3639)
Clinical/AF Trial (Independent Sub.)	Our approach, Method II, Method I	20 (1/19)	4874 (1674/3180)
DeepBeat	Our approach, Method III [15]	84 (68/16)	1124 (704/420)
UMMC Simband	Our approach, Method III, Method II, Method I	8 (8/0)	403 (403/0)
MIMICIII	Our approach, Method III, Method II	5 (5/0)	1981 (1981/0)

*One subject has both AF and non-AF segments.

of the 14 days of the clinical trial data collection. The training dataset from the AF trial was comprised of 439 30-sec PPG segments, including AF and non-AF data. In addition, we used the dataset provided by Stanford University (henceforth referred to simply as DeepBeat dataset). We randomly selected 200 25-sec PPG segments from the segments showing AF in the DeepBeat dataset.

2) Test Datasets: Clinical trial dependent test data are the left-out data originating from the subjects whose data were used for training. Independent test data are the sampled data from participants whose data were not employed in the training procedure. Table III reports the number of subjects and PPG segments used as dependent and independent test data. To sample the test data from the clinical trial dataset, the participants' data recorded during 24 hours of the day were split into 6 blocks (each includes 4 hours) based on the daytime and nighttime definition in this study. Ten 30-sec (i.e. 5 min) segments were selected from each block of a day, which yielded sixty 30-sec segments for each day of each subject. As the time duration of the data collection was 14 days, ideally, the total number of segments would be 840 segments for each subject. However, there might be blocks with no recorded data, since the daily adherence of the Pulsewatch participants to the Pulsewatch system was less than perfect. Fig. 4 illustrates the distribution of the clinical trial data sampled for testing the model during 14 days.

The DeepBeat testing dataset was randomly drawn from AF and non-AF data segments from held-out DeepBeat subjects. Further, five and eight subjects with AF from MIMIC III and

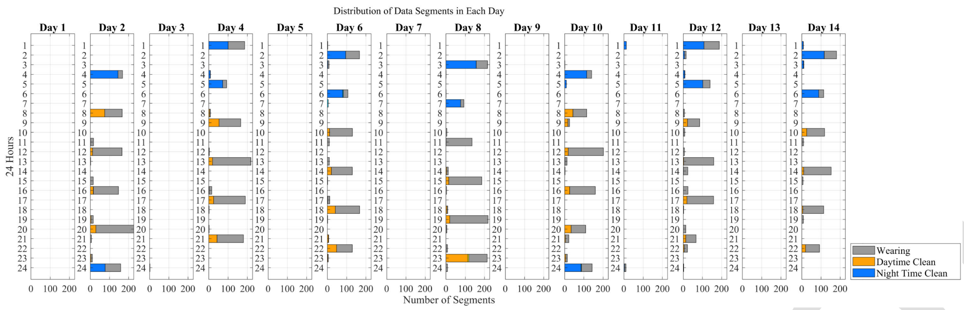

Fig. 4. Distribution of the testing data in 24 hours of day during 14 days.

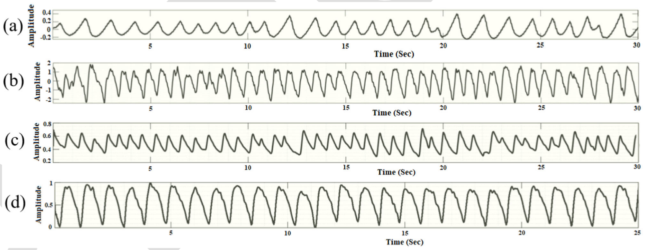

Fig. 5. PPG segments from different datasets: (a) Clinical trial, (b) UMMC Simband, (c) MIMIC III, (d) DeepBeat.

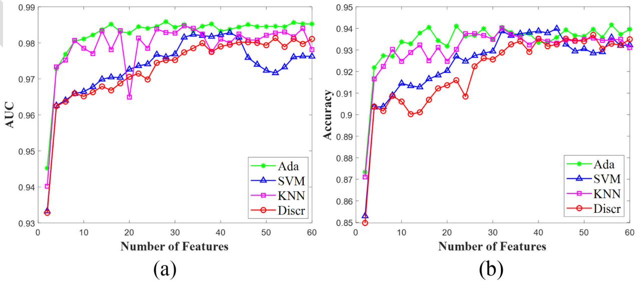

Fig. 6. Classification (a) AUC, (b) Accuracy in terms of number of selected features for different classifiers: AdaBoost (Ada), SVM, KNN, and discriminant analysis (Discr).

UMMC Simband datasets, respectively, were used as external independent test datasets in our study. Since the ICU recording for each subject in the MIMIC III dataset contained hundreds of hours of data, we only used the data from five subjects, whose data had already been prepared for an AF study, in which cardiologists adjudicated the presence of AF in those recordings [18], [31].

3) Classification Algorithms: Four classification algorithms—AdaBoost (decision trees), SVM, KNN, and discriminant analysis were compared to identify the optimal size of the feature subset. The classification abilities of the constructed model were compared by estimation of seven statistical indices: accuracy (Acc), sensitivity (Sens), specificity (Spec), positive predictive value (PPV), negative predictive value (NPV), G-mean, and F-measure (F-meas). Positives were noisy and negatives were clean segments. All algorithms were implemented in Matlab 2020b and 2021a using the Statistics and Machine Learning Toolbox (Mathworks Inc., USA).

III. Results

In this section, we give an overview of the most significant results. In the first subsection, the data preparation and feature selection results are presented, while the second subsection summarizes the performance of the signal quality assessment approach and comparison of the findings to the other methods proposed in the previous studies.

We present the results compared to three methods: Method I [18], Method II [9], and Method III, which is a deep neural network method called DeepBeat [15], via implementing their algorithms on the appropriate test datasets. We selected these methods as they are representative of the state-of-the-art heuristic (time domain), combined machine learning-heuristic (time-frequency domain), and DNN frameworks, respectively. As these approaches have been only adjusted or trained based on specific PPG data (with certain waveform) recorded using a particular device, we used the congruous data for testing to make fair comparison across the test datasets. PPG waveforms of individual datasets recorded by different devices is shown in Fig. 5.

A. Feature Subset Size

As mentioned earlier, the ensemble feature selection method does not specify the number of features, but rather a ranked list of them as the final feature list. In order to determine the optimal feature subset size, a classification model was trained and evaluated for each feature number.

Fig. 6 displays the average accuracy and Area under the ROC Curve (AUC) over a varied number of features via 5-fold cross validation by the mentioned classifiers. Performance was measured using AUC, where the receiver operating characteristic (ROC) curve itself is a plot of True Positive Rate (TPR) versus False Positive Rate (FPR). A growing trend of both AUC and accuracy can be observed at the beginning of the curves for the lower number of features in all classifiers. AdaBoost classifier achieved stable performance in terms of both AUC and accuracy (values higher than 98% and 93%, respectively) when the number of selected features was more than 12. We therefore used AdaBoost with at least 12 features as the proposed classifier and feature size, respectively.

A common drawback of feature selection algorithms is that the large subset size still shows the highest performance value

TABLE IV
PERFORMANCE OF THE QUALITY ASSESSMENT METHODS FOR CLINICAL-TRIAL DEPENDENT SUBJECTS' TEST SET

Method	PPG Type	No. of Segments	TP	TN	FP	FN	Sens	Spec	PPV	NPV	G_mean	Acc	F_meas
Our approach	Total	3924	1975	1717	44	188	91.31	**97.50**	**97.82**	90.13	**94.36**	**94.09**	**94.46**
	AF	284	176	92	13	3	98.32	**87.62**	**93.12**	96.84	**92.82**	**94.37**	**95.65**
	Non-AF	3640	1799	1625	31	185	90.68	**98.13**	**98.31**	89.78	**94.33**	**94.07**	**94.34**
Method II	Total	3924	2076	919	842	87	**95.98**	52.19	71.15	**91.35**	70.77	76.33	81.72
	AF	284	179	6	99	0	**100.00**	5.71	64.39	**100.00**	23.90	65.14	78.34
	Non-AF	3640	1897	913	743	87	**95.61**	55.13	71.86	**91.30**	72.61	77.20	82.05
Method I	Total	3924	2068	766	995	95	95.61	43.50	67.53	88.97	64.49	72.23	79.15
	AF	284	175	64	41	4	97.77	60.95	81.02	94.12	77.19	84.15	88.61
	Non-AF	3640	1893	702	954	91	95.41	42.39	66.49	88.52	63.60	71.29	78.37

TABLE V
PERFORMANCE OF THE QUALITY ASSESSMENT METHODS FOR CLINICAL/AF-TRIAL INDEPENDENT SUBJECTS' TEST SET

Method	PPG Type	No. of Segments	TP	TN	FP	FN	Sens	Spec	PPV	NPV	G_mean	Acc	F_meas
Our approach	Total	4874	2547	2022	53	252	91.00	**97.45**	**97.96**	88.92	**94.17**	**93.74**	**94.35**
	AF	1674	864	707	36	67	92.80	**95.15**	**96.00**	91.34	**93.97**	**93.85**	**94.37**
	Non-AF	3180	1683	1295	17	185	90.10	**98.70**	**99.00**	87.50	**94.30**	**93.65**	**94.34**
Method II	Total	4874	2738	742	1333	61	**97.82**	35.76	67.26	**92.40**	59.14	71.40	79.71
	AF	1674	931	33	710	0	**100.00**	4.44	56.73	**100.00**	21.07	57.59	72.40
	Non-AF	3180	1807	701	611	61	**96.73**	53.43	74.73	**91.99**	71.89	78.87	84.32
Method I	Total	4874	2707	955	1120	92	96.71	46.02	70.73	91.21	66.72	75.13	81.71
	AF	1674	918	308	435	13	98.60	41.45	67.85	95.95	63.93	73.24	80.39
	Non-AF	3180	1789	646	666	79	95.77	49.24	72.87	89.10	68.67	76.57	82.77

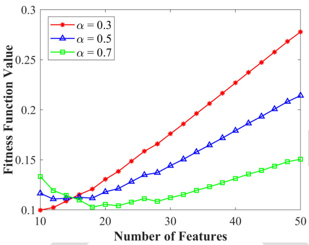

Fig. 7. Fitness function values for different numbers of features and α.

(as can be observed in Fig. 6). Hence, we used the fitness value criterion to minimize both the complexity and feature subset size as much as possible, without reducing the performance. Fig. 7 shows the obtained values of fitness function for different numbers of features and α values 0.3, 0.5, and 0.7. According to the figure, a feature subset size of 18 is the optimum value which minimizes the fitness function. Thus, we selected 18 as the optimum number of features, which was obtained with $\alpha = 0.7$. Although we cannot recommend an optimal value for α, as a general rule of thumb, we suggest that if the goal is to reduce the complexity measure at the cost of a slight increase in dimensionality, 0.7 is a suitable value for α.

B. Model Evaluation on Test Datasets

After training the AdaBoost classifier using the most suitable feature subset (18 features), we tested our model on the test dataset to quantitatively explore its performance. To compare to the previous studies,The clinical/AF trial and UMMC Simband datasets were used for Method I evaluation. Clinical/AF trial, UMMC Simband and MIMIC III were used as the test datasets for Method II. For Method III, the UMMC Simband, DeepBeat, and MIMIC III datasets were used to evaluate the model's performance (see Table III).

1) Clinical and AF Trial Test Results: Tables IV and V provide the quality assessment performance of the classifier model on clinical trial dependent and independent test datasets. The results illustrate high performance for both dependent and independent datasets (accuracy > 0.93, sensitivity > 0.90) independently of heartbeat type (AF or non-AF). Further, the specificity rate is higher than 0.98 and 0.86 for non-AF and AF segments, respectively.

Compared to Method I and Method II, our approach indicated a higher accuracy for total, AF, and non-AF segments in both dependent and independent test datasets. In comparison to Method I, our approach yielded much higher specificity and PPV, and comparable sensitivity and NPV values. Although Method II showed higher sensitivity and NPV (by reducing the false negatives), it achieved this at the loss of specificity and PPV, leading to the low values of G-mean and F-measure.

2) DeepBeat, UMMC Simband, and MIMICIII Test Results: We examined the classifier performance on held-out subjects from the DeepBeat dataset The results shown in Table VI demonstrate the consistently high performance of our approach (accuracy > 0.91) independently of heartbeat type (AF or non-AF). Further, sensitivity and specificity rates are higher than 0.86 and 0.98, respectively, across total AF and non-AF segments. A comparison with the Method III is also provided in Table VI. Although the sensitivity and NPV are higher for Method III, the high number of false positives effectively reduces the other classification indices.

Performance of the quality assessment for Simband test data can be found in Table VII. To assess the quality of the Simband data, the combination of PPG and ACC signals were used in this study. An important source of artifact in PPG signals in

TABLE VI
PERFORMANCE OF THE QUALITY ASSESSMENT METHODS FOR DEEPBEAT DATASET

Method	PPG Type	No. of Segments	TP	TN	FP	FN	Sens	Spec	PPV	NPV	G_mean	Acc	F_meas
Our approach	Total	1124	539	490	9	86	86.24	**98.20**	**98.36**	85.07	**92.02**	**91.55**	**91.90**
	AF	704	391	251	8	54	87.87	**96.91**	**97.99**	82.30	**92.28**	**91.19**	92.65
	Non-AF	420	148	239	1	32	82.22	**99.58**	**99.33**	88.19	**90.49**	**92.14**	**89.97**
Method III	Total	1124	604	395	104	21	**96.64**	79.16	85.31	**94.95**	87.46	88.88	90.62
	AF	704	430	208	51	15	**96.63**	80.31	89.40	**93.**	88.09	90.63	**92.87**
	Non-AF	420	174	187	53	6	**96.67**	77.92	76.65	**96.89**	86.79	85.95	85.50

TABLE VII
PERFORMANCE OF THE QUALITY ASSESSMENT METHODS FOR INDEPENDENT SUBJECTS WITH AF FROM UMMC SIMBAND DATASET

Method	No. of Segments	TP	TN	FP	FN	Sens	Spec	PPV	NPV	G_mean	Acc	F_meas
Our approach	403	321	59	10	13	96.11	85.51	96.98	81.94	90.65	94.29	96.54
Method III	403	318	63	6	16	95.21	**91.30**	**98.15**	79.75	**93.24**	**94.54**	**96.66**
Method II	403	326	5	64	8	**97.60**	7.25	83.59	38.46	26.59	82.13	90.06
Method I	403	326	52	17	8	**97.60**	75.36	95.04	**86.67**	85.77	93.80	96.31

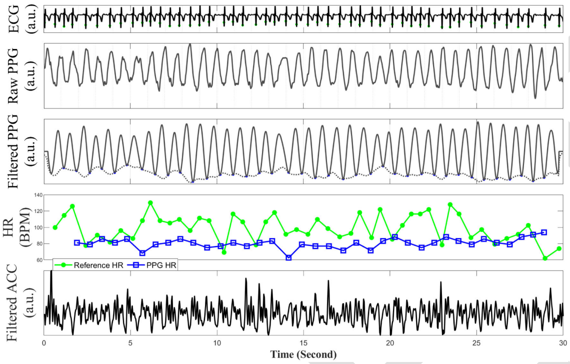

Fig. 8. An example of cyclical noise in UMMC Simband segment. The PPG segment seems to be clean, but the ACC and misaligned heart rates (of reference ECG and PPG) indicate the motion artifact corrupted signal.

wearable devices is attributed to the air gaps created between the skin and sensor during physical activity. High amplitude cyclical movement can cause quasi-periodic waves resembling the PPG signals (see Fig. 8). Therefore, it is necessary to recognize the segments which are corrupted by cyclical movement and classify them as noisy segments. Obviously, wearable device movements can be detected using an accelerometer, as the magnitude of the ACC signal changes significantly with sensor movement.

Hence, in this study, a threshold-based artifact detection approach was performed using the ACC signal to detect data segments which have been corrupted by high amplitude motion artifacts, prior to PPG-based classification. Three features were extracted from ACC signals: mean absolute deviation, sum of time-domain energy of the signal, and sum of the signal power in the frequency domain. Appropriate thresholds were estimated for each feature, based on the non-AF cohort from the UMMC Simband dataset. Derived thresholds were applied to the testing data to detect segments with significant accelerometer motion, to mark them as artifact-corrupted segments, as a primary step before the PPG-based quality assessment. Table VII represents the evaluation results on the UMMC Simband testing dataset. As can be observed, Method I, Method III and our approach exhibit comparable performance, while Method II represents very low performance in terms of specificity, NPV, and G-mean due to the high number of false positives.

The quality assessment results for MIMIC III test data are shown in Table VIII. Accuracy, sensitivity, and specificity of our approach are higher than 0.95, 0.83, and 0.98 for the PPG test subset from MIMIC III AF subjects. Accordingly, the results related to this dataset demonstrate the superiority of our approach compared to the Method II and Method III, which showed very low performance in terms of specificity and PPV.

IV. DISCUSSION AND CONCLUSIONS

Heart rhythm monitoring of cardiac patients requires reliable quality of the signals recorded from patients during their monitoring, screening, or treatment period. The main objective of this study is to provide labels demonstrating the PPG segments suitable for further processing, e.g., for HR value estimation and AF detection.

In this study, we proposed a comprehensive approach to employ the most relevant features, which have the capability to differentiate significantly between clean and corrupted PPG segments. In our approach, a combination of different types of features was used to capture various characteristics of the PPG signal. Then, the ensemble feature selection and vote threshold aggregation methods were used to provide the optimal feature subset which enhanced the resultant performance of the signal quality assessment compared to the previous studies. This is especially evident in the achievement of high performance of the quality assessment for both AF and non-AF segments from different test datasets.

Multiple studies have been conducted to assess the quality of the PPG signals using various methods, such as machine learning, deep learning, and heuristic rules-based methods. Various types of fiducial and non-fiducial features have been used in previous PPG signal analysis studies. Statistical, morphological, energy, temporal, and time-frequency attributes of the PPG signals have been widely used for PPG SQA, and arrhythmia and HR detection [8], [12], [18], [32]. The main benefit of using non-fiducial features is to eliminate the risk of fiducial

TABLE VIII
PERFORMANCE OF THE QUALITY ASSESSMENT METHODS FOR INDEPENDENT SUBJECTS WITH AF FROM MIMIC III DATASET

Algorithms	No. of Segments	TP	TN	FP	FN	Sens	Spec	PPV	NPV	G_mean	Acc	F_meas
Our approach	1984	274	1628	27	55	83.28	**98.37**	**91.03**	96.73	**90.51**	**95.86**	**86.98**
Method III	1984	295	1156	499	34	89.67	69.85	37.15	97.14	79.14	73.14	52.54
Method II	1984	328	278	1377	1	**99.70**	16.80	19.24	**99.64**	40.92	30.54	32.25

point detection errors. On the other hand, using data-driven template-based features leads to taking into account morphological characteristics of the PPG signals.

While these studies have attempted to discriminate between clean and artifact-corrupted signals, none of them have investigated the effectiveness of their techniques on a wide range of PPG signal types, including long-term real-life PPG data recordings as well as publicly available datasets recorded using various types of recording devices from patients with different types of arrhythmias, such as AF, PAC, and PVC. Pereira *et al.* [12] proposed a machine learning approach for quality assessment of the 30-sec PPG segments collected from patients admitted to the neuro and general ICU, in which the neuro ICU data included at most 22 hours of continuous PPG signals. However, their proposed model may not be appropriate for real-life PPG data, when the PPG waves might be distorted substantially by participants' physical activity and motion. Further, the heart rate variability (HRV) is considerably affected by daily physical activity during long-term recordings. To capture all these alterations, we used the template-based features, which reflect any kind of individual hemodynamic characteristics as well as waveform and HR variations occurring during the 14 days recording time.

We also, directly compared our approach to other methods proposed in previous studies, whose models have been developed based on other databases. Using congruous test datasets, Method I and Method II showed a low number of false negatives at the cost of a very high number of false positives. Our approach indicated acceptable false negatives and much lower false positives (compared to Method I and Method II), which leads us to increase the coverage (usability) of the clean PPG segments for further processing for diagnosis and treatment. We achieved more robustness and consistency against the PPG signal variations caused by different recording devices across all datasets. Particularly, the limitation for Method I as a heuristic method is that it is required to adjust the features' threshold values for each individual dataset to maintain the high performance.

The limitation of using individual datasets is more crucial for deep learning-based approaches. Deep learning model in Method III was pretrained using convolutional denoising autoencoder on over one million simulated physiological signals. Using this large amount of training data, it is expected that the model has been trained on diverse PPG waveforms and morphologies. The model is supposed to have the capability to be generalizable to apply to various PPG signals recorded by different devices. The results illustrated that even though the Method III showed high performance for UMMC Simband and DeepBeat datasets, its performance on the MIMIC III dataset was low. In addition, it failed when it was evaluated with the clinical/AF trial test dataset. This might be due to the differences of the PPG waveforms of its training dataset with the Pulsewatch and MIMIC III datasets. Therefore, the association of the model performance to the training data waveform is the main drawback of the Method III that restricts its usability to the specific data.

One limitation of the present study is that we only used a limited number of data segments from each dataset as testing data due to the annotation burden. The datasets used in this study contain thousands of data segments recorded from hundreds of subjects. The way we selected in this study to address this limitation was to use the randomized sampling of the data segments to approximate the algorithm performance on the whole dataset (as has been described in section F).

Consequently, the main contributions of this investigation can be summarized by two aspects. First, we proposed an optimized feature selection scheme to provide the feature subset as a combination of various types of features. We improved the IWSSr algorithm in the backward step to eliminate the redundant and irrelevant features and obtain the most efficient feature subset. In addition, by proposing a complexity measure based on the inverse of Fisher's ratio, the optimum number of features was estimated that maximizes the discriminative power of the feature subset.

Secondly, the study was mainly performed using the Pulsewatch dataset collected from a large number of cardiac patients with a history of stroke/transient ischemia during 14 days of recording. We examined the robustness of our approach for normal and arrhythmic PPG data recorded in real-life conditions with varied levels of noise and artifacts. In addition to the dependent dataset, our approach was evaluated on independent external test datasets to account for a wide variety of PPG wave morphologies caused by different recording devices. The results indicated the high discrimination ability of our constructed model for all test datasets and, more importantly, its reproducibility and generalizability for arrhythmic PPG signals. Particularly of note is the performance on AF PPG data, which might be potentially mis-detected as noisy non-AF PPG signals.

ACKNOWLEDGMENT

We would like to thank the generous support of Samsung Semiconductor for the use of Simbands for data collection in this study. The details of Samsung Simband can be found on: https://www.simband.io.

REFERENCES

[1] C.-Y. Lin *et al.*, "Prognostic significance of premature atrial complexes burden in prediction of long-term outcome," *J. Amer. Heart Assoc.*, vol. 4, no. 9, 2015, Art. no. e002192.

[2] A. Sološenko, A. Petrėnas, and V. Marozas, "Photoplethysmography-based method for automatic detection of premature ventricular contractions," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 5, pp. 662–669, Oct. 2015.

[3] J. A. Sukor, S. Redmond, and N. Lovell, "Signal quality measures for pulse oximetry through waveform morphology analysis," *Physiol. Meas.*, vol. 32, no. 3, 2011, Art. no. 369.

[4] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiol. Meas.*, vol. 33, no. 9, 2012, Art. no. 1491.

[5] X. Sun, P. Yang, and Y.-T. Zhang, "Assessment of photoplethysmogram signal quality using morphology integrated with temporal information approach," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 3456–3459.

[6] W. Karlen *et al.*, "Photoplethysmogram signal quality estimation using repeated gaussian filters and cross-correlation," *Physiol. Meas.*, vol. 33, no. 10, 2012, Art. no. 1617.

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

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

[9] D. Dao *et al.*, "A robust motion artifact detection algorithm for accurate detection of heart rates from photoplethysmographic signals using time-frequency spectral features," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 5, pp. 1242–1253, Sep. 2017.

[10] S. Vadrevu and M. S. Manikandan, "Real-time PPG signal quality assessment system for improving battery life and false alarms," *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 66, no. 11, pp. 1910–1914, Nov. 2019.

[11] E. Sabeti, N. Reamaroon, M. Mathis, J. Gryak, M. Sjoding, and K. Najarian, "Signal quality measure for pulsatile physiological signals using morphological features: Applications in reliability measure for pulse oximetry," *Informat. Med. Unlocked*, vol. 16, 2019, Art. no. 100222.

[12] T. Pereira *et al.*, "A supervised approach to robust photoplethysmography quality assessment," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 3, pp. 649–657, Mar. 2020.

[13] E. L. Dickson *et al.*, "Smartwatch monitoring for atrial fibrillation after stroke-the pulsewatch study: Protocol for a multiphase randomized controlled trial," *Cardiovasc. Digit. Health J.*, vol. 2, no. 4, pp. 231–241, 2021.

[14] D. Han *et al.*, "Premature atrial and ventricular contraction detection using photoplethysmographic data from a smartwatch," *Sensors*, vol. 20, no. 19, 2020, Art. no. 5683.

[15] J. Torres-Soto and E. A. Ashley, "Multi-task deep learning for cardiac rhythm detection in wearable devices," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–8, 2020.

[16] A. E. Johnson *et al.*, "MIMIC-III, A freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[17] S. K. Bashar, M. B. Hossain, E. Ding, A. J. Walkey, D. D. McManus, and K. H. Chon, "Atrial fibrillation detection during sepsis: Study on MIMIC III ICU data," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3124–3135, Nov. 2020.

[18] S. K. Bashar *et al.*, "Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, 2019.

[19] D. Han *et al.*, "Smartwatch ppg peak detection method for sinus rhythm and cardiac arrhythmia," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 4310–4313.

[20] N. Selvaraj *et al.*, "Statistical approach for the detection of motion/noise artifacts in photoplethysmogram," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 4972–4975.

[21] B. S. Everitt and A. Skrondal, *The Cambridge Dictionary of Statistics*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[22] P. Bermejo, J. A. Gámez, and J. M. Puerta, "Incremental wrapper-based subset selection with replacement: An advantageous alternative to sequential forward selection," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, 2010, pp. 367–374.

[23] M. Moradkhani *et al.*, "A hybrid algorithm for feature subset selection in high-dimensional datasets using FICA and IWSSr algorithm," *Appl. Soft Comput.*, vol. 35, pp. 123–135, 2015.

[24] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans.Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[25] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Inf. Fusion*, vol. 52, pp. 1–12, 2019.

[26] B. Pes, "Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 5951–5973, 2020.

[27] Q. Shen, R. Diao, and P. Su, "Feature selection ensemble," *Turing-100*, vol. 10, pp. 289–306, 2012.

[28] A. D. Haro García, *Scaling Data Mining Algorithms. Application to Instance and Feature Selection*, Granada: Universidad de Granada, 2012.

[29] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos, "Centralized vs. distributed feature selection methods based on data complexity measures," *Knowl.-Based Syst.*, vol. 117, pp. 27–45, 2017.

[30] X. Let, *Pattern Classification*. Wiley-Interscience Publication, 2001.

[31] S. K. Bashar *et al.*, "Noise detection in electrocardiogram signals for intensive care unit patients," *IEEE Access*, vol. 7, pp. 88357–88368, 2019.

[32] L. M. Eerikäinen *et al.*, "Comparison between electrocardiogram-and photoplethysmogram-derived features for atrial fibrillation detection in free-living conditions," *Physiol. Meas.*, vol. 39, no. 8, 2018, Art. no. 084001.