Expert Systems With Applications 237 (2024) 121611

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Atrial fibrillation detection on reconstructed photoplethysmography signals collected from a smartwatch using a denoising autoencoder

Fahimeh Mohagheghian [a,*], Dong Han [a], Om Ghetia [a], Darren Chen [a], Andrew Peitzsch [a], Nishat Nishita [b], Eric Y. Ding [c], Edith Mensah Otabil [c], Kamran Noorishirazi [c], Alexander Hamel [c], Emily L. Dickson [d], Danielle DiMezza [c], Khanh-Van Tran [c], David D. McManus [c], Ki H. Chon [a]

[a] Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA
[b] Department of Public Health Sciences, University of Connecticut Health, Farmington, CT, USA
[c] Division of Cardiology, University of Massachusetts Medical School, Worcester, MA, USA
[d] College of Osteopathic Medicine, Des Moines University, Des Moines, IA, USA

ARTICLE INFO

Keywords:
Photoplethysmography
Atrial fibrillation
Deep learning
Denoising
Autoencoder

ABSTRACT

Photoplethysmography (PPG) signals collected by wearables have been shown to be effective in accurate detection of atrial fibrillation (AF), provided that the data are devoid of motion and noise artifacts (MNA). Many studies have been previously conducted to detect AF arrhythmia using PPG data; however, the subjects were mostly in clinics or controlled settings with data collection lasting several minutes to at most several hours with minimal MNA. Our study, Pulsewatch, differs from previous AF studies in that PPG data from smartwatches prescribed to stroke survivors were continuously collected for two weeks in real-life conditions, which invariably included a significant amount of MNA. Our aim is to provide a framework for a novel use of a denoising autoencoder to reconstruct motion-artifact-removed PPG signals so that we can improve the AF detection performance and to increase the amount of analyzable data.

We used more than 30,000 25-sec PPG segments from 129 subjects randomly selected from Pulsewatch and Stanford University's datasets. The training and testing datasets from these two databases came from smartwatches from different vendors with varying sampling frequencies and time duration of recordings in diverse and realistic settings. In this study, the highly corrupted PPG data were automatically detected and discarded, but those segments contaminated with low-to-moderate motion and noise artifacts (MNA) were subjected to a convolutional denoising autoencoder (CDA). To reconstruct the artifact-removed PPG segments, we proposed to employ two distinct CDA models for AF and non-AF data groups initially classified as AF or non-AF. Using the proposed approach, we significantly improved the performance of detecting occult AF. We achieved classification accuracy, sensitivity, and specificity of 91.02%, 91.54%, and 90.85%, respectively, for out-of-sample test data from both databases. By sanitizing data from low-to-moderate MNA, we were able to increase the usable data coverage by 21%.

1. Introduction

As one of the most prevalent cardiac arrhythmias, atrial fibrillation (AF) may increase the likelihood of stroke and heart failure if not promptly diagnosed and appropriately treated. It is generally acknowledged that the incidence and prevalence of AF are growing due to the increasing number of elderly people in the U.S. (Colilla et al., 2013). It is estimated that 6–12 million people in the US and 17.9 million people in Europe will suffer from AF in the year 2050 and 2060, respectively (Lippi, Sanchis-Gomar, & Cervellin, 2021). AF often occurs fleetingly with minimal symptoms, making diagnosis time intensive and challenging. It is not feasible for cardiologists to manually screen the large

* Corresponding author.
E-mail addresses: fahimeh.mohagheghian@uconn.edu (F. Mohagheghian), dong.han@uconn.edu (D. Han), om.ghetia@uconn.edu (O. Ghetia), darren.3.chen@uconn.edu (D. Chen), andrew.peitzsch@uconn.edu (A. Peitzsch), nishita@uchc.edu (N. Nishita), eric.ding@umassmed.edu (E.Y. Ding), edith.mensahotabil@umassmed.edu (E. Mensah Otabil), alexander.hamel@umassmed.edu (A. Hamel), emily.l.dickson@dmu.edu (E.L. Dickson), danielle.dimezza@umassmed.edu (D. DiMezza), Khanh-Van.Tran@umassmemorial.org (K.-V. Tran), david.mcmanus@umassmed.edu (D.D. McManus), ki.chon@uconn.edu (K.H. Chon).

https://doi.org/10.1016/j.eswa.2023.121611
Received 9 December 2022; Received in revised form 17 August 2023; Accepted 11 September 2023
Available online 17 September 2023

amount of data collected during long-term AF monitoring, such as that recorded from a Holter or mobile cardiac outpatient telemetry system (Seet, Friedman, & Rabinstein, 2011). Thus, a smart computer-aided system is required to automatically screen the data and identify the episodes that contain AF.

Photoplethysmography (PPG) signals extensively collected by wearable devices are able to support the diagnosis of cardiac arrhythmias, including atrial fibrillation (AF). Smartwatches are widely available and provide us with an efficient method for long-term AF detection using PPG data. This capability makes them appropriate for long-term monitoring of patients who are at risk of cardiac arrhythmias or have a history of heart disease (Pereira et al., 2020). However, the PPG data recorded by wearables are highly prone to movement artifacts, which can significantly affect the morphology of PPG waveforms and consequently disrupt accurate arrhythmia detection, e.g., they can lead to false AF alerts. There have been reports of generating anxiety in smartwatch users after receiving false AF alerts from a smartwatch (Rosman, Gehi, & Lampert, 2020), hence, development of more reliable AF detection algorithms for a smartwatch PPG signal is warranted.

To date, several studies have attempted to diagnose AF rhythms using photoplethysmogram (PPG) signals. In one study (Bonomi, Schipper, Eerikäinen, Margarito, Aarts, Babaeizadeh, de Morree, & Dekker, 2016), a wrist-based PPG AF detection algorithm was compared with clinical adjudications of AF episodes collected using a standard Holter ECG monitor. The algorithm achieved high performance during 24-hour monitoring of the patients. The authors concluded that accurate detection of daily life rhythm irregularities caused by AF using wrist-based wearable PPG device is feasible.

In a previous study by Pereira et al. (Pereira et al., 2020), the authors provided a review of signal processing methods and machine learning and statistical approaches to detect AF from PPG signals. They discussed the limitations and challenges of using PPG signals recorded from wearable devices in clinical applications.

In another study (Georgieva-Tsaneva, Gospodinova, & Cheshmedzhiev, 2022), the authors evaluated their proposed algorithm using simultaneous ECG and PPG recordings to support PPG-based clinical applications in diagnosis of cardiovascular diseases. The high performance of the PPG signal accuracy in terms of the number of peaks and mean RR intervals in comparison to the simultaneous ECG signal confirmed the validity of using PPG signals in cardio-diagnostics in clinical practice.

In most similar studies to the above, the data collection duration was limited to several minutes with minimal motion artifacts. The main drawback of such studies is that they do not consider the PPG signal waveform variations that occur during different activities in real-life conditions, and more importantly, the significant amount of motion artifacts that need to be dealt with. Another study attempted longer-duration AF detection using a smartwatch data limited to several hours duration (Shen, Voisin, Aliamiri, Avati, Hannun, & Ng, 2019). AF data were collected in ambulatory free-living conditions from a wrist-worn device recorded in 8- and 3-hour durations on average, for their two datasets. Using a 50-layer convolutional neural network, they achieved an area under the ROC curve (AUC) of 95% on test data in the presence of motion artifacts.

Using a commercial smartwatch on actual stroke patients in their daily lives for an approximately 2-week period is a radically harder data analysis challenge than the above-mentioned studies. Hence, in our present study on this real-world data, the aim was to provide a framework for removing low-to-moderately-corrupted motion artifact data to enhance the accuracy of AF detection and retain more usable data for AF analysis. As expected, we encountered many challenges, including sensing issues, from various types of motion artifacts. In our approach to address the sensing issues, we employed a denoising autoencoder to reconstruct corrupted PPG segments caused by subjects' activities and movements. Reconstruction of the corrupted PPGs is crucial, as a classifier is not able to accurately identify AF/non-AF segments when the data contain motion artifacts. Hence, reconstruction of artifact-removed PPG data not only increases AF detection accuracy, but it also enhances the usable data coverage (Pereira et al., 2020). This is important especially for paroxysmal AF where episodes occur randomly and can be brief. Therefore, we want to maximize the opportunity to detect these short and random-onset AF episodes even when the data segments contain motion artifacts.

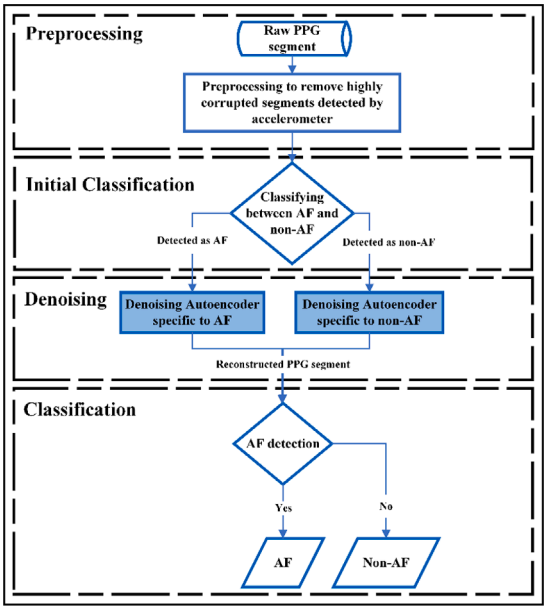

Fig. 1. Flowchart of our proposed approach.

In this study, we propose a deep convolutional denoising autoencoder (CDA) architecture to reconstruct low-to-moderately corrupted PPG segments so that we enhance both usable data coverage and the accuracy of AF detection. Our main contribution is that we employ two distinct CDA models, one specifically trained for AF and the other for non-AF data. For example, to train the CDA specific to AF data, the training data was comprised of 95% AF and 5% normal sinus rhythm (NSR) and vice versa, for the CDA specific to NSR data. Thus, in testing, a CDA model is specifically applied to either AF or non-AF data segments to effectively recover the corrupted PPG segments according to their rhythms. To identify which CDA model to use on a test segment, we initially classify the segment as either AF or non-AF so that the appropriate CDA model can be used to reconstruct the segment.

2. Methods

2.1. Methods overview

Fig. 1 shows the flowchart of our proposed approach. It contains three main components: preprocessing, using the denoising autoencoder, and AF/non-AF classification. First, 25-sec PPG segments were filtered using a bandpass filter. As intense movement fluctuations may lead to dramatic altering of the PPG waveforms and their frequency

Table 1
Number of clean/corrupted segments recognized by ACC signal analysis from Pulsewatch dataset.

Data	Total No. of segments	Highly corrupted segments (%)	Clean/Low-to-moderately corrupted segments (%)
Non-AF	38,145	20,375 (53.41%)	17,770 (46.59%)
AF	9,694	5,751 (59.32%)	3,943 (40.67%)
Total	47,839	26,126 (54.61%)	21,713 (45.39%)

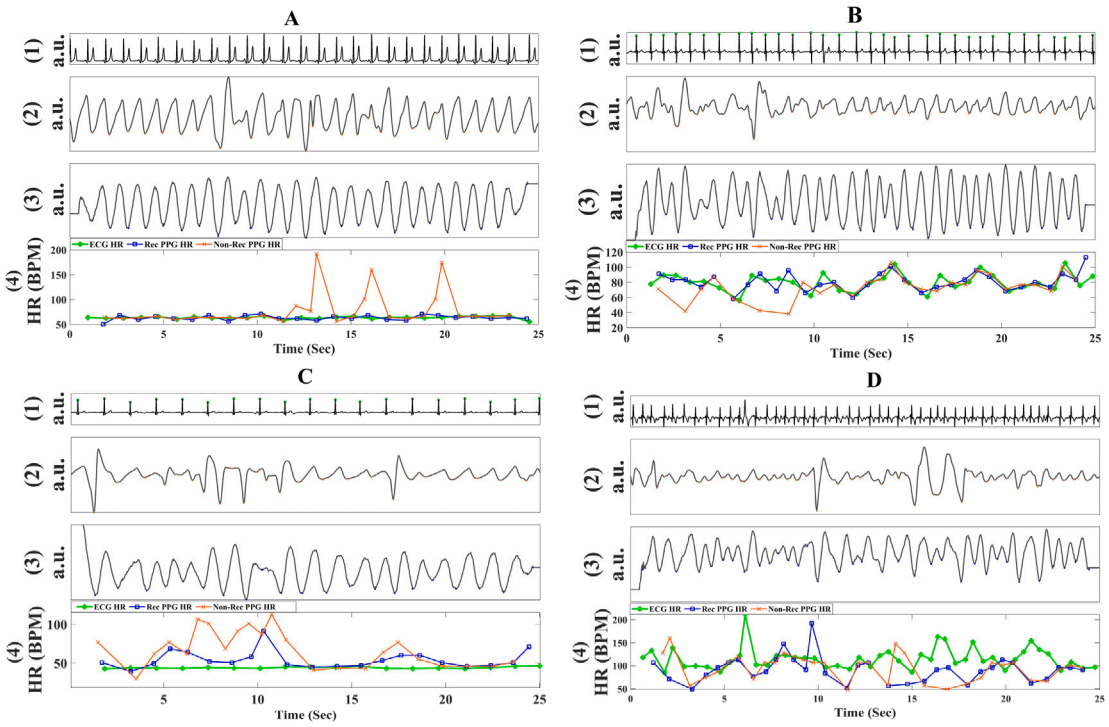

Fig. 2. (2) Non-reconstructed PPG segments (2nd rows), (3) reconstructed PPG segments (3rd rows), (1) corresponding ECG (1st rows), and (4) the extracted HRs (4th rows), A. Recoverable non-AF PPG segment; B. Recoverable AF PPG segment; C. Non-recoverable non-AF PPG segment; D. Non-recoverable AF PPG segment.

content, highly motion artifact-corrupted PPG segments were recognized via analyzing the corresponding time-aligned accelerometer (ACC) signal from the Pulsewatch dataset. Those PPG segments were then considered unrecoverable segments and were excluded from further processing. Hence, only the segments which were clean or low-to-moderately corrupted by motion artifacts were analyzed (see Table 1). We define low-to-moderately corrupted as those portions of PPG data that are intermittently corrupted with < 5 sec in duration noise within the 25-sec segment length. Fig. 2 shows representative recoverable and non-recoverable PPG segments, corresponding ECG, and the extracted HRs for both AF and non-AF PPG segments. The figure illustrates the reconstructed (Rec) and non-reconstructed (Non-Rec) PPG segments.

Afterward, an initial 1D deep learning-based classifier was applied to the bandpass-filtered PPGs to identify the AF and non-AF segments. The proposed specific CDA models were applied to the AF/non-AF segments to reconstruct the low-to-moderately-corrupted signals. The reconstructed PPG data from AF/non-AF denoising models were fed to the classifier to provide the final AF/non-AF designation for each PPG segment. The following sections describe the main components of our approach.

2.2. Dataset description

2.2.1. Data Collection: Clinical and AF trials

The data (including PPG, ACC, and ECG data) were collected in a multi-phase NIH-funded clinical trial called Pulsewatch (NCT number: NCT03761394, registered on December 3, 2018). Details of the Pulsewatch study can be found in (Dickson et al., 2021; Ding et al., 2021; Tran, Filippaios, Noorishirazi, Ding, Han, Mohagheghian, Dai, Mehawej, Wang, & Lessard, 2022). Only the data collected in Phase I of the clinical trial were used in this study. Phase I of the clinical trial aimed to validate the accuracy of AF detection algorithms implemented on a smartwatch by comparing them to a gold-standard ECG patch over 14 days. In total, 120 participants that had a previous history of stroke or transient ischemia (TI) were enrolled and the intervention group (90 patients) was randomized to use the Pulsewatch system. To be eligible to enroll in Pulsewatch, participants needed to be at least 50 years old, have had an ischemic stroke in the last decade, and have no major contraindications to anticoagulation therapy (Tran et al., 2022). Formal ethical approval for the Pulsewatch study was obtained from the Institutional Review Boards (IRBs) of the University of Massachusetts Chan Medical School (UMass Chan) and the University of Connecticut (UConn) (approval number H00016067). Informed consent was provided by all participants. The subjects' characteristics for the clinical trial dataset are shown in Table 2.

Under the ambulatory monitoring ("free living") conditions of the trial, participants wore a Galaxy Watch 3 or Samsung Gear S3 (Samsung, San Jose, CA, USA) smartwatch on their wrist, as well as a Cardea SOLO ECG patch (Cardiac Insight Inc., Bellevue, WA, USA) on their chest (Han et al., 2023). The single channel ECG data were sampled at 250 Hz and were used as the reference to adjudicate cardiac arrhythmia. The

Table 2
Subjects' characteristics for the clinical trial dataset.

Dataset	Age Mean ± STD years	Male/ Female (%)	Cardiac arrhythmias (%)	Hyperlipidemia (%)	Anti-arrhythmic drugs (%)	Beta-blockers (%)	Ca-blockers (%)	Anticoagulant (%)	Hypertension drug (%)
Clinical trial	65.1 ± 9.3	53/37 (58.9/ 41.1%)	12 (13.3%)	77 (85.5%)	2 (2.2%)	40 (44.4%)	19 (21.1%)	11 (12.2%)	51 (56.7%)

Table 3
Training and test datasets.

	Dataset	Total No. of Subjects	No. of AF Subjects	No. of Non-AF Subjects	Total No. of Segments	No. of AF Segments	No. of Non-AF Segments
Training	Pulsewatch	18	2	16	8193	3352	4841
	DeepBeat	21*	20	4	1188	1011	177
	Total	**39**	**22**	**20**	**9381**	**4363**	**5018**
Test	Pulsewatch	32*	23	10	21,713	4411	17,302
	DeepBeat	78**	28	75	3827	1024	2803
	Total	**110**	**51**	**85**	**25,540**	**5435**	**20,105**

* Training dataset: Three subjects from DeepBeat dataset have both AF and non-AF segments.
** Test dataset: One subject from Pulsewatch and 25 subjects from DeepBeat dataset have both AF and non-AF segments.

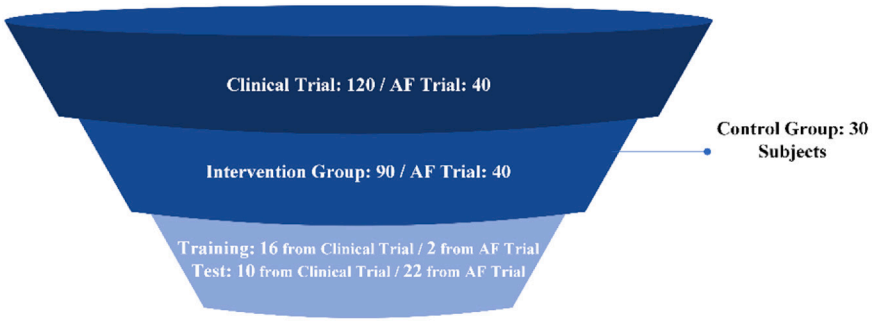

Fig. 3. Total number of subjects in Pulsewatch study (Clinical & AF trials) and the number of subjects used for training and test.

smartwatch data were composed of a single channel PPG signal and the magnitude of the accelerometer signal; both were automatically partitioned into 30-sec data segments at 50 Hz sampling frequency.

Since the AF population was relatively low in the clinical trial (only 5 participants presented with AF, i.e., ~5%, with a mean duration of 240 s), to have more balanced AF/non-AF data, our partner at UMass Chan conducted an additional AF trial to collect more AF data on 40 patients with confirmed persistent AF. Participants used the same devices under the same ambulatory conditions as the clinical trial; the only difference was duration of the experiment. Time duration was ~ 20 min for the first 20 participants, then increased to 7 days for the remaining 20 participants. This AF trial was approved by UMass Chan and UConn IRBs with number H00009953.

2.2.2. *Stanford University's PPG dataset*

The PPG data collected from participants undergoing elective stress or elective cardioversion (CV) tests have been provided by Stanford University as an open access PPG database. These data were used to develop a multi-task convolutional neural network (CNN) model called DeepBeat for noise and AF event detection (Torres-Soto & Ashley, 2020). The data were recorded at 128 Hz sampling frequency using a wrist-based wearable device (Simband) and segmented into overlapping 25-sec data segments. Data were collected from CV patients and the participants enrolled in an elective exercise stress test at Stanford hospital. Average recording time was about 20 min prior- and 20 min post-CV procedure for 132 participants confirmed with AF. Average data collection time was about 45 min for 42 subjects in the elective exercise stress test.

2.3. *Training and test datasets*

Table 3 shows the details of the data used for training and testing from both databases (Pulsewatch and Stanford), in this study. As Pulsewatch and Stanford University datasets contain a huge number of PPG segments (millions of data segments) from a large number of subjects, we randomly selected a limited number of subjects to evaluate our approach. From the Pulsewatch dataset, 29,906 PPG segments were used as training and testing data from 50 subjects. These 50 subjects were randomly selected from the clinical trial (intervention group) and AF trial. The testing subjects were independent, as they were randomly selected from the left-out subjects whose data were not used in the training stage. Fig. 3 indicates the number of subjects from the Pulsewatch study used for training and test. The number of testing data segments are based on those clean/low-to-moderately corrupted segments remaining after the ACC-based noise detection approach was applied (described in section 2.5.2).

Table 4
HR and time duration for training/test, AF/non-AF groups and the subject with PAC/PVC from Pulsewatch dataset.

Metrics	Train	Test	AF	Non-AF	PAC/PVC -Total Time Duration
	Mean (std)				
ECG duration-hours	318.63 (93.17)	283.26 (100.64)	179.34 (95.00)	331.73 (73.52)	389.2481
PPG duration-hours	174.00 (157.92)	54.97 (60.00)	63.32 (40.82)	135.45 (148.78)	158.75
AF duration-sec	17.25 (49.74)	80.96 (212.12)	240.64 (265.88)	0 (0)	0
HR from Cardea SOLO-BPM Mean (std)	71.62 (12.27)	69.55 (12.93)	79.94 (10.30)	68.59 (12.11)	65.45 (7.73)

In this study, the number of collected data segments from the Pulsewatch study varied from subject to subject (depending on a watch's wearing time). In addition, test data subjects were randomly selected from hold-out subjects (independent from subjects used for training). Hence, training and testing datasets were subject-oriented and not data segment-oriented.

As the Stanford University dataset (Torres-Soto & Ashley, 2020) (we called it DeepBeat dataset in this study) does not contain ACC signals, testing (AF and non-AF) data segments were randomly selected. Subsequently, 3,827 PPG segments visually annotated as not highly

Table 5
Total time duration of AF and non-AF PPG data segments in training and test datasets.

	Train AF	Train non-AF	Test AF	Test non-AF	Total
Total number of PPG segments	4363	5018	5435	20,105	34,921
Total time for PPG segments (hour)	34.95	41.57	43.87	163.65	284.04

corrupted (i.e., those with<5 sec of severe motion artifacts) were used for testing.

As can be seen in Table 3, two subjects with persistent AF from Pulsewatch were used for training. One subject with paroxysmal AF and twenty-two subjects with persistent AF from Pulsewatch were used in the testing dataset. One subject from the testing dataset had PAC/PVC (about 1% of total testing data segments). Table 4 indicates the HR and time duration for training/test, AF/non-AF groups and the subject with PAC/PVC from Pulsewatch dataset. Table 5 demonstrates the total time duration of AF and non-AF PPG data segments in our training and test datasets from both Pulsewatch and DeepBeat datasets.

2.4. Signal annotation

The data collected from the Pulsewatch study (clinical and AF trials) and the DeepBeat dataset consist of four different classes of PPG rhythms including NSR, PAC, premature ventricular contractions (PVC), and AF—based on the annotations adjudicated independently by PPG signal experts with arrhythmia discernibility. Annotations were performed manually by two people with hundreds of hours of experience reviewing PPG signals. The final segments' annotation was based on the consensus of adjudication of the two experts. When the two experts were in disagreement about a segment annotation, the opinion of a third expert was sought, and the final annotation was based on the view of the majority. The experts adjudicated the ECG data segments from the Pulsewatch dataset while the DeepBeat dataset was adjudicated by the visual inspection of the PPG pulse waveforms.

2.5. Preprocessing

2.5.1. Filtering

A bandpass filter of 0.5–6 Hz was applied to the PPG signals to remove the baseline wandering and high frequency noise, since the frequency range of PPG signals is usually within 0.5 to 5 Hz. A sixth-order zero-phase IIR Butterworth filter was used for bandpass filtration. Zero-mean unit-variance standardization of the PPG data was performed prior to further processing. As the DeepBeat dataset contained 25-sec PPG segments, we discarded the first 5 s of each 30-sec segment data of the Pulsewatch dataset so that the two datasets have the same length. PPG signal peaks were detected using the waveform envelope peak detection (WEPD) algorithm (Han, Bashar, Lazaro, Ding, Whitcomb, McManus, & Chon, 2019) (Han et al., 2020) (Han et al., 2022a) to enable comparison to the previous AF detection study (Bashar et al., 2019).

2.5.2. Discarding highly corrupted segments

A significant source of artifacts in PPG recordings by wearable devices is the air gaps created between the sensor and skin during movement and physical activities. Cyclical movements can generate quasi-periodic waveforms that may resemble PPG signal. In addition, sensor's movements may significantly deteriorate the waveforms so that they are not usable for analysis. Accordingly, an algorithm based on the ACC data is required to identify and exclude the unusable PPG segments which are designated as highly corrupted by both cyclical and non-cyclical movements.

Table 6
Performance of the ACC-based noise detection algorithm.

Dataset	Accu. (%)	Sens. (%)	Spec. (%)	PPV (%)	NPV (%)
Pulsewatch Validation	93.93	77.36	97.80	89.13	94.87

Wearable device movement can be detected using an embedded triaxial ACC, as the amplitude of the accelerometer signal alters significantly with sensor movement. To assess the motion-corrupted PPG data in this study, the ACC signals captured concurrently to the PPG recordings in our Pulsewatch data collection were used to detect the segments with high motion artifact content. A threshold-based artifact identification approach was taken on ACC signals to exclude the corresponding PPG segments with high levels of movement artifacts prior to the classification. Appropriate thresholds for ACC signals were derived from Pulsewatch training data to identify the high amplitude ACC signals with a time duration of more than 5 sec. We selected greater than 5 s out of the 25 s of each PPG data segment because the accuracy of most AF detection algorithms degrades when this condition occurs (Bashar et al., 2019). Performance of the ACC-based noise detection algorithm on the validation dataset is shown in Table 6.

2.6. Denoising autoencoder

In this study, we employed the convolutional denoising autoencoder (CDA) framework comprised of the fully convolutional and deconvolutional networks applied to 2D short-time Fourier transform (STFT) images. The encoder section eliminates the corruptions of the input data and projects it to the latent representation (Suk, Lee, & Shen, 2015). The decoder section then maps back to recover the denoised data from the latent representation to estimate the input data.

To improve the performance of our denoising autoencoder model, we used skip connections (also called residual connections) from the encoder layer to the corresponding symmetric decoder layer, in which the encoder layer feature maps are summed to the feature maps in the symmetric decoder layer. The parameters of the skip connections are learned during training to control the amount of information passed through the skip connections (He, Zhang, Ren, & Sun, 2016).

A skip connection with the multiplication factor of 1 (identity skip connection) can be represented as:

$$x_{l+1} = f_l(x_l) + x_k \tag{1}$$

where x_k and x_l denote the inputs of layers k and l, respectively; f_l is the activation function of layer l.

Skip connections provide more effective training of the mapping as they help the back-propagation of the gradients to the bottom layers of the network. Further, they simplify detail passing to the top layers to recover the spatial information lost by downsampling (Mao, Shen, & Yang, 2016). More specifically, autoencoders suffer from loss of information in image reconstruction. The main reasons for performance degradation in image restoration are: first, a large amount of image details might be lost or corrupted while passing through several convolutional layers, making image recovery an under-determined problem. Second, gradient vanishing often occurs during optimization in deep networks (Mao et al., 2016). As we applied the time–frequency representation of the PPG signals as the input and output of the autoencoder, we considered the reconstruction as an image restoration problem.

As the PPG waveforms and the frequency content of the AF and non-AF data groups are significantly different, it is more appropriate to train the CDA model specifically for each group. To be able to reconstruct AF and non-AF PPG segments, CDAs were trained using specified combinations of AF and non-AF training data, as described in section 3.1. This strategy allowed us to reconstruct the testing PPG waveforms based on

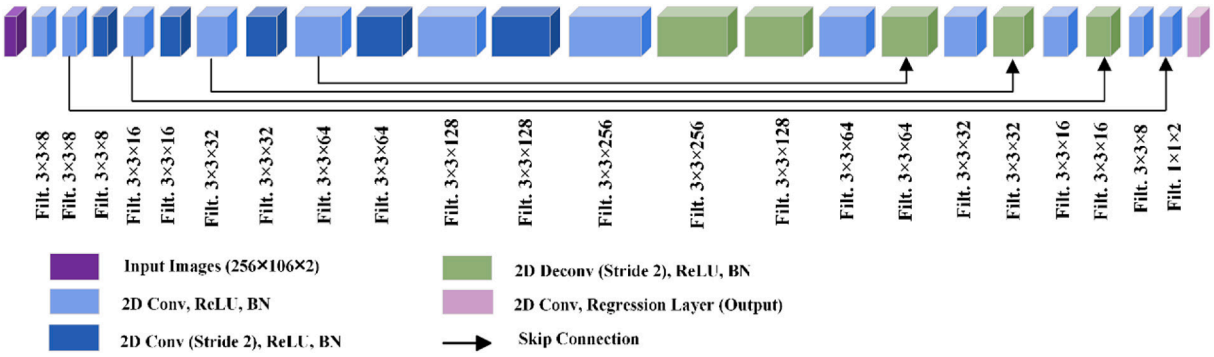

Fig. 4. The proposed CDA architecture.

the model trained using optimum proportions of AF and non-AF data segments. Accordingly, in testing, we employed two CDA models for each class of AF and non-AF PPG segments classified based on an initial PPG classification.

The final classification is applied to the denoised PPG signals reconstructed by CDAs to predict the final AF/non-AF labels of the PPG segments. It is expected that the final predicted AF/non-AF labels are more accurate than the labels predicted by the initial classifier, since the final classification is based on the recovered or denoised PPG signals.

The proposed CDA architecture is shown in Fig. 4. The encoder part consists of a series of convolutional layers, where each individual convolutional layer is followed by a rectified linear units (ReLU) function and a batch normalization (BN) to prevent overfitting. The input and output of the network are images sized 256 × 106 × 2, where the dimension shows the width, height, and number of input channels, respectively. In the CNN layers, the convolutional filters of size [3 3] are applied to the input to create feature maps. A series of convolutional and transposed convolutional layers (also known as deconvolutional layers) formed the encoder part. Each convolutional and deconvolutional layer is followed by ReLU and BN. A stride of 2 was applied to every-other convolution layer to down-sample the input image in the encoder part.

PPG signals consist of a diverse range of time–frequency information, as they are a combination of heart activity, vascular relaxation processes, and the microcirculation system status (Liang, Chen, Ward, & Elgendi, 2018). As the PPG is a nonstationary signal in which instantaneous frequency alters with time, its characteristics cannot be fully described using only the frequency domain information. Among time–frequency analysis methods, STFT, which explores both the instantaneous frequency and instantaneous amplitude of the nonstationary time-dependent pulses (Huang, Chen, Yao, & He, 2019), is known to be a particularly effective method for time–frequency analysis of PPG data. For a discretized signal, the mathematical formulation shown in Equation (1),

$$STFT\{x[n]\} = X(m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn} \tag{2}$$

where $x[n]$ and $w[n]$ are the signal and window functions, respectively.

To estimate the time–frequency content of the PPG segments, we applied STFT using the Hanning window of length 64, which is defined as below:

$$w(n) = \begin{cases} 0.5\left[1 - cos\left(\frac{2\pi n}{M-1}\right)\right], & 0 \le n \le M-1 \\ 0, & otherwise \end{cases} \tag{3}$$

where $M = 64$.

For each PPG segment, we extracted magnitude and phase images, which contain the absolute and imaginary values of the transformed signal coefficients, respectively. The STFT images were resized into 256 × 106 pixels for each 25-sec PPG segment. We used the simple skip connection with a multiplication factor of 1.

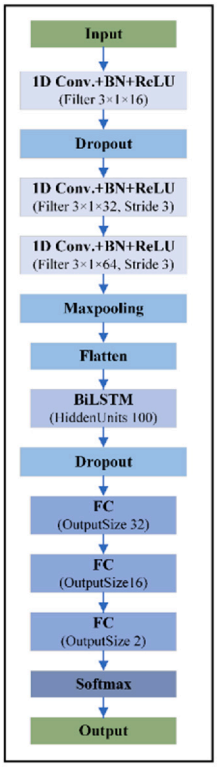

Fig. 5. Architecture of the designed AF/non-AF classifier.

2.7. AF/non-AF PPG data classification

The AF detection network is composed of convolutional and bidirectional Long Short-Term Memory (biLSTM) layers to extract morphological and temporal features. The PPG signal is fed as the input data into the three 1D convolutional layers with 3 × 1 convolutional kernels, followed by the BN and rectified linear layer as the activation function. The last convolutional layer is connected to a Long Short-Term Memory (LSTM) layer through a maxpooling layer to reduce sensitivity to the location of the features, by down sampling the feature maps. The LSTM layer consists of a bidirectional LSTM with 100 neurons connected to a softmax layer through a drop-out layer and three fully connected layers. The number of neurons in the fully connected layers is 32, 16, and 2, respectively. A drop out with p = 0.5 is used to enhance the generalization performance of the network. Finally, a sofmax function is applied to normalize the outputs and convert them from weighted sum values into probabilities interpreted as the probability of containing AF.

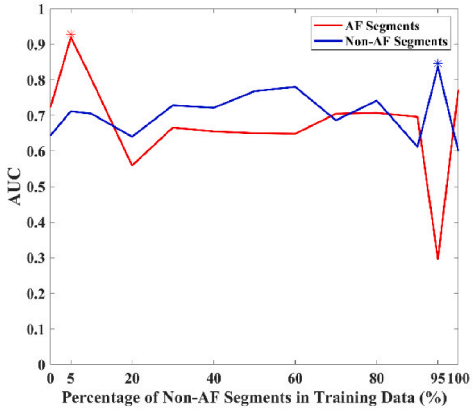

Fig. 6. AUC values of the classifiers obtained by AF/non-AF data segments, which were reconstructed by trained CDA using different percentage of non-AF segments in the training dataset. Red and blue asterisks show the highest AUC values obtained for AF and non-AF data segments, respectively.

Fig. 5 illustrates the architecture of the described DL-based classifier in this study.

2.8. Evaluation metrics

We evaluated the performance of AF/non-AF classification models using accuracy (Accu.), sensitivity (Sens.), specificity (Spec.), positive prediction value (PPV), negative prediction value (NPV), and F2-measure (F2_meas.). Among all performance measures for classifiers, the F-measure is a combined performance measure, which reflects the classifier performance on the minority class; it performs well on fairly balanced data (Akosa, 2017). The F_β-measure represents the weighted harmonic mean between PPV and sensitivity, however, it is sensitive to changes in data distributions.

We compared the classification performance of our proposed approach with the rule-based method (Method I) (Bashar et al., 2019) and DL-based method (Method II) (Torres-Soto & Ashley, 2020), which were developed in previous studies. Further, we calculated the increase of usable data coverage, by comparing the recovered PPG segments using our approach and the clean segments detected using two noise detection algorithms, called Noise-detection-Method I (Mohagheghian et al., 2022) and Noise-detection-Method II (Bashar et al., 2019).

3. Results

3.1. Optimization of the AF/non-AF data proportion used for CDA training

In this study, we applied a classifier on low-to-moderately corrupted PPG data to initially identify the AF and non-AF data classes. Then, the appropriately designated CDA models that were trained to reconstruct either the corrupted AF or non-AF PPGs (based on the initial classification determination) were applied to the data to improve the probability of accurate classifications in the final step. There is some likelihood of misclassification of AF and non-AF data (in initial classification) commensurate with the degree of motion artifacts. Hence, there is some chance that AF designation may largely be due to motion artifacts. Thus, to account for this scenario, the training dataset for CDA specific to AF data should mostly contain AF and a small proportion of non-AF segments. Similarly, the training dataset for CDA specific to non-AF data should mostly contain non-AF and a small proportion of AF segments. These conditions dictated that we explore the optimized CDAs trained based on the appropriate proportion of AF and non-AF data segments. To optimize the CDA models for reconstruction of AF and non-AF data groups, we trained the CDA model using varying percentage of AF and non-AF segments. Then, the CDA models were evaluated using the classification performance of the reconstructed data in each group.

Fig. 6 indicates the AUC (Area under the ROC Curve) values of the classifier applied to the reconstructed AF/non-AF data groups when CDA training dataset contains varying percentage of non-AF and AF data. According to Fig. 6, the highest AUC for the reconstructed AF data was achieved when the CDA model was trained dominantly by AF data segments (equaling 95% of training data) along with a small percentage of non-AF segments (equaling 5% of training data). Similarly, the highest AUC value for the reconstructed non-AF data group was obtained when the CDA was trained using a training dataset containing 95% non-AF and %5 AF data segments. Accordingly, the CDA is optimum for a data type when it is dominantly trained using the congruous data type. All testing results were based on these optimized CDA models.

3.2. Evaluation on test data sets

In this section, we summarize the performance of the AF detection methods (including our proposed approach, Method I (Bashar et al., 2019), and Method II (Torres-Soto & Ashley, 2020)) on Pulsewatch and DeepBeat testing datasets. Table 7 demonstrates the performance of classifiers on non-reconstructed and reconstructed Pulsewatch PPG segments. The results from our approach showed higher classification performance of the (denoised) reconstructed PPGs (Accu. = 91.98%, Sens. = 91.88 %, and Spec. = 92.01%) compared to the non-reconstructed signals (Accu. = 83%, Sens. = 90.07 %, and Spec. = 81.2%). Though PPV is sensitive to the imbalanced data (Tharwat, 2020), a noticeable improvement of PPV (an increase of about 20%, from 54.98% to 74.56%) was observed on the reconstructed PPG signals due to the significant reduction of false positive (FP) values after reconstruction. Accuracy and specificity on the reconstructed signals also increased compared to the non-reconstructed PPGs. Accuracy showed an increase from 83% to 92%, and specificity improved from 81% to 92% after signal reconstruction. To summarize the classification performance, we used the F2-measure as a measure of PPV, and sensitivity to reflect the improvement on reconstructed data compared to the non-reconstructed.

To examine the two other methods, both Method I and Method II resulted in a reduction of FP values on the reconstructed PPG segments compared to on the non-reconstructed signals. However, the false negative (FN) values are high on both reconstructed and non-reconstructed PPG segments, leading to very low sensitivity values for both methods. Further, both Method I and Method II scored low on the F2-measure scale. Figs. 7 and 8 display the bar plots for performance measures and heatmap confusion matrices of the classification methods

Table 7
Performance of the classification methods on Pulsewatch dataset.

		Number of Segments	TP	FP	TN	FN	Accu.	Sens.	Spec.	PPV	NPV	F2_meas.
Our approach	**Non-reconstructed Seg.**	21,713	3973	3253	14,049	438	83.00	90.07	81.20	54.98	96.98	79.88
	Reconstructed Seg.	21,713	4053	1383	15,919	358	**91.98**	**91.88**	92.01	74.56	**97.80**	**87.80**
Method 1	**Non-reconstructed Seg.**	21,713	1367	312	16,990	3044	84.54	30.99	98.20	81.42	84.81	35.37
	Reconstructed Seg.	21,713	955	256	17,046	3456	82.90	21.65	**98.52**	**78.86**	83.14	25.32
Method 2	**Non-reconstructed seg.**	21,713	821	6217	11,085	3590	54.83	18.61	64.07	11.67	75.54	16.63
	Reconstructed Seg.	21,713	531	538	16,764	3880	79.65	12.04	96.89	49.67	81.21	14.19

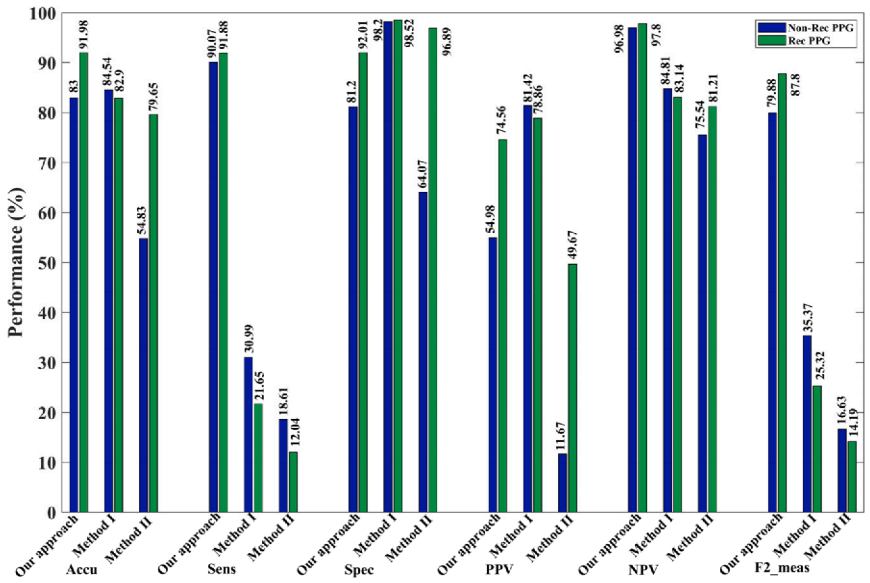

Fig. 7. Performance of the classification methods on Pulsewatch test dataset.

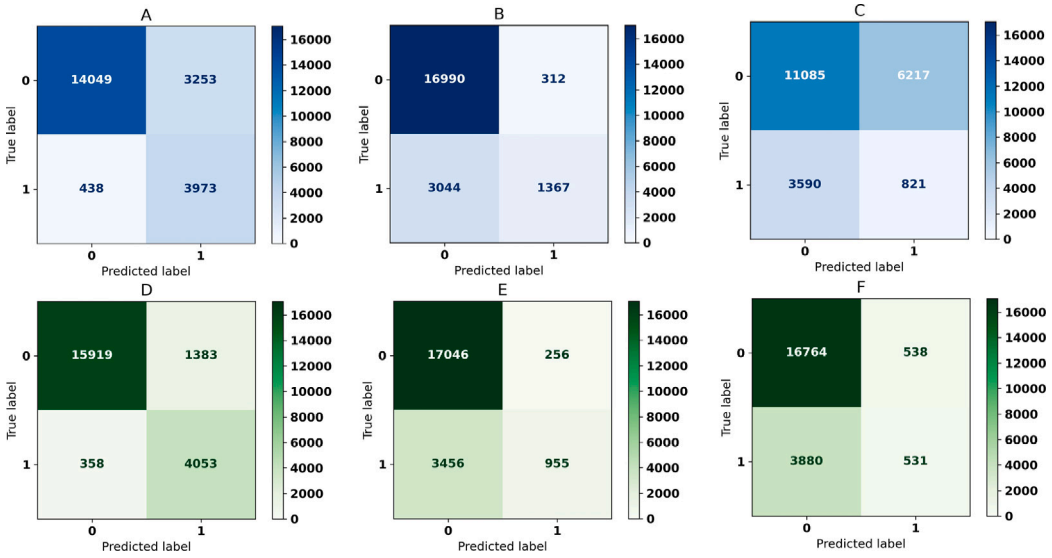

Fig. 8. Confusion matrices of the classification methods on Pulsewatch test dataset pre-CDA: A. Our approach, B. Method I, C. Method II; post-CDA: D. Our approach, E. Method I, F. Method II.

Table 8
Performance of the classification methods on DeepBeat dataset.

		Number of Segments	TP	FP	TN	FN	Accu.	Sens.	Spec.	PPV	NPV	F2_meas.
Our approach	**Non-reconstructed Seg.**	3827	1746	615	1249	217	78.26	88.95	67.01	73.95	85.20	85.48
	Reconstructed Seg.	3827	1782	371	1493	181	**85.58**	**90.78**	80.10	82.77	**89.19**	**89.06**
Method 1	**Non-reconstructed Seg.**	3827	452	1	1863	1511	60.49	23.03	**99.95**	**99.78**	55.22	27.21
	Reconstructed Seg.	3827	548	187	1677	1415	58.14	27.92	89.97	74.56	54.24	31.91
Method 2	**Non-reconstructed Seg.**	3827	696	1568	1235	328	50.46	67.97	44.06	30.74	79.01	54.72
	Reconstructed Seg.	3827	160	486	2317	864	64.72	15.63	82.66	24.77	72.84	16.87

compared in Table 7.

In addition to the Pulsewatch data, we used DeepBeat data as the independent testing dataset. Table 8 shows the classifiers' performance on the DeepBeat dataset. According to the table, our proposed approach achieved superior performance compared to other methods, as all metrics represent significant enhancement after signal reconstruction. It is notable that Method I achieved the highest specificity and PPV on non-reconstructed data, due to very low FP value. From Table 8, it is observed that the high FP value obtained from Method II on DeepBeat data correlates to low specificity and PPV for non-reconstructed segments, while a high FN value results in low sensitivity for reconstructed PPGs. Figs. 9 and 10 show the performance bar plots and heatmap confusion matrices of the classification methods compared in Table 8.

Table 9 shows the increase of data coverage (usable data) percentage

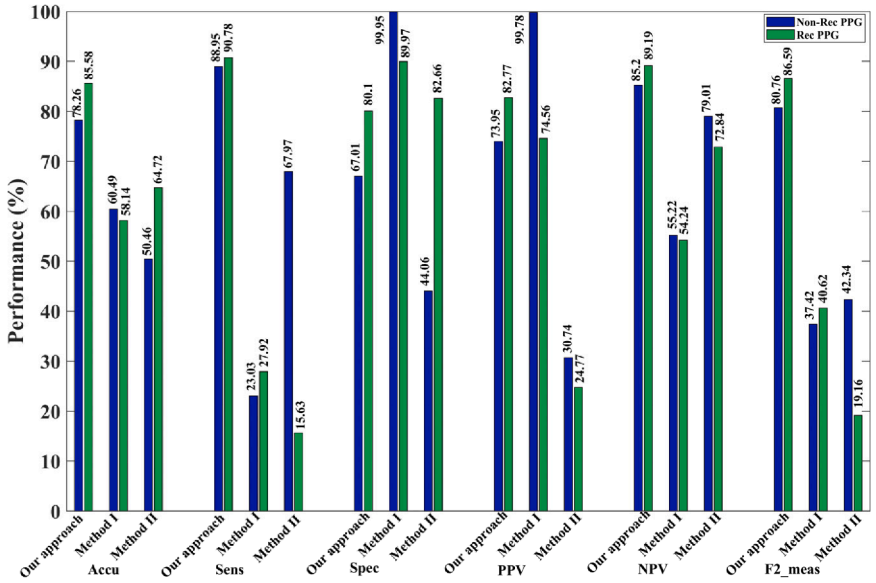

Fig. 9. Performance of the classification methods on DeepBeat test dataset.

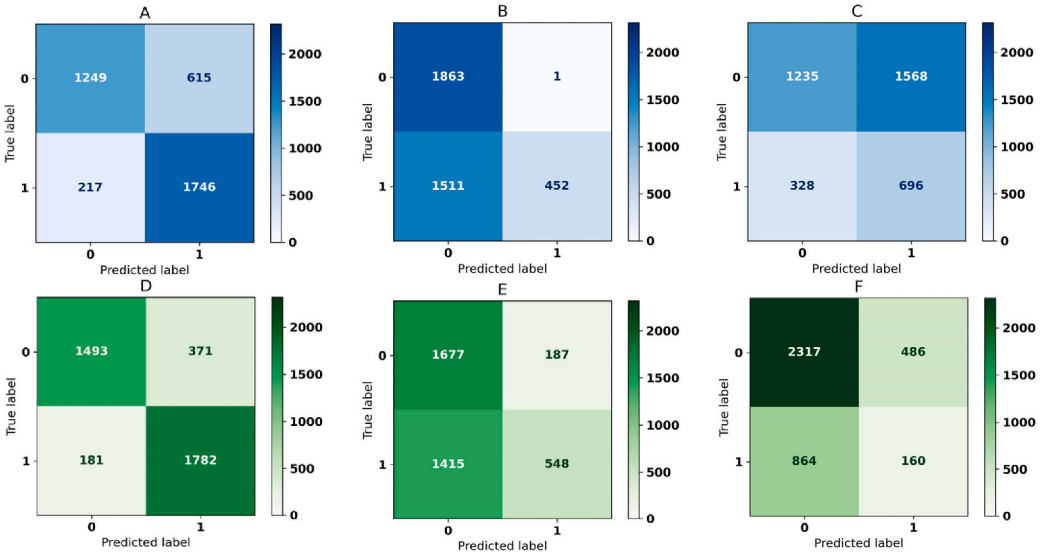

Fig. 10. Confusion matrices of the classification methods on DeepBeat test dataset pre- CDA: A. Our approach, B. Method I, C. Method II; post-CDA: D. Our approach, E. Method I, F. Method II.

Table 9
PPG data coverage obtained by our approach in comparison to the PPG noise detection algorithms.

	Number of Recovered PPG Seg.	Number of Clean PPG Seg.		Coverage Percentage Increase (%)	
	Our approach	**Noise-detection-Method I**	**Noise-detection-Method II**	**Compared to Noise-detection-Method I**	**Compared to Noise-detection-Method II**
AF Segments	4411	3437	3175	22.08	28.02
Non-AF Segments	17,302	14,738	13,830	14.81	20.07
Total	21,713	18,175	17,005	16.29	21.68

using our denoising approach evaluated by comparing the number of CDA-recovered segments with the clean segments detected using PPG noise detection algorithms. As shown in the table, our proposed approach achieved an increase of about 16% and 21% in total recoverable data (both AF and non-AF) coverage compared to the noise detection algorithms, Noise-detection-Method I (Han, Mohagheghian, & Chon, 2022b; Mohagheghian et al., 2022), and Noise-detection-Method II (Bashar et al., 2019), respectively. The usable data coverage represents more increase for AF segments (greater than 20%), which is especially important for long-term monitoring of cardiac patients.

4. Discussion

PPG signals collected by wearable devices during long-term monitoring can improve the detection of AF and allow for timely treatment as well as save the time and resources of clinicians (Shen et al., 2019). However, the signal quality of the PPG affects the accuracy of identification of AF segments, especially in real-life applications with long-duration monitoring. Previous studies have reported that a large proportion of the PPG data recorded by wearable devices is corrupted and cannot be used to detect cardiac arrhythmias, e.g., atrial fibrillation (Liang et al., 2018) (Selder et al., 2020) (Väliaho et al., 2021) (Bonomi et al., 2016).

In a prior study by Kwon et al. (Kwon et al., 2019), the recurrent neural network (RNN) and 1D convolutional neural network (1D-CNN) were used to detect either AF or NSR in the presence of premature atrial contractions (PAC) after successful cardioversion (CV). In the mentioned study, 75 patients who had had successful direct-current CV underwent ECG and PPG data recording for about 15 min. During the measurements, they were required to rest in the supine position so that motion artifacts were minimized. The authors stated that the accuracy of AF detection versus NSR was 99.32% versus 95.85% from 1D-CNN and 98.27% versus 96.04% from the RNN method.

In another study (Shashikumar, Shah, Li, Clifford, & Nemati, 2017), a deep neural network approach was proposed to detect AF episodes in PPG data. The authors applied a continuous wavelet transform to the PPG signal and then trained a CNN model on the extracted spectrograms to identify AF. The authors stated that the combination of using CNN and features calculated based on signal quality and beat-to-beat variability provided a high accuracy. Their data were collected from 98 patients (45 subjects with AF and 53 subjects with other rhythms) using wrist-based Simband watches. Their proposed approach showed a pooled AUC of 0.95 (Accuracy = 91.8%) for the cross-validation results.

In a previous study by Torres-Soto et al. (Torres-Soto & Ashley, 2020), the authors developed a multitask deep learning method to perform signal quality assessment and AF arrhythmia detection. They used a CNN-based AF detection model called DeepBeat to detect AF in data from wearable PPG devices (Simband watches). They collected the data from 132 patients with confirmed AF undergoing elective CV or elective stress tests during about 20 min before and 20 min after the CV procedure for the treatment of AF. Their algorithm performance showed sensitivity, specificity, and F1 score of 0.98, 0.99, and 0.93, respectively.

The authors in a prior study (Eerikäinen et al., 2019) used a Random Forest (RF) model, which combined features from PPG, ACC, and inter-pulse interval (IPI) data, to classify AF, atrial flutter (AFL), and other cardiac rhythms. PPG, ACC, and ECG data were collected from 40 patients using a data logging device during patient monitoring over 24 h as part of routine clinical care. The RF model classified AF/AFL/other rhythms with sensitivity of 97.6/84.5/98.1% and specificity of 98.2/99.7/92.8%, respectively.

This study provides the first comprehensive framework to accurately detect AF and non-AF rhythms from two large PPG datasets captured using smartwatches from different vendors and varying models. We developed our novel denoising and subsequent classification approach on smartwatch PPG signals continuously collected for two weeks from stroke survivors or patients with a history of transient ischemia. Further, we used a PPG dataset provided by Stanford University, containing millions of segments, for independent testing of our proposed approach. Our aims were to improve the data coverage, i.e., the amount of usable data, as well as to enhance AF classification performance of the PPG data.

In the application of wearable devices for arrhythmia detection, achieving good PPG data quality remains challenging, as a high-fidelity signal is required to achieve acceptable sensitivity and specificity (Dörr et al., 2019). In (Bashar et al., 2019), the authors have reported that among 650 subjects whose PPGs were recorded using a commercially available smartwatch (Samsung Gear Fit 2), 142 subjects (21.8%) were excluded from the study due to their insufficient signal quality. One study showed that 43 (24%) of 180 PPG recordings of wristband-derived data were not classifiable because of poor signal quality (Selder et al., 2020). Another study demonstrated that 69.5% of the wristband PPG data collected from 173 patients during the daytime were discarded due to poor quality of the data (Väliaho et al., 2021). Other researchers reported almost 37% of the collected PPG signals from wrist-wearable devices during a 24-hour monitoring period were unreliable in detecting AF, due to motion artifacts (Bonomi et al., 2016). Overall, these studies emphasized how significant portions of data were unusable for confirming the presence of AF due to poor quality of the PPG signal.

4.1. Performance of the proposed approach

To avoid excluding a high amount of PPG data, we employed the convolutional denoising autoencoder model to recover the PPG segments which were low-to-moderately corrupted by motion artifacts. We define low-to-moderately corrupted as those portions of PPG data that are intermittently corrupted with noise < 5 s in duration within the 25-sec segment length. To reconstruct the recoverable segments, we developed a novel optimized CDA model to reconstruct the data in each initially identified AF/non-AF group. Accordingly, to successfully recover the waveform characteristics and dynamics of both AF and non-AF data, we proposed to use two CDAs trained using specific AF and non-AF data proportions determined via a validation dataset so that optimum performance of the classification could be achieved.

According to the results, the proportion of highly corrupted data during two weeks of smartwatch data recordings was 54%. As shown in Table 1, the amount of corrupted AF data was slightly higher than for non-AF; discarded highly corrupted data proportions were 53% and 59% for non-AF and AF data segments, respectively. However, an increase in data recovery percentage of 21.68% was observed due to our signal reconstruction approach for total AF and non-AF data. The improvement was even higher (28.02%) for the AF data group. In addition to the enhancement of the PPG data recovery, we aimed to significantly improve the classification performance of the PPG signal after reconstruction. Our approach achieved high performance in all measures, particularly in PPV which showed a notable enhancement for the reconstructed PPG data in both Pulsewatch and DeepBeat datasets.

4.2. Comparison to the existing methods

In addition, we demonstrated that our approach substantially outperforms existing techniques to detect AF segments from low-to-moderately noise-contaminated PPG data. For existing methods, the major concern in AF detection is that corrupted PPG segments containing only NSR might be incorrectly detected as AF segments (Pereira et al., 2020). The most common and simplest way to deal with corrupted PPG signals is to discard them and use only the high-quality segments, a method used in (Torres-Soto & Ashley, 2020) for the ambulatory testing data. In (Torres-Soto & Ashley, 2020), only a very small portion of the total testing PPG data (15 subjects) was from ambulatory cohort: the data from 11 subjects collected over the course of one week had no confirmed AF episodes. Only 4 subjects had AF events (929 25-sec segments). The authors reported very high performance (sensitivity and specificity of 98% and 99%, respectively), however, their reported performance was based on only the high-quality data segments. If our method only used the MNA-free data, we would see performance metrics similar to (Torres-Soto & Ashley, 2020). Including low-to-moderate MNA-corrupted data segments and other confounding rhythms such as premature atrial and ventricular contractions, our AF detection resulted in lower performance metrics. The fact that AF detection performance metrics are greater than 91% and usable data coverage can be increased by 21% with our approach, suggests that long-term monitoring of AF for stroke patient is feasible using smartwatches.

As a direct comparison, the rule-based method using RMSSD-Sample

entropy (Method I) resulted in a lower FP value for both non-reconstructed and reconstructed PPGs, however, the high FN value was its main drawback for both Pulsewatch and DeepBeat datasets. It is apparent that the algorithm in Method I tended to reduce the FP value at the cost of higher FN values, which led to adverse consequences for the sensitivity and NPV for both non-reconstructed and reconstructed PPGs. The low sensitivity and NPV due to a high FN value indicates its unreliability in identifying true AF segments, even after signal reconstruction. The general limitation of the rule-based methods is their dependency on the prior adjusted threshold values for classifications, which make them unreliable for screening applications where the new data may not conform to the chosen threshold values of the model.

Further, fine-tuned deep learning architectures, which infer the rhythm classifications directly from the signal waveforms, may become inaccurate in AF identification commensurate with the amount of artifact due to movements and other types of noise sources. The high FP rate obtained from the DL-based Method II is due to the artifact-corrupted non-AF segments, thereby leading to low PPV and specificity of the non-reconstructed PPG data. Note that the FP value reduced after signal reconstruction via CDA, which confirms the inefficiency of Method II when it is confronted with motion and noise artifacts. In addition, the inability of Method II to correctly detect AF segments is especially seen with the Pulsewatch dataset, as it showed a high number of FN values. These findings illustrate the low robustness and generalizability of Method II, as generalizability of a model is assessed by its performance on datasets other than the ones used for training.

To demonstrate the efficacy of our technique to detect AF, we compared the performance of AF detection on reconstructed PPG data using CDA models versus some of the alternative denoising approaches. The AF detection results on data reconstructed through EMD, Wavelet, biLSTM, and CDA-based denoising methods are presented in Table A1 of the appendix. Our proposed reconstruction approach demonstrated superior Accu., Sens., NPV, and F-meas. while maintaining comparable Spec. and PPV when compared to the other methods. Comprehensive details of this comparison can be found in (Mohagheghian et al., 2023, In Press), where we extensively evaluate our proposed approach to alternative reconstruction methods.

In binary classification problems, the minority class is usually the class of interest, which needs to be accurately predicted (Maratea, Petrosino, & Manzo, 2014). As our testing dataset used in this study has an imbalanced distribution, albeit not severely, we evaluated the performance of the classification by the F2-measure, which accentuates the FN values. According to F2-measures from all approaches, our proposed approach showed the highest improvement, which subsequently resulted in higher classification performance on the AF data group.

In addition to motion and noise corruption, which occur more in long-duration PPG recordings, the presence of non-AF rhythms other than NSR poses a challenge for existing AF detection algorithms (Pereira et al., 2020). The Pulsewatch dataset used in this study contained three types of cardiac arrhythmias—AF, PAC, and PVC—with substantial variability of these rhythms among individuals.

As a limitation in this study, the ACC-based noise detection approach could not be applied to the Stanford University ("DeepBeat") database, since it does not contain ACC recordings. Thus, the low-to-moderate corrupted data segments were manually annotated and they were used as the testing data.

Patients with persistent AF are likely to have different PPG characteristics compared to new-onset AF, due to different ventricular rates and cardiac outputs. The HR tends to be lower in the persistent AF group, as the arrhythmia has sustained longer or due to medications such as beta-blockers. The current results are based on both types of AF patients, paroxysmal and persistent AF. Our work as well as others have shown that persistent AF can also be accurately detected using PPG.

PPG signal analysis using wearable devices can provide continuous monitoring of stroke patients to enhance the probability of detecting paroxysmal AF and better assessment of AF burden, whereas intermittent ECG evaluation during clinical visits has a low likelihood of paroxysmal AF detection. Moreover, developing PPG-based AF detection algorithms is an effective strategy to reduce the morbidity and mortality rate of stroke patients, as very long-term ambulatory monitoring with PPG devices is more feasible than with ECG devices. However, the current PPG technology itself is not sufficient to diagnose AF according to the current AF guidelines, hence, abnormal rhythms detected by PPG-based smart devices need to be validated using an ECG monitor. As the research community develops more advanced AI-based algorithms and further tests them using more clinical data, patients and clinicians will both be more confident to accept the validity of PPG-based sensors in the near future.

5. Conclusion

The main contributions of this investigation can be summarized as follows. First, we achieved an optimized framework based on smartwatch PPG signals to accurately diagnose AF vs. non-AF cardiac rhythms with a high sensitivity, which is the prerequisite of a monitoring system, and high specificity, endorsing the feasibility of our approach for long-term screening. It is particularly evident when our test dataset included real-life data accounting for high heart rates due to various physical activities. Second, we significantly increased the PPG data coverage (usable data for AF screening) by reconstructing PPG data that were defined to be low-to-moderately noise corrupted, which consequently improved the AF detection performance in the presence of other arrhythmias like PAC and PVC.

In addition, the results illustrate the robustness and generalizability of our approach to identify AF/non-AF rhythms from different smartwatches and the capability of our algorithm to differentiate other non-AF rhythms including PAC and PVC, as well as NSR. Our approach allows not only AF detection but more accurate representation of AF burden, which is clinically important for subsequent oral anticoagulation therapy. Our approach to reduce motion-corrupted segments can lead to better assessment of AF burden.

CRediT authorship contribution statement

F.M.: Formal analysis, Investigation, Methodology, Validation, Writing-original draft, F.M., D.H., and A.P.: Software, Writing – review & editing, D.D.M. and K.H.C.: Conceptualization and Project administration, K.H.C.: Investigation, Supervision, Writing-review & editing, O. G., D.C, A.P., N.N.: Data curation, E.Y.D., E.M.O., K.N., A.H., E.L.D., D. D., K.V.T.: Data curation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial

Table A1
Performance comparison of AF detection algorithm on data reconstructed using EMD, Wavelet, biLSTM, and CDA denoising methods.

Denoising Method	Number of Segments	TP	FP	TN	FN	Accu.	Sens.	Spec.	PPV	F_meas.	NPV
EMD	**1,200**	379	160	489	172	0.72	0.69	0.75	0.70	0.69	0.74
Wavelet	**1,200**	432	122	527	119	0.80	0.78	**0.81**	**0.78**	0.78	0.82
biLSTM	**1,200**	456	173	476	95	0.78	0.83	0.73	0.72	0.80	0.83
CDA	**1,200**	483	151	498	68	**0.82**	**0.88**	0.77	0.76	**0.85**	**0.88**

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgement

This work was supported by grant NIH R01 HL137734.

Appendix A

.

References

Akosa, J., 2017. Predictive accuracy: A misleading performance measure for highly imbalanced data. Presented at the Proceedings of the SAS global forum, pp. 1–4.

Bashar, S. K., Han, D., Hajeb-Mohammadalipour, S., Ding, E., Whitcomb, C., McManus, D. D., & Chon, K. H. (2019). Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches. *Scientific Reports, 9*, 1–10.

Bonomi, A.G., Schipper, F., Eerikäinen, L.M., Margarito, J., Aarts, R.M., Babaeizadeh, S., de Morree, H.M., Dekker, L., 2016. Atrial fibrillation detection using photoplethysmography and acceleration data at the wrist. Presented at the 2016 computing in cardiology conference (cinc), IEEE, pp. 277–280.

Colilla, S., Crow, A., Petkun, W., Singer, D. E., Simon, T., & Liu, X. (2013). Estimates of current and future incidence and prevalence of atrial fibrillation in the US adult population. *The American journal of cardiology, 112*, 1142–1147.

Dickson, E. L., Ding, E. Y., Saczynski, J. S., Han, D., Moonis, M., Fitzgibbons, T. P., … McManus, D. D. (2021). Smartwatch monitoring for atrial fibrillation after stroke—The Pulsewatch Study: Protocol for a multiphase randomized controlled trial. *Cardiovascular Digital Health Journal, 2*, 231–241.

Ding, E. Y., Han, D., Dickson, E. L., DiMezza, D., Scott, J., Mohagheghian, F., … Fitzgibbons, T. P. (2021). Use of a smartwatch and app designed by stroke survivors for atrial fibrillation detection in older adults after stroke/transient ischemic event: Preliminary findings from an ongoing randomized clinical trial. *Circulation, 144*, A9886–A.

Dörr, M., Nohturfft, V., Brasier, N., Bosshard, E., Djurdjevic, A., Gross, S., … Eckstein, J. (2019). The WATCH AF trial: SmartWATCHes for detection of atrial fibrillation. *JACC: Clinical Electrophysiology, 5*, 199–208.

Eerikäinen, L. M., Bonomi, A. G., Schipper, F., Dekker, L. R., de Morree, H. M., Vullings, R., & Aarts, R. M. (2019). Detecting atrial fibrillation and atrial flutter in daily life using photoplethysmography data. *IEEE Journal of Biomedical and Health Informatics, 24*, 1610–1618.

Georgieva-Tsaneva, G., Gospodinova, E., & Cheshmedzhiev, K. (2022). Cardiodiagnostics based on photoplethysmographic signals. *Diagnostics, 12*, 412.

Han, D., Bashar, S.K., Lazaro, J., Ding, E., Whitcomb, C., McManus, D.D., Chon, K.H., 2019. Smartwatch PPG peak detection method for sinus rhythm and cardiac arrhythmia. Presented at the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, pp. 4310–4313.

Han, D., Bashar, S. K., Lázaro, J., Mohagheghian, F., Peitzsch, A., Nishita, N., … Scott, J. (2022a). A real-time PPG peak detection method for accurate determination of heart rate during sinus rhythm and cardiac arrhythmia. *Biosensors, 12*, 82.

Han, D., Bashar, S. K., Mohagheghian, F., Ding, E., Whitcomb, C., McManus, D. D., & Chon, K. H. (2020). Premature atrial and ventricular contraction detection using photoplethysmographic data from a smartwatch. *Sensors, 20*, 5683.

Han, D., Ding, E. Y., Cho, C., Jung, H., Dickson, E. L., Mohagheghian, F., … McManus, D. D. (2023). A smartwatch system for continuous monitoring of atrial fibrillation in older adults after stroke or transient ischemic attack: Application design study. *JMIR cardio, 7*, e41691.

Han, D., Mohagheghian, F., Chon, K.H., 2022b. Recent advances involving hardware and algorithmic approaches to combat motion artifacts in photoplethysmographic data.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, J., Chen, B., Yao, B., & He, W. (2019). ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. *IEEE Access, 7*, 92871–92880.

Kwon, S., Hong, J., Choi, E.-K., Lee, E., Hostallero, D. E., Kang, W. J., … Oh, S. (2019). Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: Algorithms development study. *JMIR mHealth and uHealth, 7*, e12770.

Liang, Y., Chen, Z., Ward, R., & Elgendi, M. (2018). Photoplethysmography and deep learning: Enhancing hypertension risk stratification. *Biosensors, 8*, 101.

Lippi, G., Sanchis-Gomar, F., & Cervellin, G. (2021). Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge. *International Journal of Stroke, 16*, 217–221.

Mao, X., Shen, C., Yang, Y.-B., 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. Advances in neural information processing systems 29.

Maratea, A., Petrosino, A., & Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences, 257*, 331–341.

Mohagheghian, F., Han, D., Ghetia, O., Peitzsch, A., Nishita, N., Pirayesh Shirazi Nejad, M., … Chon, K. H. (2023). *Noise Reduction in Photoplethysmography Signals using a Convolutional Denoising Autoencoder with Unconventional Training Scheme*. IEEE Transactions on Biomedical Engineering: In Press.

Mohagheghian, F., Han, D., Peitzsch, A., Nishita, N., Ding, E., Dickson, E., … Scott, J. (2022). Optimized signal quality assessment for photoplethysmogram signals using feature selection. *IEEE Transactions on Biomedical Engineering*.

Pereira, T., Tran, N., Gadhoumi, K., Pelter, M. M., Do, D. H., Lee, R. J., … Hu, X. (2020). Photoplethysmography based atrial fibrillation detection: A review. *NPJ Digital Medicine, 3*, 1–12.

Rosman, L., Gehi, A., & Lampert, R. (2020). When smartwatches contribute to health anxiety in patients with atrial fibrillation. *Cardiovascular Digital Health Journal, 1*, 9.

Seet, R. C., Friedman, P. A., & Rabinstein, A. A. (2011). Prolonged rhythm monitoring for the detection of occult paroxysmal atrial fibrillation in ischemic stroke of unknown cause. *Circulation, 124*, 477–486.

Selder, J., Proesmans, T., Breukel, L., Dur, O., Gielen, W., van Rossum, A. C., & Allaart, C. (2020). Assessment of a standalone photoplethysmography (PPG) algorithm for detection of atrial fibrillation on wristband-derived data. *Computer Methods and Programs in Biomedicine, 197*, Article 105753.

Shashikumar, S.P., Shah, A.J., Li, Q., Clifford, G.D., Nemati, S., 2017. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. Presented at the 2017 IEEE EMBS international conference on biomedical & health informatics (BHI), IEEE, pp. 141–144.

Shen, Y., Voisin, M., Aliamiri, A., Avati, A., Hannun, A., Ng, A., 2019. Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning. Presented at the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1909–1916.

Suk, H.-I., Lee, S.-W., & Shen, D. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure and Function, 220*, 841–859.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

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

Tran, K.-V., Filippaios, A., Noorishirazi, K., Ding, E., Han, D., Mohagheghian, F., Dai, Q., Mehawej, J., Wang, Z., Lessard, D., 2022. False Atrial Fibrillation Alerts from Smartwatches are Associated with Decreased Perceived Physical Well-being and Confidence in Chronic Symptoms Management.

Väliaho, E.-S., Lipponen, J. A., Kuoppa, P., Martikainen, T. J., Jäntti, H., Rissanen, T. T., … Laitinen, T. M. (2021). Continuous 24-h photoplethysmogram monitoring enables detection of atrial fibrillation. *Frontiers in Physiology, 12*.