PAPER • **OPEN ACCESS**

A supervised machine learning semantic segmentation approach for detecting artifacts in plethysmography signals from wearables

To cite this article: Zhicheng Guo *et al* 2021 *Physiol. Meas.* **42** 125003

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
6 August 2021

REVISED
9 November 2021

ACCEPTED FOR PUBLICATION
18 November 2021

PUBLISHED
29 December 2021

PAPER

A supervised machine learning semantic segmentation approach for detecting artifacts in plethysmography signals from wearables

Zhicheng Guo[1], **Cheng Ding**[2], **Xiao Hu**[2,3] and **Cynthia Rudin**[1,2]

[1] Department of Computer Science, Duke University, United States of America
[2] Department of Electrical and Computer Engineering, Duke University, United States of America
[3] Division of Health Analytics, School of Nursing, Biomedical Engineering, Pratt School of Engineering, Departments of Neurology, Biostatistics & Bioinformatics, Surgery, School of Medicine, Duke University, United States of America

E-mail: cynthia@cs.duke.edu

Keywords: plethysmography, PPG, wearables, signal artifacts

Abstract

Objective. Wearable devices equipped with plethysmography (PPG) sensors provided a low-cost, long-term solution to early diagnosis and continuous screening of heart conditions. However PPG signals collected from such devices often suffer from corruption caused by artifacts. The objective of this study is to develop an effective supervised algorithm to locate the regions of artifacts within PPG signals. *Approach*. We treat artifact detection as a 1D segmentation problem. We solve it via a novel combination of an active-contour-based loss and an adapted U-Net architecture. The proposed algorithm was trained on the PPG DaLiA training set, and further evaluated on the PPG DaLiA testing set, WESAD dataset and TROIKA dataset. *Main results*. We evaluated with the DICE score, a well-established metric for segmentation accuracy evaluation in the field of computer vision. The proposed method outperforms baseline methods on all three datasets by a large margin (≈7 percentage points above the next best method). On the PPG DaLiA testing set, WESAD dataset and TROIKA dataset, the proposed method achieved 0.8734 ± 0.0018, 0.9114 ± 0.0033 and 0.8050 ± 0.0116 respectively. The next best method only achieved 0.8068 ± 0.0014, 0.8446 ± 0.0013 and 0.7247 ± 0.0050. *Significance*. The proposed method is able to pinpoint exact locations of artifacts with high precision; in the past, we had only a binary classification of whether a PPG signal has good or poor quality. This more nuanced information will be critical to further inform the design of algorithms to detect cardiac arrhythmia.

1. Introduction

Wearable health monitoring devices equipped with plethysmography (PPG) sensors were shown to have strong potential in improving the cardiovascular disease monitoring (McConnell *et al* 2018, Ioannidis *et al* 2019, Raja *et al* 2019). PPG-enabled devices contain an optical sensor that detects blood volume changes through the skin (Castaneda *et al* 2018). Beams of light are emitted from the sensor, and changes in light absorption by the blood are recorded as PPG signals. The sensor is often embedded on the back of wrist-worn smart devices. These devices are non-invasive (because they are optical), cost-effective, easy to use, and can collect signals over long periods of time as their wearers go about their daily activities. For these reasons, PPG has become a standard feature in up to 71% of consumer smart wearables (Henriksen *et al* 2018). PPG monitoring can provide detailed physiological measurements of the user (blood oxygen saturation, blood pressure, heart rate, respiration, etc.), and can enable early detection of atrial fibrillation, hypertension, vascular aging, chronic kidney disease, atherosclerosis, and other serious conditions that otherwise might go undetected (Allen *et al* 2006, Liang *et al* 2018, Saritas *et al* 2019, Pereira *et al* 2020, Ouyang *et al* 2020, Dall'Olio *et al* 2020).

One problem with wearable devices is that they are influenced by the wearer's motion and environmental 'noise' (e.g. ambient light, sweat, pressure applied to the sensor Sañudo *et al* 2019) that impacts the signal and could lead to false positive rates that are unacceptably high for identifying heart conditions. Thus, the detection

of artifacts is a requirement for PPG monitoring. Accurate detection and localization of artifacts could also help preserve as much useful PPG signal as possible for detection of heart conditions.

We formulate the problem as a 1D supervised segmentation task, where we aim to segment artifacts from non-artifacts. To generate a dataset for training the algorithm, we first built a software annotation tool, which makes it easy for humans to identify and record the artifacts. Equipped with our new segmentation data that we developed using the annotation tool, we trained a novel deep neural network whose loss is a combination of segmentation accuracy—measured using a loss function called the active contour loss—and a smoothness term that encourages the model to generate fewer transitions between artifacts and non-artifacts in its segmentation predictions. Our deep neural network architecture builds on that of U-Net, which has been known to yield high-quality segmentation for images (Ronneberger *et al* 2015). Our method modifies U-Net in that it operates on 1D signals and places residual structures inside encoder and decoder blocks. Its loss function allows it to segment PPG signals into artifact and non-artifact segments.

We conducted extensive comparative experiments of this approach and several baseline approaches on multiple datasets. Importantly, we trained the model on one dataset, and tested it on three other datasets from other sources where subjects were performing many different activities and were in a variety of emotional and physical states. On all three test datasets, the results from our approach were substantially better than those of the state-of-the-art baselines we compared with. These results indicate that our approach, which is the first to use supervised segmentation, could be a promising development of using wearable technologies in widespread early detection of heart conditions.

2. Related work and relevance to our work

2.1. Artifact reduction

Previous studies have focused on reducing artifacts in PPG signals. Naraharisetti and Bawa (2011), Lee *et al* (2004), Kim *et al* (2007), Schack *et al* (2015), Chong *et al* (2014), Ram *et al* (2012) denoise signals by removing certain high frequencies from the signal with a focus on preserving heart rate information. (More generally, there is a subfield that aims to 'reconstruct' signals to improve estimation of heart rate.) Even though these methods preserve heart rate information, they potentially cause distortion and loss of morphological information, which changes the timing of pulse features and limits our ability to detect abnormal heart conditions. Thus, rather than *denoising* the signal in this work, we aim to *detect artifacts* and preserve as much of the original signal as possible.

2.2. Leverage additional/auxiliary information

Some approaches also utilize multiple channels and additional sensors such as accelerometers (Foo and Wilson 2006, Lee *et al* 2010, Bashar *et al* 2019, Zhang *et al* 2019). These methods are promising, but additional hardware is not always available. The dependence on additional hardware reduces the methods' compatibility and renders them less suitable for wide deployment under normal daily conditions.

2.3. Artifact detection with sliding windows and/or handcrafted features

Typical approaches for artifact detection suffers from one or both of the following disadvantages: (1) they rely on sliding windows rather than direct localization to detect artifacts, (2) they often rely on fixed features (that are not learned, but only computed). Let us go into more detail.

2.4. Sliding windows

In order to detect the precise location of artifacts, one would typically define a window/sub-sequence length and evaluate whether an artifact appears in that sub-sequence. Doing this has the unfortunate side effect of limiting the resolution at which artifacts can be detected because of computational reasons. If we would like to use a sliding window that starts at each time-step, it would be enormously computationally expensive, since the model needs to evaluate many sub-sequences with a large amount of overlap. Sliding- window-based methods also use only local features within a window, they do not take into account global features of the whole signal, or even features that extend beyond the edge of the window.

2.5. Handcrafted features

Most methods use features that are pre-computed. These methods implicitly make a strong assumption that clean signals have a set of non-person-specific criteria that differ from those of signals with artifacts. However, these criteria actually can vary dramatically between subjects and with different subject activities, which means any statistics that are calculated from pre-defined 'clean' signals will generally not be reliable. These statistical features would require constant adjustments for different subjects in order to be effective. In addition, some of

these non-learned features rely on peak detection (e.g. template creation, pulse segmentation, peak to peak feature calculation, random distortion testing). However, due to the nature of PPG, reliable and accurate detection of peaks is challenging thus introducing additional inaccuracy.

Handcrafted features include statistics such as entropy, signal skewness, Kurtosis, peak and valley magnitudes, peak-to-peak time interval, and slope ratios (Krishnan *et al* 2008, 2008, Selvaraj *et al* 2011, Tabei *et al* 2018, Vandecasteele *et al* 2018, Athaya and Choi 2020). A number of studies have used waveform-derived features such as heart rate, amplitude, waveform morphology, or spectral features, for sub-sequence artifact detection (Chong *et al* 2014, Cherif *et al* 2016, Dao *et al* 2017, Fischer *et al* 2017, Papini *et al* 2017, Lim *et al* 2018). These features are often used for classification models on sliding windows.

Leverage SQI: The signal quality index (SQI) of PPG signals can be used as a feature for artifact detection. SQI reflects the amount of corruption within a signal without pinpointing the location of the artifacts. Many studies have developed methods for SQI estimation based on the above non-learned features, including those of, Sukor *et al* (2011), Li *et al* (2012), Karlen *et al* (2012), Li and Clifford (2012). By estimating the SQI of each sliding window, one can locate the artifacts. However, this approach suffers from the above-mentioned disadvantages of sliding-window-based and non-learned feature-based methods.

2.6. Leverage learned features

Related work on learned features: Several studies have used learned features with a sliding window classification setup (Pereira *et al* 2019, Liu *et al* 2020, Goh *et al* 2020). In these works, classification models were trained on signal sub-sequences or sub-sequence encodings (e.g. 2D Gramain Angular Field), and the algorithms were able to locate the sub-sequences containing artifacts. Even using learned features, these methods still have shortcomings stemming from the use of sliding windows.

Besides the typical approaches mentioned above, one could resort to using a black box classification model on the PPG signal (clean or artifact binary classification), combined with a post-hoc 'explanation' analysis that uses saliency maps to identify possible artifacts. The goal of saliency map explanation techniques is to determine where the model was focusing its attention when it classified a signal as an anomaly. However, as we will show in this paper, this type of approach typically yields poor results.

2.7. Our work

How our work relates to past work: Our work differs from previous work in several ways and successfully avoids the previously discussed disadvantages. First, we treat the problem as a *segmentation* problem, aiming to segment out artifacts in the signal. This avoids the problems inherent to sliding windows. Second, we use learned features with a deep neural architecture that has not previously been used for PPG signal analysis. This neural network approach eliminates the need to rely on parameter adjustments made for different users, and it does not require peak detection or other pre-defined features that may not hold across subjects. Third, our study focuses on producing an algorithm that is widely deployable, reliable and usable for daily life, rather than inpatient or laboratory settings, where signals are much cleaner. To pursue this, we used datasets recorded from wrist sensors in ambulatory conditions. Our subjects have widely-varying activities throughout the recording.

Our deep neural architecture is an extension of the U-Net architecture. U-Net was first proposed by Ronneberger *et al* (2015) for 2D image segmentation featuring 'skip channels' that pass information from encoders to decoders. We will discuss our architecture in more detail in section 4.2.

For evaluation of results, we use the DICE score to produce a more direct and realist measure of the algorithms' performance. DICE evaluates the similarity between human annotations and model annotations of the artifact regions in the context of actual complete signals. The DICE score produces information for each time-step of the signal, instead of measuring the classification accuracy of each window.

3. Datasets

Three datasets were used in this study. The dataset used for training is the PPG-DaLiA dataset (Reiss *et al* 2019), which contains multimodal signals of 15 subjects performing a various real-life activities. Data including ECG signals (chest recorded, 700 Hz), three-axis acceleration (chest recorded, 700 Hz), PPG signals (wrist recorded 64 Hz), electrodermal activities record (4 Hz) and subject information (age, gender, height weight, skin color, fitness level—i.e. how often the subject participates in sports) were used in our study. This dataset was selected for training because its data collection setting is the most comprehensive and representative of daily life conditions among existing PPG datasets.

We used two independent datasets for evaluation: the WESAD dataset (Schmidt *et al* 2018) and TROIKA dataset (Zhang 2015).

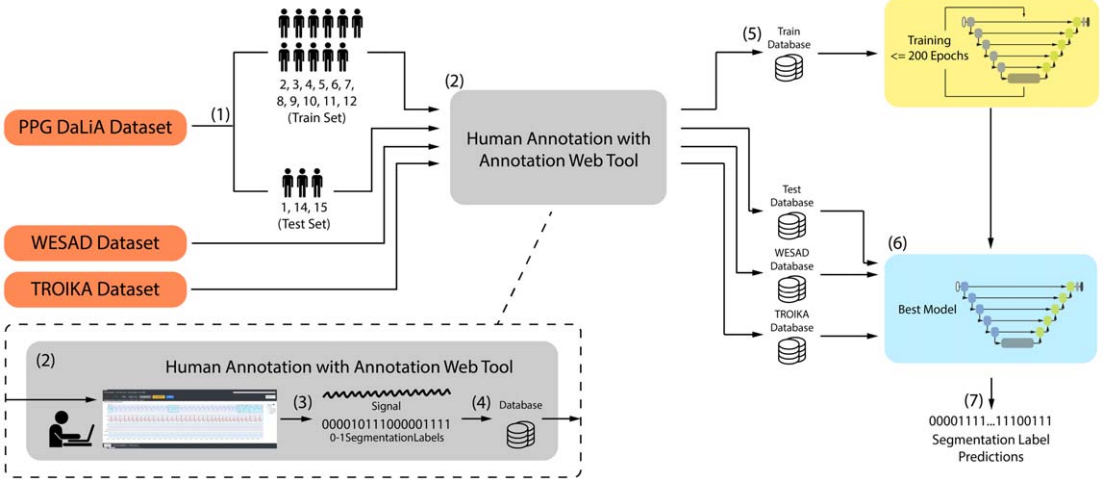

Figure 1. The overall pipeline of our work. Step (1): 30 s signal segments from PPG DaLiA were split into train and test set by randomly chosen subject IDs. Step (2): datasets were uploaded into the web annotation tool for human annotation. Steps (3) and (4): the tool transcribes human annotations into binary segmentation label masks (ground truth) and 30 s signal segments were stored into the back-end database. Step (5): ground truth and 30 s signal segments were then used to train the model. Step (6): predictions by the trained model are made on the PPG DaliA test set, WESAD and TROIKA datasets. Step (7): the trained model's predictions are provided for evaluation against human-annotated ground truth.

The WESAD dataset was recorded from both wrist- and chest-worn devices, from 15 subjects (age ranging from 21 to 55 years old, median 28 years old) during a lab study under different emotional states including neutral, stress, and amusement. Subjects were allowed to move freely while performing tasks. Data including ECG signals (chest recorded, 700 Hz), three-axis acceleration (chest recorded, 700 Hz), PPG signals (wrist recorded 64 Hz), electrodermal activities record (4 Hz) and subject information were used for our study. The PPG signals in WESAD dataset are also recorded from the wrist. This dataset is a good representation of PPG signals under a relatively small amount of movement, so it is ideal for testing the model's generalization ability to handle specific settings.

The TROIKA data was recorded from subjects with ages between 18 and 35. During data recording, each subject ran on a treadmill with changing speeds. The PPG signal from channel one (wrist recorded, 125 Hz), three-axis acceleration signals (wrist recorded, 125 Hz), and ECG signals (chest recorded, 125 Hz) were used. This dataset represents PPG signals affected by frequent and large movements, thus it was chosen for evaluating performance of the model under high motion intensity and extremely poor signal quality conditions.

In both the WESAD and the PPG DaLia dataset, the chest-worn ECG recording device is RespiBAN and the wrist-worn PPG recording device is Empatica E4. The TROIKA dataset used a bespoke device to collect the data.

For more details regarding subject information and activities in the above three datasets, please see appendix.

3.1. Pre-processing

During pre-processing of the PPG-DaLiA dataset, signals were sliced into 4305 30 s non-overlapping segments and normalized to range [0, 1]. The 30 s window size was chosen based on our downstream AF detection task which uses a 30 s window. Thirty seconds is also the time length required to identify an AF episode by accepted convention (Kirchhof *et al* 2016). Subjects had IDs from 1 to 15. Subjects were randomly selected for training and testing to avoid leakage of information from training into test. 3436 segments from 12 subjects (ID 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13) were reserved for training; 869 segments from the remaining subjects were reserved for testing as illustrated as Step (1) of figure 1.

A bandpass filter with a low end cutoff of 0.9 Hz and a high end cutoff of 5 Hz was applied to the segments of both PPG-DaLiA, WESAD dataset. This bandpass filter setup was also chosen based on our existing AF detection algorithm. The TROIKA dataset was pre-processed by its original author with bandpass from 0.4 Hz to 5 Hz (Zhang 2015). Signals from TROIKA and WESAD dataset were sliced into 30 s non-overlapping segments and generated 113, 2886, 2683 segments respectively. We down-sampled TROIKA dataset signals to 64 Hz to comply with the resolution of the training data. In addition, all the signals are converted into [0,1] by min–max normalization.

3.2. Annotation

One challenge of analyzing PPG data is the lack of publicly available fine-grained labeled data. In particular, *prior to this work, there did not exist a public dataset where artifacts are labeled in each PPG signal*. We created a dataset

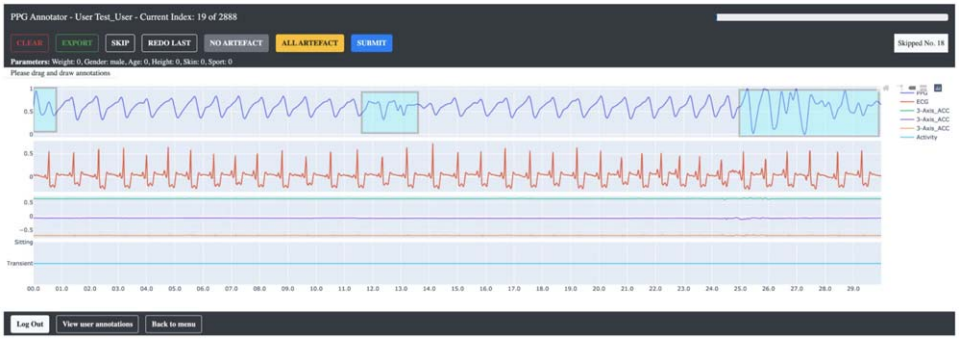

Figure 2. Screenshot of our annotation web tool with example user annotations.

that we made publicly available in the supplement of this manuscript. To create this dataset, we built a specially-designed web-based tool for annotation of PPG data. The tool allows the user to precisely select segments of the signal and mark them as artifacts as shown in Step (2) to Step (4) of figure 1. This tool automatically transcribes users' annotations into binary segmentation label masks. Figure 2 shows a screenshot of the annotation tool's annotation interface. The four rows are PPG signal, ECG signal, three-axis acceleration signal and recorded subject activity. All data in the four rows are time-synchronized, and users can zoom in on the signals via mouse scroll. Miscellaneous subject information including subject's weight, height, age, skin color and fitness level is also displayed. In cases when such information is missing, a placeholder of 0 will be displayed. For efficiency purposes, an annotator can mark the whole signal as 'No Artifact' or 'All Artifact' with the click of a button. After making selections on each PPG signal, the user would click 'Submit' to record the selections. Visualization of annotations are also built into this web-tool, so users could inspect their annotations and make adjustments accordingly.

We next describe how the annotations were transformed into labels for machine learning. The annotation tool assigns each signal a 1D segmentation label mask with the same dimension as the signal itself. Each data point in the original signal will be assigned a binary label (0 for clean and 1 for artifact) by an annotator. To create these binary labels, the annotator referenced the three-axis acceleration signal, observed the correlation between ECG heart beat and PPG heart beats and the regularity of the PPG signals to determine whether there was an artifact in the PPG signal. If there is no artifact present, the annotator will mark the signal as 'No Artifact'. The following are two scenarios we consider for artifact annotations:

(1) If the accelerometer shows motion and PPG signal shows irregularities that correspond with the accelerometer data, the signal segment will be marked as an artifact.

(2) If the accelerometer shows no obvious motion, and ECG shows normal sinus rhythm; but the PPG shows irregularities, the segment will be marked as an artifact.

Here we give three examples of the annotation process. As shown in figure 3(a), and the corresponding label mask will be assigned a vector of 0's. In this case, the beats in the ECG and PPG signals match and the waveform has a regular pattern, and there were no sudden changes in the three-axis acceleration signal; that is, the combination of PPG, ECG and acceleration signals indicated that there are no artifacts. In contrast, figure 3(b) shows a clear conflict between the ECG signal, acceleration signal and the PPG signal. There is no movement and the ECG is regular, so the irregularity in the PPG signal would be labeled by the annotator as an artifact. Figure 3(c) is an example of 'All Artifact.' Here, according to the acceleration signal, the sensors experience frequent large movements. The ECG shows the impact of the motions and the PPG signal has many artifacts.

Each signal was annotated by at least one annotator. During the early annotation trial phase, 50 30 s signals were randomly selected and annotated by three annotators independently. We used the intersection over union (IoU) metric to measure the inter annotator agreement. IoU is popular metric to measure overlap between segmentation masks, it is computed as following:

$$IoU = \frac{a_1 \cap a_2}{a_1 \cup a_2},$$

where $a_1,\ a_2$: $\mathbb{R}^d \rightarrow \mathbb{R}, a_1, a_2 \in \{0, 1\}^d$, and d is the dimension of the data (which in our case is 1920). a_1 and a_2 are two annotators' annotation of the same signal. The group's annotations had an average pairwise IoU (intersection over union) score of 0.68. The annotations from each pair of annotators were then compared and

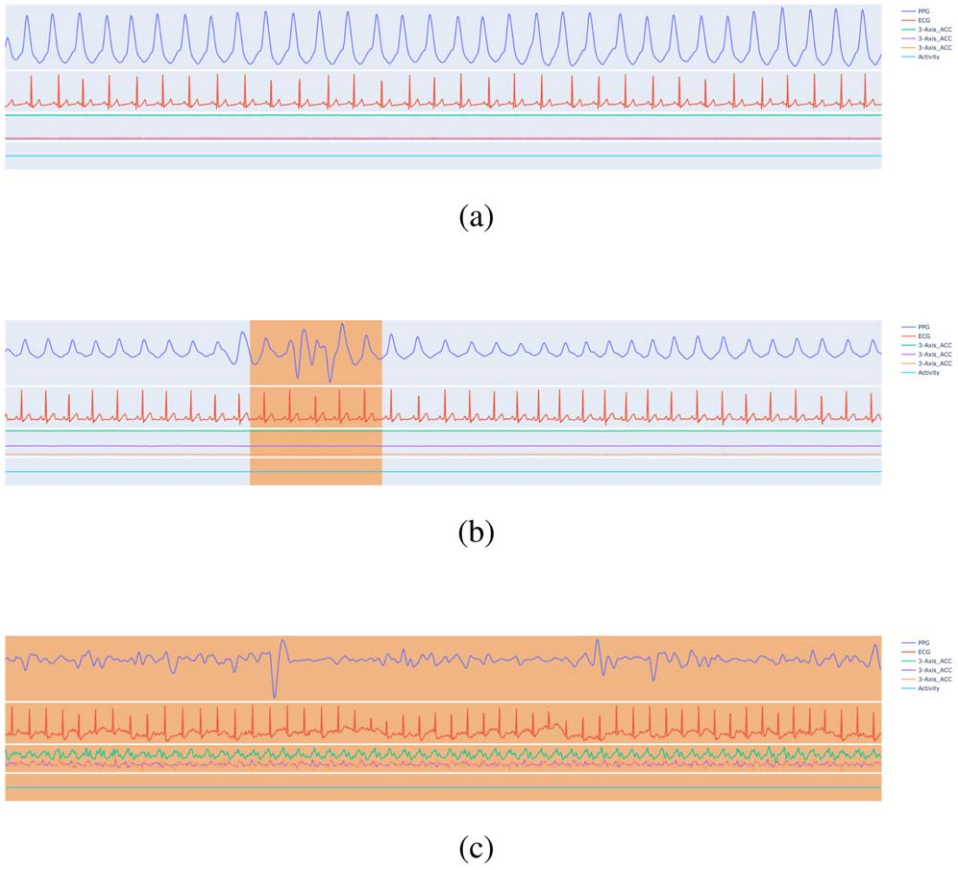

Figure 3. PPG signal annotation examples from the 'Annotation Inspection' interface, the orange overlay indicates the human annotation. (a) 'No Artifact' (b) partial artifact (c) 'All Artifact.'

analyzed, and the group of annotators jointly made decisions on the correct annotations. This permitted better agreement on the correct way to annotate these signals. The rest of the data were annotated by a single annotator afterwards.

4. Method

4.1. System overview

As illustrated by figure 1, the PPG DaLiA training set was used for training. The PPG DaLiA test set, WESAD and TROIKA datasets were used for evaluation.

4.2. Proposed model

We treated this problem as a semantic segmentation problem. Our proposed machine learning method is called *Segmentation-based Artifact Detection*, denoted Segade. Segade leverages U-Net's model architecture, based on U-Net's general high-quality performance on semantic segmentation problems in computer vision (Ronneberger *et al* 2015). However, each block of Segade is different from U-Net's blocks. Segade is shown in figure 4. Unlike U-Net (which uses plain convolutions), our encoder and decoder blocks consist of residual blocks; we also used 1D convolution layers instead of 2D convolution layers with padding to preserve dimension of the signals.

Each encoder block is composed of one convolution layer and a skipping convolutional layer with kernel size of 1, forming a residual structure, followed by a maxpooling layer. We used five blocks, with filter sizes of 16, 32, 64, 128 and 256, and kernel sizes of 80, 40, 20, 10, 5 from top to bottom. The initial layer kernel size of 80 was chosen in order for the kernel to cover at least 1–2 heart beats.

Each decoder block has the same residual structure as in the encoder block, followed by an upsampling layer. There are 5 decoder blocks in our network, with filter sizes of 256, 128, 64, 32, 16, respectively, and kernel sizes of 5, 10, 20, 40, 80 from bottom to top.

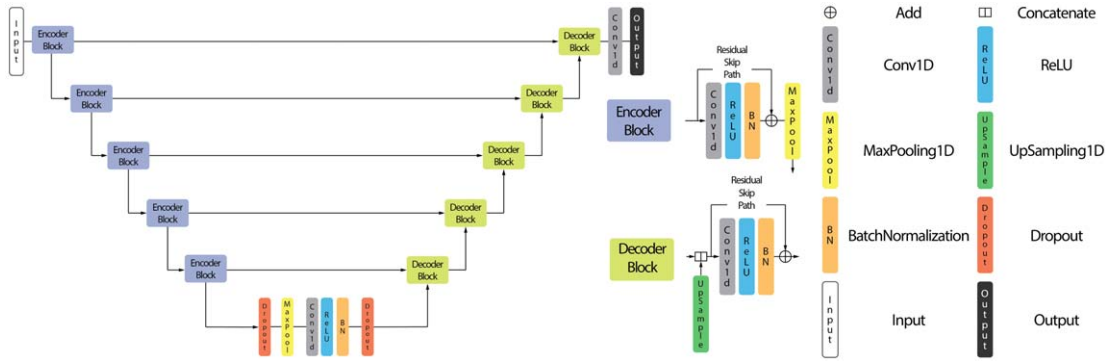

Figure 4. Proposed network. We used 1D convolution layers in each encoder and decoder block. Skip paths were added to form residual blocks inside each encoder and decoder block. Input dimensions of each convolution layer is preserved via zero-padding.

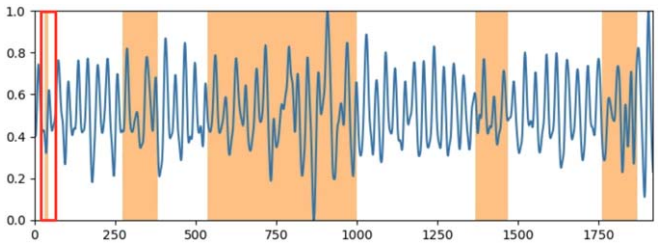

Figure 5. Sub-pulse-segment example. Orange overlays are proposed model's segmentation labels, the first segment highlighted by the red bounding box is smaller than one pulse, and is thus not useful.

Each model input is a 1D signal, which in our case consists of a 1920-dimensional vector (30 s of PPG measurements taken at a 64 Hz sampling rate). The output of the model has the same dimension, and each element of the output vector is a sigmoid score between 0 and 1. A threshold of 0.5 was applied to convert the model to generate a binary segmentation label.

4.3. Training loss, hyper-parameter selection, and implementation

The model was trained with the PPG-DaLiA training set. The hyper-parameters were selected via 10-fold cross validation on the training set. When choosing the loss function and selecting hyper-parameters, both segmentation accuracy and segmentation usefulness were taken into consideration. Segmentation accuracy is measured by the **DICE** score, defined as follows for Boolean operations:

$$\text{DICE Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$

In the original setting of DICE, the TP, FP and FN meant the 'True Positive,' 'False Positive' and 'False Negative' of the pixel classification; where a pixel with label 1 is positive, and a pixel with label 0 is negative. 'True Positive' means a pixel with label 1 is predicted as 1, 'False Positive' means a pixel with label 0 is predicted as 1, 'False Negative' means a pixel with label 1 is predicted as 0. In our setting, we have time steps instead of pixels (that is each data point in a PPG signal) and we can reformulate the DICE score as shown below

$$\text{DICE Score}(a,\, b) = \frac{2 \times \|a \cdot b\|_0}{\|a\|_0 + \|b\|_0},$$

where $a, b \in \{0, 1\}^d$ and d is the dimension of the data. Segmentation usefulness is a measure of how useful or interpretable a segmentation is; segmentation that generate unrealistically small artifacts too often are low quality. Segmentation usefulness is measured by the sub-pulse-segment quantity measure (SQM), which measures how often a model generates an artifact segment that is smaller than one pulse. Artifacts should not be less than one pulse, so these types of segments do not have physical meaning.

An example showing a segment that is smaller than a pulse is shown in the red bounding box of figure 5. Normal adult human heart rate ranges from 60 to 100 bpm, for comparison purpose we use an average 80 bpm (0.755 s per beat) as the threshold to identify segments that are smaller than a pulse.

Since we cannot easily optimize the DICE score directly because it is discrete, we used the 1D active contour loss (AC loss) within our algorithm's objective, which helps to promote both segmentation accuracy and segmentation usefulness. The active contour was first proposed by Kass *et al* (1988), and its first use with deep learning for biomedical image segmentation was proposed by Chen *et al* (2019); here we adapt it for 1D signal segmentation. The idea behind the active contour is to frame the problem of image segmentation as a minimization problem. The 'contour' being the boundary between foreground (artifact signal segments) and background (clean signal segments).

The active contour method employs contours that evolve over time. When the contour is outside the artifact, it would shrink; when the contour is inside the artifact, it would expand. The evolution of the contour is constrained by a minimization objective and the signal's ground truth values. The objective to be minimized is formulated as the error of the contour, taking into account both artifact timesteps that are outside the contour and non-artifact timesteps that are inside the contour. When this objective is minimized, the contour will reach the correct boundary as there will be few timesteps on the wrong side of the contour.

Our 1D AC loss has two terms: the magnitude of the transition (marked 'transition' in the equation below) and an energy term (marked 'region' in the equation). Specifically, let us define $\mathbf{u} \in (0,1)^d$, $\mathbf{v} \in \{0,\ 1\}^d$, and d is the dimension of the data (which in our case is 1920). Here, $\mathbf{v}$ and $\mathbf{u}$ are the ground truth labels and predicted segmentation labels respectively. v_j is 1 where the ground truth signal contains an artifact at position j, and 0 otherwise. Our **active contour loss** is expressed as follows:

$$L_{\mathrm{AC}}(\mathbf{u},\ \mathbf{v}) = \lambda_t \cdot \text{transition} + \text{region}, \tag{1}$$

where

$$\text{transition} = \frac{1}{d-1}\sum_{j=1}^{d-1}\sqrt{\nabla \mathbf{u}_j^2} = \frac{1}{d-1}\sum_{j=1}^{d-1}|\nabla \mathbf{u}_j|, \tag{2}$$

where we denote ∇ as the subtraction of two $d-1$-dimensional vectors, namely $\mathbf{u}[2{:}\ d] - \mathbf{u}[1{:}\ d-1]$. The transition term is a proxy for segmentation usefulness; it motivates fewer transitions between artifact and non-artifact, and encourages the model to generate less sub-pulse (i.e. smaller than a pulse) segments.

The second term is a type of classification error:

$$\text{region} = \frac{1}{d}\sum_{j=1}^{d}(1-v_j)^2 \cdot u_j + v_j^2 \cdot (1-u_j). \tag{3}$$

The first term in the region term sum handles cases where $v_j = 0$, meaning there is no artifact at location j. The penalty is proportional to how far u_j is from 0.

In our experiments, the trade-off term λ_t's value was selected by cross-validation for the combination of the segmentation accuracy (which is controlled by both terms) and segmentation usefulness (controlled by the transition term) as shown in figure 10.

The model was optimized using the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.005 and batch size of 64. During training, early stopping was used with a patience of 15 epochs on validation loss; i.e. after 15 epochs, if the validation accuracy does not decrease, the training will be stopped and the epoch with the lowest validation loss will be saved for evaluation. The maximum training epoch limit was set to 200. The learning rate was also scheduled to decrease to 0.001 after 15 epochs, and further decrease to 0.0005 after 35 epochs to achieve high-quality training results. We portrayed a simplified version of this process in Stage (5) in figure 1.

5. Experiment setup

In this section, we introduce the four baseline approaches that were used for comparison.

5.1. Baseline 1: convolution neural networks sliding window

For our first baseline, we used a convolutional classifier (classifying each 30 s interval in a sliding window) to produce segmentation masks. For each time step, we consider all the sliding windows that it sits in. If any of them are labeled by a machine learning method as 'artifact,' then we predict that this time step is part of an artifact. Our version of this baseline was based on work of Goh *et al* (2020). The machine learning model (figure 6) used for detecting artifacts contains three convolution-batch_normalization-maxpooling blocks. The convolution layers have filter sizes of 64, 64 and 128 and kernel sizes of 10, 5, and 3. We randomly selected 5000 3 s windows (2500 clean and 2500 artifact windows) from the PPG-DaLiA training set to train the classifier. Each of the windows was assigned a binary label based on whether any artifact appeared in that window. In testing, the 3 s windows were generated in a traditional sliding window fashion with a 1 s interval (2 s of overlap). The predicted binary segmentation labels were used for evaluation against ground truth segmentation labels created by human

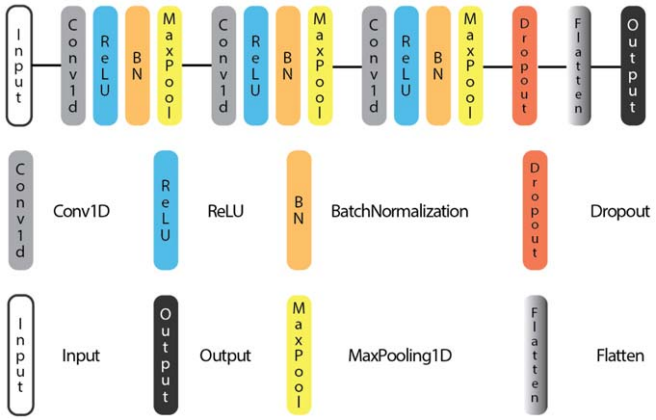

Figure 6. Structure of the CNN sliding window classifier.

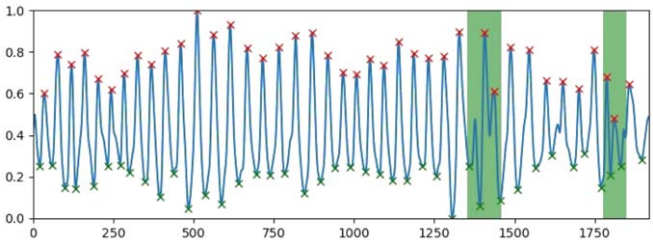

Figure 7. Pulse segmentation example. Top peaks and bottom peaks are indicated as red and green crosses. Between each pair of adjacent green crosses is one pulse. Green overlays are the human annotation of artifacts.

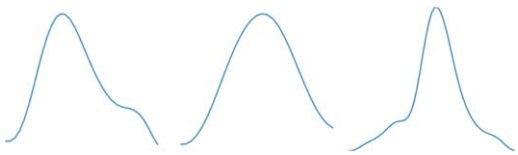

Figure 8. Three clean template examples (containing no artifacts).

annotators. The classifier was trained with binary cross-entropy loss and the Adam optimizer (Kingma and Ba 2014). The maximum number of training epochs was limited to 200.

5.2. Baseline 2: pulse segmentation template matching

This baseline method, inspired by Lim *et al* (2018)'s work, does not involve any machine learning. The signals from the PPG-DaLiA training set were first segmented into pulses via peak detection as shown in figure 7. From these pulses, 10 of them that have all clean (non-artifact) timesteps were chosen as templates for comparison with test pulses (i.e. figure 8). Each pulse in a test signal is compared against all 10 templates by calculating the dynamic time warping (DTW) distance between the template pulse and the test pulse. We used the fast DTW implementation based on Salvador and Chan (2004)'s work for our experiments. The minimum distance (among comparisons of the test pulse to the 10 templates) was calculated. A threshold of 1 was applied to this minimum DTW distance to generate a binary label (0 for clean and 1 for artifact). That is, we calculate:

$$a = \min_{\text{templates}} \text{distance}_{\text{DTW}}(\text{template}, \text{testpulse}).$$

If $a >$ threshold then we classify the pulse (that is, all time steps within the pulse) as an artifact.

Table 1. Methods results (DICE score) on datasets.

	PPG-DaLiA test set	WESAD	TROIKA
Baseline 1	0.8068 ± 0.0014	0.8446 ± 0.0013	0.7247 ± 0.0050
Baseline 2	0.6974 ± 0.0323	0.6954 ± 0.0309	0.6748 ± 0.0122
Baseline 3	0.7129 ± 0.0010	0.7372 ± 0.0024	0.6989 ± 0.0034
Baseline 4	0.6748 ± 0.0000	0.6634 ± 0.0000	0.6849 ± 0.0001
Proposed model	0.8734 ± 0.0018	0.9114 ± 0.0033	0.8050 ± 0.0116

5.3. Baselines 3 and 4: segmentation via post-hoc explanation techniques

In this baseline experiment, we explored the possibility to extract segmentation labels from a classification model. We used the Resnet-34 architecture proposed by Dai *et al* (2016) for 1D signal binary classification ('clean' and 'artifact'). This is a classic image classification architecture that was adapted for time series. Since we have a relatively small amount of training data, we performed transfer learning on a pre-trained Resnet34-1D PPG signal quality classifier by Zhang *et al* (2021). This pre-trained model was trained on the UCSF PPG dataset by the authors of Pereira *et al* (2019). We retrained the last two residual blocks, global average pooling and the last dense layer.

Here, we have switched to classification loss, so we need to define classification labels. Thus, to create the training ground truth labels, if there were any artifact timesteps in a signal, the signal was labeled as an artifact, and all other signals were labeled as non-artifact. After this, the training set contained 175 clean signals and 3261 artifact signals.

The model was trained with the Adam optimizer (Kingma and Ba 2014) and the binary cross-entropy loss. The initial learning rate was 10^{-5}, scheduled to decrease to 5×10^{-6} after 10 epochs and decrease further to 1×10^{-6} after 50 epochs. The maximum number of training epochs was set to 100.

The Resnet34-1D network by itself is only a classifier. In order to generate segmentation labels, we made the assumption that the model would focus its attention on artifacts in order to make the prediction for an artifact signal. Thus, two popular post-hoc explanation approaches were used to generate the model's attention, described next.

5.3.1. ResNet-34 with Grad-CAM (Baseline 3)

The gradient-weighted class activation mapping (Grad-CAM) was introduced by Selvaraju (2016). It is an approach to estimate attention for a deep convolutional neural network's prediction by generating a localization of class-discriminative attention, meaning that it aims to determine which part of an observation the network is paying attention to. In our experiments, after the Grad-CAM values were calculated and normalized, timesteps with Grad-CAM values above 0 were predicted as artifacts, and other timesteps were predicted as clean; this is how we generated a binary segmentation prediction for artifact signals.

5.3.2. ResNet-34 with SHAP (Baseline 4)

SHAP (Lundberg and Lee 2017) is also an approach to estimate attention for black-box models. The 'shap' library of Lundberg and Lee (2017) was used for this experiment. In a classification task, the SHAP algorithm uses baselines to calculate the marginal contribution of each feature. During the calculation of SHAP values, clean (non-artifact) signals from the training set were used as baselines. Positive SHAP values from 'artifact' predictions and negative SHAP values from 'clean' predictions were added to generate the final artifact SHAP values. The SHAP values were then normalized and smoothed by a Gaussian filter because they tended to be non-smooth, leading to many very small intervals with 'artifact' predictions. Finally, timesteps with SHAP values above 0 were assigned as 'artifact' and other timesteps were assigned as 'non-artifact' to generate a binary segmentation label for artifact signals.

6. Results

The test split from the PPG-DaLiA, TROIKA and WESAD datasets were used for evaluation purposes for all algorithms. We compared the model-predicted segmentation labels with the ground truth segmentation labels and calculated DICE scores.

Our main result is *the proposed model outperformed all baseline models by a large margin on all datasets.* Table 1 shows that our model exceeded all other methods by around 7 percentage point on PPG DaLiA, WESAD and TROIKA datasets. We also provide visual comparisons of segmentation results between the proposed method and baselines in appendix.

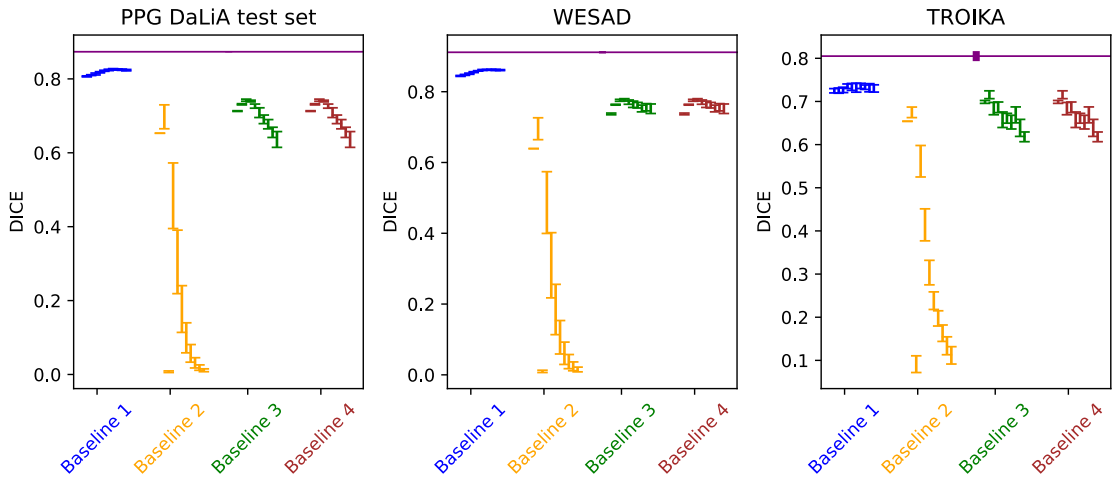

Figure 9. Sensitivity analysis results. DICE scores of different models under different threshold settings for PPG-DaLiA test set, TROIKA dataset, WESAD dataset. Intervals were created via 10-fold cross validation. The purple horizontal line represents the performs of the proposed model.

Table 2. DICE scores and sub-pulse-segment quantity measure (SQM) on validation folds for different loss functions. Higher DICE score is better, lower SQM score is better.

	DICE	SQM
Binary cross-entropy loss	0.8999 ± 0.0058	9.39 ± 1.30
DICE Loss	0.9061 ± 0.0133	5.41 ± 2.47
AC Loss($\lambda_t = 0$)	0.9172 ± 0.0021	2.12 ± 0.22
AC Loss($\lambda_t = 4$)	0.9127 ± 0.0072	0.18 ± 0.03

For our task, the active contour loss function outperforms other popular loss functions including binary cross-entropy and the DICE loss in both segmentation accuracy and segmentation usefulness, as shown in table 2, where DICE score is a measure of segmentation accuracy and segmentation usefulness is measured by SQM as described in section 4.3, hyper-parameter selection, and implementation.

The value of λ_t was set to 4 where DICE and SQM reached their best values, as shown in figure 10.

We conducted a sensitivity analysis to check robustness of the results. In particular, we tested different values for the thresholds (threshold used for determining binary classification labels) for Baseline 1, 3 and 4, ranging from 0 (which is the default value) to 0.9 with a step of 0.1. Different DTW thresholds ranging from 0 to 10 with a step of 1 were also tested for Baseline 2. As shown in figure 9, each bar represents the DICE score result of the models under different threshold settings. Our proposed model's DICE score is represented as the purple horizontal line, since it does not require any threshold setting. The proposed model still out-performed the baselines regardless of threshold parameter settings on all datasets.

7. Discussion

In this study, we demonstrated the performance of our proposed model for PPG signal artifact segmentation. By comparison to the four baseline methods, the proposed model has shown superiority in identifying the areas of artifacts within a 30 s PPG signal strip.

The methods we studied can be divided into two categories, global methods (proposed model Segade, Baseline 3 and 4) and local methods (Baseline 1 and 2). Comparing these two categories, local methods all performed worse than the proposed model. Baseline 3 and 4 performed the worst among all methods. We analyzed the difference between proposed model and local methods with example visualizations. Our proposed model could identify small segments of clean signals even if they are surrounded by artifacts, as demonstrated by figure 14. The proposed method takes not only the local features of waveforms into account, but also the global features of the whole signal, which makes it superior to Baselines 1 and 2. (For instance, figures 11 and 12 show how both Baseline 1 and Baseline 2 segmented artifacts incorrectly.) Both Baseline 1 and Baseline 2 are local methods; they rely on sub-sequence/sliding window classifications. Local methods tend to be less

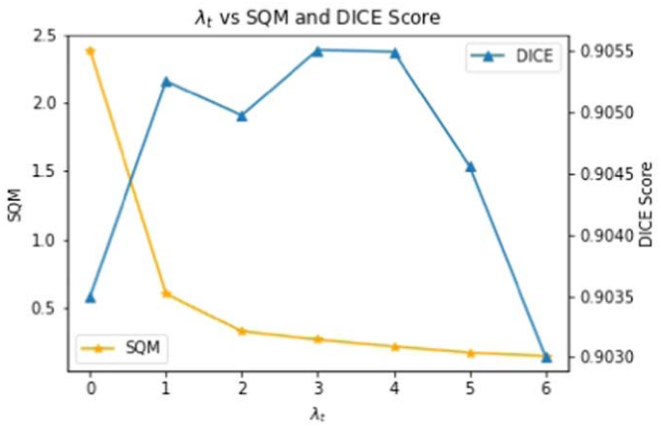

Figure 10. λ_t vs sub-pulse-segment quantity measure (SQM) plot on validation set. The average amount of sub-pulse-segments per signal decreases as λ_t increases. λ_t values ranging from 0 to 6 were tested. We selected $\lambda_t = 4$ based on the high DICE score and low SQM at $\lambda_t = 4$.

Table 3. PPG DaLiA subject information (Reiss *et al* 2019).

Subject ID	Gender	Age (years)	Height (cm)	Weight (kg)	Skin type	Fitness
S1	m	34	182	78	3	6
S2	m	28	189	80	3	5
S3	m	25	170	60	3	5
S4	m	25	168	57	4	5
S5	f	21	180	70	3	4
S6	f	37	176	70	3	1
S7	f	21	168	58	3	2
S8	m	43	179	70	3	5
S9	f	28	167	60	4	5
S10	f	55	164	56	4	5
S11	f	24	168	62	3	5
S12	m	43	195	105	3	5
S13	f	21	170	63	3	6
S14	f	26	170	67	3	4
S15	m	28	183	79	2	5

computationally efficient than global methods, as they need to classify all the sub-sequences to generate results. Sliding-window-based methods (Baseline 1) have another disadvantage in that they must label all timesteps in a sub-window as either artifact or signal, whereas our approach generates labels for individual timesteps, which enables high-precision artifact segmentation and localization. When comparing the two local methods, Baseline 1 has a better DICE score than Baseline 2. Baseline 1 is a more sophisticated learning-based method while Baseline 2 has more limited capacity as it relies on a limited number of templates and a fix threshold. Baselines 3 and 4 were based on the same hypothesis that the model would focus its attention on artifact time-steps when classifying artifact signals. We explored the possibility of extracting segmentation masks from classification model. However, the performance of the methods was poor. Baselines 3 and 4's transfer learning model was trained on an extremely imbalanced dataset (175 clean signals and 3261 artifact signals), which potentially contributed to the poor performance. When comparing across different datasets, all methods except Baseline 4 performed the worst on the TROIKA dataset compared to the PPG-DaLiA and WESAD dataset. We believe this was caused by the high motion intensity and poor signal quality from this dataset.

The limitation of our study mainly lies in the lack of skin color diversity in the training set. PPG sensors are also known to be sensitive to skin color. According to the information provided by the PPG DaLiA dataset authors (Reiss *et al* 2019) and the Fitzpatrick scale (Fitzpatrick 1988) as shown in table 3, the subjects all have beige, olive and light brown skin. Future studies could address this issue by incorporating data with both darker and paler skin colors via transfer learning and further evaluations on a more comprehensive dataset.

The implementation of the annotation tool has greatly accelerated and assisted our study. To our knowledge, previously there did not exist an open source tool built for PPG signal annotations, our tool successfully allowed

the annotators to perform the task of annotating segments from PPG signals. This tool is specialized, light and easy to operate and deploy.

In complicated signal processing problems, we are used to feeding large amounts of coarsely-labeled data into a black box and expecting it to transform into a model that generates accurate predictions (e.g. 78278 30-seconds segments used for training in Pereira *et al* (2019)'s study on PPG signal quality classification). Our work, which leveraged an annotation tool, showed that even with a small amount of finely labeled data (and a carefully-designed architecture), we can predict better.

This illustrates a message that generalizes across domains, which is that a small amount of fine-grained annotated data is often helpful to achieve better results faster.

8. Conclusion

In this study, we proposed an effective supervised segmentation model, paired with a novel loss function to accurately locate and segment artifacts from PPG signals. We compared the proposed method's performance against four baseline approaches including convolutional neural networks sliding windows, pulse segmentation template matching, and segmentation via Grad-CAM and SHAP 'explainers.' We evaluated the proposed method and all baseline approaches on three datasets with comprehensive representations of different levels of artifacts in different environments. Our proposed model outperforms other baseline methods. It provides an end-to-end solution for artifact detection and segmentation without the need to adjust additional parameters. For reproducibility, the baselines and proposed algorithm implementation along with annotated data for have been published at https://github.com/chengstark/Segade and the annotation tool has been published at https://github.com/chengstark/Segade-Annotation-Tool.

Appendix.

A.1. Dataset information details

A.1.1. PPG DaLiA dataset. 15 subjects participated in this study. Table 3 shows detailed subject information from PPG DaLiA dataset. Skin type (according to the Fitzpatrick scale (Fitzpatrick 1988)) fitness level (how often does the subject do sports; on a scale 1–6 where 1 refers to less than once month and 6 refers to 5–7 times a week) (Reiss *et al* 2019). Table A1 shows detailed subjects' activity information from PPG DaLiA dataset. Each activity is measured by minutes, there could be transitional activity in between two activities listed in the table. For more detailed information please visit PPG DaLiA website https://archive.ics.uci.edu/ml/datasets/PPG-DaLiA#.

A.1.2. WESAD dataset. 17 subjects participated in this study, subject ID S1 and S12's data was discarded due to sensor malfunction (Schmidt *et al* 2018). Table A2 contains subject information for the WESAD dataset. These are the information extracted from the dataset subject readme file. Schmidt *et al* (2018) indicated in their experiment the subjects were either standing or sitting, but movements or activities were not clearly stated in other stages. Subjects performed the following tasks: (1) sitting or standing at a table and neutral reading material (magazines) was provided; (2) watched a set of eleven funny video clips; (3) public speaking and a mental

Table A1. PPG DaLiA activity information (Reproduced from Reiss *et al* 2019 CC BY 4.0.).

Activity	Duration (min)
Sitting still	10
Ascending/Descending starirs	5
table scoccer	5
Cycling	8
Driving car	15
Lunch break	30
Walking	10
Working	20

Table A2. WESAD subject information (Reiss *et al* 2019).

Subject ID	Age (years)	Height (cm)	Weight (kg)	Gender	Dominant hand
S2	27	175	80	male	right
S3	27	173	69	male	right
S4	25	175	90	male	right
S5	35	189	80	male	right
S6	27	170	66	male	right
S7	28	184	74	male	right
S8	27	172	64	female	left
S9	26	181	75	male	right
S10	28	178	76	male	right
S11	26	171	54	female	right
S13	28	181	82	male	right
S14	27	180	80	male	right
S15	28	186	83	male	right
S16	24	184	69	male	right
S17	29	165	55	female	right

arithmetic task; (4) guided meditation. For more detailed information please visit WESAD website https://archive.ics.uci.edu/ml/datasets/WESAD+%28Wearable+Stress+and+Affect+Detection%29.

A.1.3. TROIKA dataset. The TROIKA Dataset contains 12 subjects' data. These subjects has age from 18 to 35, detailed per subject information such as skin color, weight or height was not provided. Subjects ran on treadmill with changing speed, they performed two types of speed sequences indicated as following: (1) rest(30 s) → 8 km h^{-1}(1 min) → 15 km h^{-1}(1 min) → 8 km h^{-1}(1 min) → 15 km h^{-1}(1 min) → rest(30 s); (2) rest(30 s) → 6 km h^{-1}(1 min) → 12 km h^{-1}(1 min) → 6 km h^{-1}(1 min) → 12 km h^{-1}(1 min) → rest(30 s) (Zhang 2015). For more details about TROIKA dataset or TROIKA Framework please refer to 'TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise' and visit https://sites.google.com/site/researchbyzhang/ieeespcup2015.

A.2. Segmentation results visualization

Here we show three of the visualizations of segmentation results of different approaches in figures 11 12, 13, 14, and offer side to side comparisons between methods.

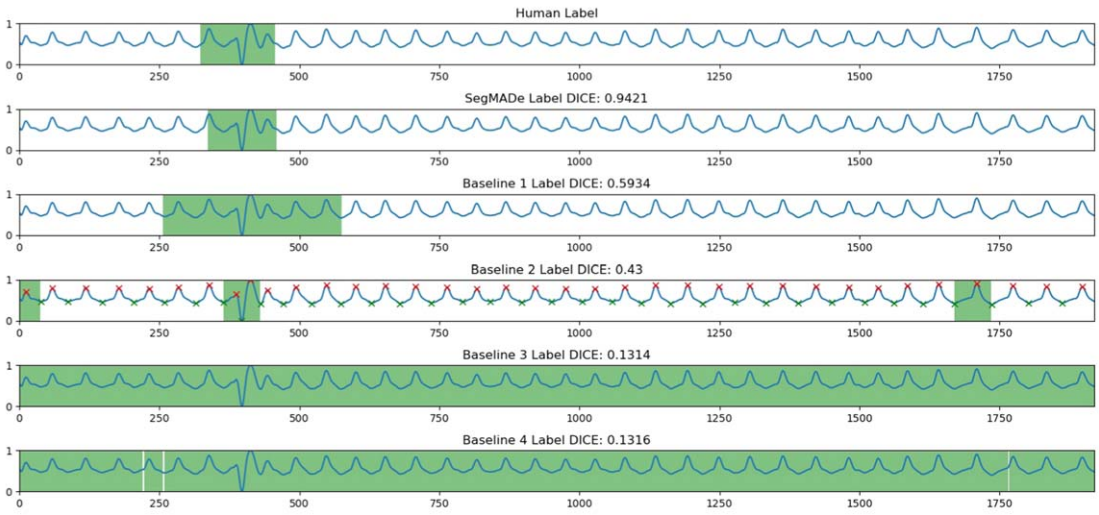

Figure 11. Example from PPG DaLiA test set.

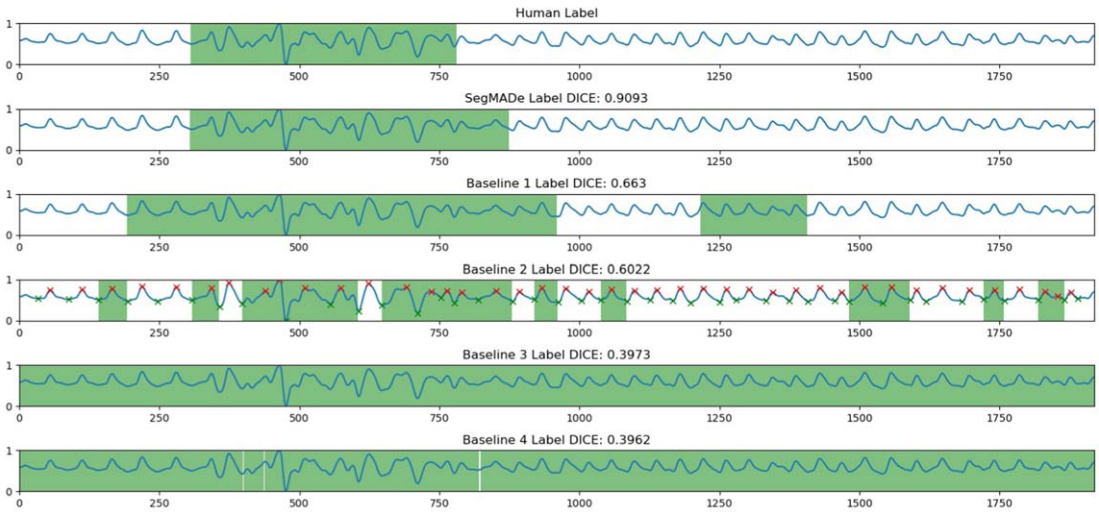

Figure 12. Example from PPG DaLiA test set.

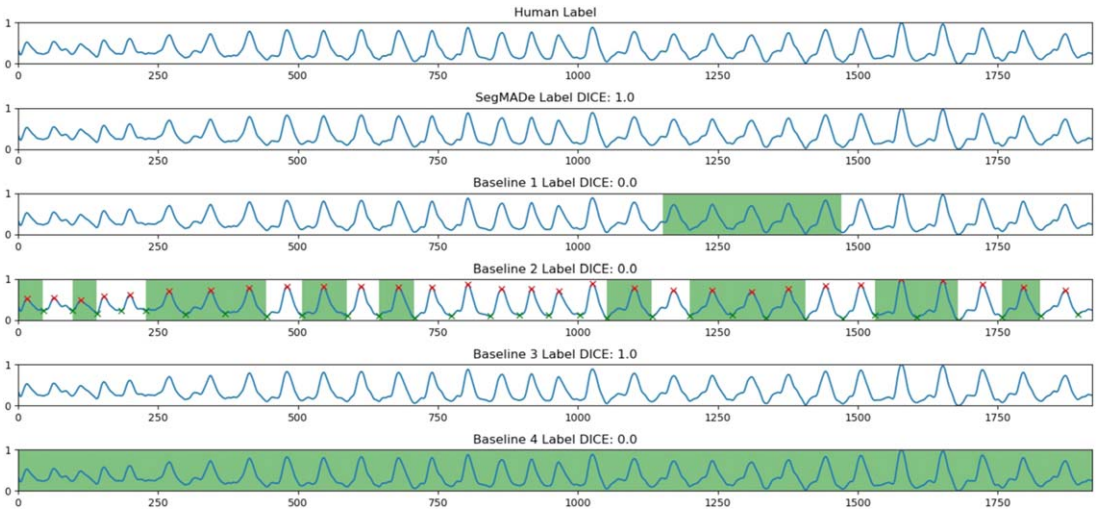

Figure 13. Example from WESAD dataset.

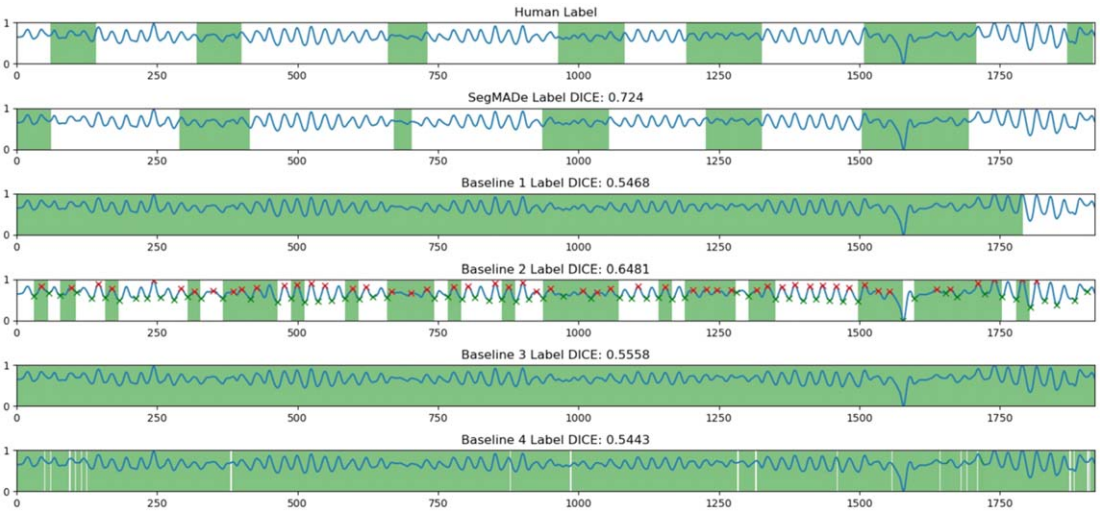

Figure 14. Example from TROIKA dataset.

ORCID iDs

Xiao Hu https://orcid.org/0000-0001-9478-5571

References

Allen J, Overbeck K, Stansby G and Murray A 2006 Photoplethysmography assessments in cardiovascular disease *Meas. Control* **39** 80–3

Athaya T and Choi S 2020 Evaluation of different machine learning models for photoplethysmogram signal artifact detection *2020 Int. Conf. on Information and Communication Technology Convergence (ICTC)* pp 1206–8

Bashar S K, Han D, Ding E, Whitcomb C, McManus D D and Chon K H 2019 Smartwatch Based Atrial Fibrillation Detection from Photoplethysmography Signals *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**

Castaneda D, Esparza A, Ghamari M, Soltanpur C and Nazeran H 2018 A review on wearable photoplethysmography sensors and their potential future applications in health care *Int. J. Biosens. Bioelectron.* **4** 195–202

Chen X, Williams B M, Vallabhaneni S R, Czanner G, Williams R and Zheng Y 2019 Learning active contour models for medical image segmentation *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 11632–40

Cherif S, Pastor D, Nguyen Q T and L'Her E 2016 Detection of artifacts on photoplethysmography signals using random distortion testing *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (Oct. 2016)*

Chong J W, Dao D K, Salehizadeh S M A, McManus D D, Darling C E, Chon K H and Mendelson Y 2014 Photoplethysmograph signal reconstruction based on a novel hybrid motion artifact detectionreduction approach: I. Motion and noise artifact detection *Ann. Biomed. Eng.* **42** 2238–50

Dai W, Dai C, Qu S, Li J and Das S 2016Very deep convolutional neural networks for raw waveforms *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

Dall'Olio L, Curti N, Remondini D, Safi Harb Y, Asselbergs F W, Castellani G and Uh H-W 2020 Prediction of vascular aging based on smartphone acquired ppg signals *Sci. Rep.* **10** 19756

Dao D, Salehizadeh S M A, Noh Y, Chong J W, Cho C H, McManus D, Darling C E, Mendelson Y and Chon K H 2017 A robust motion artifact detection algorithm for accurate detection of heart rates from photoplethysmographic signals using time-frequency spectral features *IEEE J. Biomed. Health Inform.* **21** 1242–53

Fischer C, Domer B, Wibmer T and Penzel T 2017 An algorithm for real-time pulse waveform segmentation and artifact detection in photoplethysmograms *IEEE J. Biomed. Health Inform.* **21** 372–81

Fitzpatrick T B 1988 The validity and practicality of sun-reactive skin types I through VI *Arch Dermatol.* **124** 869–71

Foo J Y A and Wilson S J 2006 A computational system to optimise noise rejection in photoplethysmography signals during motion or poor perfusion states *Med. Biol. Eng. Comput.* **44** 140–5

Goh C H, Tan L K, Lovell N H, Ng S C, Tan M P and Lim E 2020 Robust PPG motion artifact detection using a 1-D convolution neural network *Comput. Methods Programs Biomed.* **196** 105596

Henriksen A, Haugen Mikalsen M, Woldaregay A Z, Muzny M, Hartvigsen G, Hopstock L A and Grimsgaard S 2018 Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables *J. Med. Internet Res.* **20** e110

Ioannidis D C, Kapasouri E M and Vassiliou V S 2019 Wearable devices: monitoring the future? *Oxford Med. Case Rep.* **2019** 492–4

Karlen W, Kobayashi K, Ansermino J M and Dumont G A 2012 Photoplethysmogram signal quality estimation using repeated gaussian filters and cross-correlation *Physiol. Meas.* **33** 1617–29

Kass M, Witkin A and Terzopoulos D 1988 Snakes: active contour models *Int. J. Comput. Vision* **1** 321–31

Kim S H, Ryoo D W and Bae C S 2007 Adaptive Noise Cancellation Using Accelerometers for the PPG Signal from Forehead *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*

Kingma D and Ba J 2014 A method for stochastic optimization *Int. Conf. on Learning Representations*

Kirchhof PE. S. D.Group *et al* 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS *Eur. Heart J.* **37** 2893–962

Krishnan R, Natarajan B and Warren S 2008 Analysis and detection of motion artifact in photoplethysmographic data using higher order statistics *2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing* pp 613–6

Krishnan R, Natarajan B and Warren S 2008 Analysis and detection of motion artifact in photoplethysmographic data using higher order statistics *2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing* pp 613–6

Lee B, Han J, Baek H J, Shin J H, Park K S and Yi W J 2010 Improved elimination of motion artifacts from a photoplethysmographic signal using a kalman smoother with simultaneous accelerometry *Physiol. Meas.* **31** 1585–603

Lee J, Jung W, Kang I T, Kim Y and Lee G 2004 Design of filter to reject motion artifact of pulse oximetry *Comput. Standards Interfaces* **26** 241–9

Li K, Warren S and Natarajan B 2012 Onboard tagging for real-time quality assessment of photoplethysmograms acquired by a wireless reflectance pulse oximeter *IEEE Trans. Biomed. Circuits Syst.* **6** 54–63

Li Q and Clifford G D 2012 Dynamic time warping and machine learning for signal quality assessment of pulsatile signals *Physiol. Meas.* **33** 1491–501

Li Q, Mark R G and Clifford G D 2007 Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter *Physiological Measurement* **29** 15

Liang Y, Chen Z, Ward R and Elgendi M 2018 Hypertension assessment via ecg and ppg signals: An evaluation using mimic database *Diagnostics* **8**

Lim P K, Ng S C, Lovell N H, Yu Y P, Tan M P, McCombie D, Lim E and Redmond S J 2018 Adaptive template matching of photoplethysmogram pulses to detect motion artefact *Physiol. Meas.* **39** 105005

Liu X, Hu Q, Yuan H and Yang C 2020 Motion artifact detection in ppg signals based on gramian angular field and 2-d-cnn *2020 13th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* pp 743–7

Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *NIPS* pp 4765–74

McConnell M V, Turakhia M P, Harrington R A, King A C and Ashley E A 2018 Mobile health advances in physical activity, fitness, and atrial fibrillation: moving hearts *J. Am. Coll. Cardiol.* **71** 2691–701

Naraharisetti K V P and Bawa M 2011 Comparison of different signal processing methods for reducing artifacts from photoplethysmograph signal *2011 Ieee International Conference on Electro/information Technology*

Ouyang V, Ma B, Pignatelli N, Sengupta S, Sengupta P, Mungulmare K and Fletcher R R 2020 The use of multi-site photoplethysmography (PPG) as a screening tool for coronary arterial disease and atherosclerosis *Physiol. Meas.* **42** 064006

Papini G B, Fons P, Aubert X L, Overeem S, Bergmans J W M and Vullings R 2017 *Photoplethysmography Beat Detection and Pulse Morphology Quality Assessment for Signal Reliability Estimation* (Institute of Electrical and Electronics Engineers Inc.) pp 117–20

Pereira T, Ding C, Gadhoumi K, Tran N, Colorado R A, Meisel K and Hu X 2019 Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation *Physiol. Meas.* **40** 125002

Pereira T, Tran N, Gadhoumi K, Pelter M M, Do D H, Lee R J, Colorado R, Meisel K and Hu X 2020 Photoplethysmography based atrial fibrillation detection: a review *Npj Digital Med.* **3** 3

Raja J M, Elsakr C, Roman S, Cave B, Pour-Ghaz I, Nanda A, Maturana M and Khouzam R N 2019 Apple watch, wearables, and heart Rhythm: where do we stand? *Ann. Transl. Med.* **7** 417

Ram M R, Madhav K V, Krishna E H, Komalla N R and Reddy K A 2012 A Novel Approach for Motion Artifact Reduction in PPG Signals Based on AS-LMS Adaptive Filter *IEEE Transactions on Instrumentation and Measurement* **61** 1445–57

Reiss A, Indlekofer I, Schmidt P and Van Laerhoven K 2019 Deep ppg: large-scale heart rate estimation with convolutional neural networks *Sensors* **19**

Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Int. Conf. on Medical image computing and computer-assisted intervention* (Berlin: Springer) pp 234–41

Salvador S and Chan P 2004 Toward accurate dynamic time warping in linear time and space *Intelligent Data Analysis* **11** 561-580

Sañudo B, De Hoyo M, Muñoz-López A, Perry J and Abt G 2019 Pilot study assessing the influence of skin type on the heart rate measurements obtained by photoplethysmography with the apple watch *J. Med. Syst.* **43** 195

Saritas T, Greber R, Venema B, Puelles V G, Ernst S, Blazek V, Floege J, Leonhardt S and Schlieper G 2019 Non-invasive evaluation of coronary heart disease in patients with chronic kidney disease using photoplethysmography *Clin. Kidney J.* **12** 538–45

Schack T, Sledz C, Muma M and Zoubir A M 2015 *A New Method for Heart Rate Monitoring During Physical Exercise Using Photoplethysmographic Signals* (Institute of Electrical and Electronics Engineers Inc.) pp 2666–70

Schmidt P, Reiss A, Duerichen R, Marberger C and Van Laerhoven K 2018 Introducing wesad, a multimodal dataset for wearable stress and affect detection *Proc. 20th ACM Int. Conf. on Multimodal Interaction* (ser. ICMI '18, Boulder, CO, USA: Association for Computing Machinery) pp 400–8

Selvaraj N, Mendelson Y, Shelley K H, Silverman D G and Chon K H 2011 Statistical approach for the detection of motion/noise artifacts in Photoplethysmogram *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2011** 4972–5

Selvaraju R R, Das A, Vedantam R, Cogswell M, Parikh D and Batra D 2016 Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization *2017 IEEE International Conference on Computer Vision (ICCV)*

Sukor J A, Redmond S J and Lovell N H 2011 Signal quality measures for pulse oximetry through waveform morphology analysis *Physiol. Meas.* **32** 369–84

Tabei F, Kumar R, Phan T N, McManus D D and Chong J W 2018 A novel personalized motion and noise artifact (mna) detection method for smartphone photoplethysmograph (ppg) signals *IEEE Access* **6** 60498–512

Vandecasteele K, Lázaro J, Cleeren E, Claes K, Van Paesschen W, Van Huffel S and Hunyadi B 2018 Artifact detection of wrist photoplethysmograph signals *SciTePress* **4** 182–9

Zhang O, Ding C, Pereira T, Xiao R, Gadhoumi K, Meisel K, Lee R J, Chen Y and Hu X 2021 Explainability metrics of deep convolutional networks for photoplethysmography quality assessment *IEEE Access* **9** 29736–45

Zhang Y, Song S, Vullings R, Biswas D, Simões-Capela N, Van Helleputte N, Van Hoof C and Groenendaal W 2019 Motion artifact reduction for wrist-worn photoplethysmograph sensors based on different wavelengths *Sensors* **19** 673

Zhang Z, Pi Z and Liu B 2015 Troika: a general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise *IEEE Trans. Biomed. Eng.* **62** 522–31