Atrial Fibrillation Detection From PPG

Yoav Meiri[1]
[1] Technion - Israel Institute of Technology
Haifa 32000 Israel

Abstract

Atrial fibrillation (AF) is the most prevalent arrhythmia, resulting in varying and irregular heartbeats. AF increases risk for numerous cardiovascular diseases including stroke, heart failure and as a result, computer aided efficient monitoring of AF is crucial, especially for intensive care unit (ICU) patients.

A subset of the Medical Information Mart for Intensive Care (MIMIC) III database containing 35 subjects is used in this study. We compare the AF detection performance of several classifiers for both the training and blinded test data. We present two feature sets which combine both local and global information (both morphological and heart rate variability features) and evaluate their predictive ability in the AF detection task.

1. Introduction

Atrial fibrillation (AF) is a prevalent heart rhythm disorder that affects a significant number of adults in the United States, ranging from 2.7 million to 6.1 million individuals. It is projected that this figure will double over the next 25 years [5]. AF poses various risks, including blood clots, cognitive impairment, heart failure, and stroke. The standard method of diagnosis involves observing the electrical activity of the heart through an electrocardiogram (ECG), typically performed using devices such as a cardiac event recorder, Holter monitor, or chest patch. However, these ECG devices are primarily used reactively rather than proactively, leading to many cases of undetected subclinical or silent AF. Regrettably, this lack of detection contributes to a significant proportion of ischemic strokes.

An emerging technology known as photoplethysmography (PPG) allows for non-invasive measurement of heart rhythm using optical sensing. PPG sensors detect changes in blood volume within the microvascular tissue bed by utilizing low-intensity light. The optical mechanism employed by PPG sensors enables their integration into wearable devices like smartwatches.

Utilizing PPG sensors for AF detection offers several advantages over ECG sensors. Unlike ECG event recorders, PPG sensors can continuously monitor heart rhythm without requiring active involvement from the user. ECG recorders need to be manually activated by the user when symptoms arise. Consequently, PPG sensors can more accurately quantify the burden of AF, which refers to the percentage of time an individual's heart rhythm is in AF. This burden serves as a more reliable risk factor for heart attacks compared to a simple presence or absence of AF. Moreover, PPG sensors are already incorporated into widely-used smartwatches, making PPG-based monitoring seamless and comfortable for extended periods of time. In contrast, ECG-based monitors can be inconvenient and less comfortable. Therefore, continuous AF monitoring through PPG sensors in mainstream smartwatches presents a convenient and cost-effective solution for proactive AF screening. This approach has the potential to detect challenging AF cases such as paroxysmal and silent AF, which often go undiagnosed through reactive ECG-based screening methods.

The application of deep learning algorithms to PPG waveforms is particularly promising, as deep learning algorithms can learn highly predictive models from raw data.

2. Methods

2.1. Data

The MIMIC PERform AF Dataset contains 20 minutes of data from 19 patients in AF, and 16 patients in normal sinus rhythm (non-AF). It was used to compare between AF and normal sinus rhythm and classify between these two states. Labels of AF were obtained from manual annotations by cardiologists [4] [3]. The PPG equipment used for this dataset is a bedside monitor at 125 Hz (mostly finger PPG recordings). In number of beats, in total this dataset contains 29,592 AF beats and 22,477 non-AF beats.

We first split the data into train (80%) and test (20%) subject sets while keeping the AF and non-AF ratio (16:19). We then split the training set to train (97.8%) and validation (12.5%) sets where for each train subject, the last 12.5% of his signal is taken as validation and the rest is train, as demonstrated in 1. Notice that the train and validation sets are taken from a subject set which is disjoint to the set of subjects the test set was taken from.

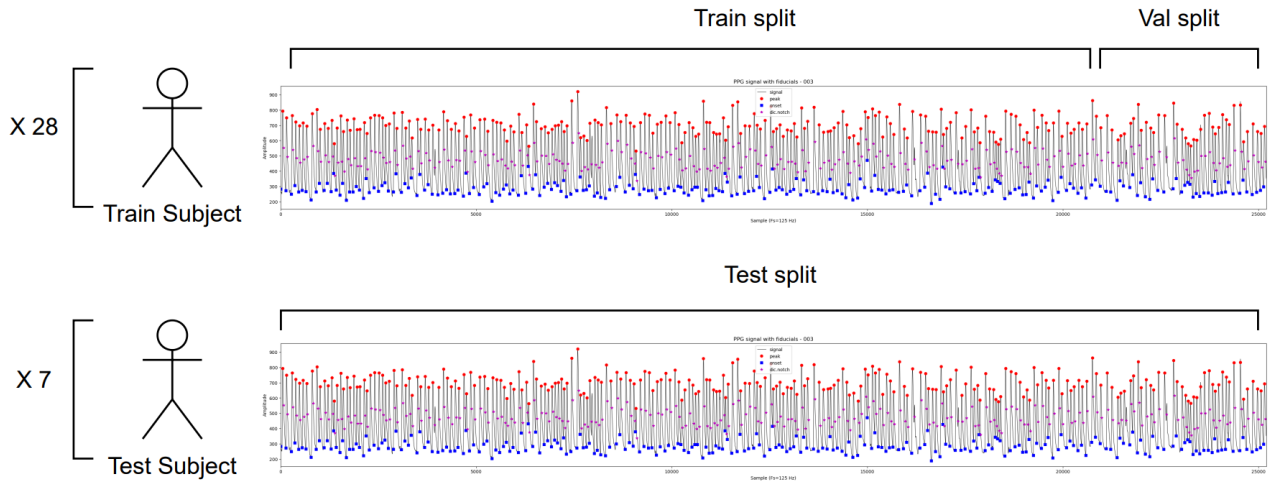

Figure 1. A demonstration of the data splitting process.

2.2. Prefiltering and Fiducial Points Detection

For the prefiltering and feature extraction in this work we use the pyPPG library [6]. Before computing the PPG morphological biomarkers, prefiltering of the raw PPG time series is performed to remove the baseline wander as well as remove high-frequency noise. The following filters have been applied to the raw signals:

- **4th order Chebyshev Type II 0.5-12 Hz band-pass filtering for original signal:** The 12 Hz low-pass filtering has two main reason. The first one was to avoid the time-shifting of a given fiducial point, particularly the systolic onset, and dicrotic notch. The second reason was to eliminate the unwanted frequency contents from the PPG derivatives. The 0.5 Hz high-pass filtering was used to minimize the baseline wondering of the PPG signal.
- **20 ms moving average filtering (MAF) for band-pass filtered signal:** In the case of very noisy signals, some high-frequency content can remain in the band-pass filter signal.
- **10 ms MAF for the PPG derivatives:** To eliminate the high-frequency content in the PPG derivatives, a 10 ms MAF with 45 Hz cut-off frequency has been applied.

For the fiducial points detection, we extracted the following types: pulse onset, systolic peak, dicrotic notch, diastolic peak, pulse offset. The pyPPG library uses the improved Automatic Beat Detector presented in *Aboy et al. 2005* [1].

2.3. Problem Formulation and Evaluation

Our goal is to detect AF episodes in a continuous PPG signal collected from free-living subjects. We defined our task as:

classifying consecutive 30-cycle records between AF and non-AF.

For each 30-cycle record x, our model outputs a binary score $y \in \{0, 1\}$ indicating respectively the absence or presence of AF.

To evaluate the success of our model we use the $F1$ metric. The $F1$ score is the harmonic mean of precision and recall, providing a balanced evaluation of the model's performance by considering both false positives and false negatives simultaneously.

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$$

Naturally we which to maximize the $F1$ score on the test set prediction results, because is expresses the generalization ability of our model which is highly important in the context of medical condition detection.

Also, to further evaluate the generalization ability of our models we will fit and test them across multiple splits of train subjects and test **subjects**. The variance between different splits is an indication for the model robustness.

To match this task, we need to further split our data to consecutive 30-cycle-long signals. It is possible due to the fiducial points detection which determined the onset of each cycle.

First, for the test and validation sets, we simply split each full signal to consecutive 30-cycle-long sub-signals. For the training set, instead of simply splitting it to sub-signals, we wanted to augment our data and sample B 30-cycle-long sub-signals where each has a random starting point. We chose B such that there is no too much overlap in the dataset ($\sim$100) (it might increase the training time without a significant contribution to the overall performance).

Over all the number of samples (30-cycle-long signal) in each split is 2600, $\sim$120, $\sim$320 for the train, validation and test sets respectively (for the validation and test sets the number has a $\sim$ sign because over different train/test subject splitting configurations these numbers changed a bit).

2.4. Feature Extraction

In this work we focus on the feature extraction framework, in which the models don't get the raw signal as input. For each sample we will extract two sets of features, morphological features and hrv (heart rate variability) features. The hrv features are extracted in a full signal level and the morphological features are extracted in a single cycle level. The motivation to combine these two sets of features comes from our hypothesis that combining both 'local' and 'global' features can lead to more informative representations.

2.4.1. HRV features

HRV features are used to analyze and measure the variations in time intervals between successive heartbeats (and in our case we applied the same tools to intervals between successive PPG peaks). **The HRV features are extracted in the full sample level** (one features vector for a single sample) The feature set includes various types grouped into different domains. The time domain features involve statistical measures such as mean and standard deviation of the PP-intervals (interval between consecutive PPG peaks), differences between adjacent PP-intervals, and the square root of the mean of squared differences between adjacent NN-intervals (normalized PP intervals). The frequency domain features analyze the power spectrum of HRV and provide information about different frequency bands. These include total power density spectral, variance in HRV in the very low frequency, low frequency, and high frequency ranges. Poincare plot features involve the analysis of the scatter plot of consecutive NN-intervals. The full feature list appears in B.

2.4.2. Morphological features

Morphological features in PPG data refer to characteristics related to the shape, amplitude, and duration of the PPG waveform. These features capture specific patterns and abnormalities in the PPG signal that can indicate the presence of AF. **The morphological features are extracted in a single cycle level** (30 feature vectors will be extracted from a single 30-cycle-long signal). Some of the features we use are: the time between left onset and the systolic, Systolic width, Diastolic width, sum of Systolic and Diastolic width, and the area under the curve. The full feature list appears in C.

3. Models

3.1. Baseline: Classical ML Approach

We employed a conventional machine learning (ML) pipeline as the baseline approach. In this pipeline, each individual sample was associated with a distinct feature vector, which was subsequently utilized as input for a classifier that was trained on our training dataset. As previously mentioned, a sample was represented by both a hrv feature vector and a morphological feature matrix. Notably, the morphological features were extracted at the level of individual cycles, resulting in the acquisition of 30 feature vectors for a single sample (one for each cycle).

In all our baseline models, we computed the average of the morphological feature vectors across all cycles, yielding a consolidated vector. This averaged morphological feature vector was then concatenated with the HRV feature vector to form a final composite vector that effectively characterizes the given sample.

For classification, we focused on 3 types of classifiers: KNN, Random Forest and regularized SVM with gaussian kernel. By focusing on these three types of classifiers, we aimed to provide a comprehensive analysis and comparison of different classical ML approaches.

3.2. LSTM

The utilization of Long Short-Term Memory (LSTM) networks in this context is driven by the need to effectively handle the morphological feature set. Each sample consists of a consecutive sequence of 30 cycles, with each cycle being mapped to a feature vector. Considering this feature set as a time series calls for the employment of LSTM, which is specifically designed to capture both long-term and short-term dependencies within the data.

In contrast to the baseline approach, where morphological feature vectors were averaged into a single representation for each sample, we adopt a distinct LSTM model to process these 30 vectors individually. Subsequently, we extract the last hidden state vectors from the LSTM as the morphological feature vector for the corresponding sample. The last hidden state of the LSTM is recognized to encapsulate information about the entire series and is commonly employed as a representative vector characterizing the entire series.

In addition to the aforementioned approach, we further explore the use of Bidirectional LSTM (Bi-LSTM) with multiple layers to enhance the modeling of temporal dependencies in the morphological feature extraction process. Bi-LSTM is a variant of LSTM that processes the

input sequence in both forward and backward directions, capturing information from past and future contexts simultaneously. By employing multiple layers of Bi-LSTM, we aim to capture hierarchical representations of the morphological feature sequences.

Following this, the obtained morphological feature vector from the LSTM and the HRV feature vector are concatenated to form a unified feature vector. This combined vector is then inputted into a fully connected classifier with non-linear activation functions. The training objective of this model is the binary cross-entropy loss, which is a standard loss function for binary classification tasks.

4. Results

We evaluate our models w.r.t three possible feature set configurations: *only hrv, only hrv, morph & hrv*. In the *only hrv* configuration each sample is represented only using the hrv feature vector (note that for this configuration the LSTM model we use is equivalent to a MLP classifier because there is no use for the morphological features). In the *only hrv* configuration each sample is represented only using the morphological feature vector, and in the *only hrv* a concatenated vector of both representations is used.

4.1. Baseline

In the present analysis, noticeable variations in the test performance among diverse feature set configurations are evident (refer to 2). Specifically, when solely considering hrv features, the test results exhibit proximity to 1 for both the KNN and random forest algorithms. However, when incorporating morphological and hrv features together, the test results deteriorate, accompanied by a higher degree of variance. Moreover, the utilization of exclusively morphological features yields the poorest outcomes, with all models achieving an approximate value of 0.75. These results are substantially inferior to the state-of-the-art outcomes achieved in this specific task, as documented in [2].

4.2. LSTM

Here 3 the training results demonstrate a comparable pattern for both *only hrv* configuration, as well as for the *only hrv* configuration. In both cases, the results converge towards a value of 1. However, a significant decline in performance is observed when evaluating these models on the test set (-0.2 for *only hrv* and 0.08 for *only hrv*). Specifically, the *only hrv* configuration exhibits pronounced overfitting. Among all the configurations considered, it is noteworthy that the *only hrv* configuration manifests the best performance on the test set, attaining an approximate value of 0.95.

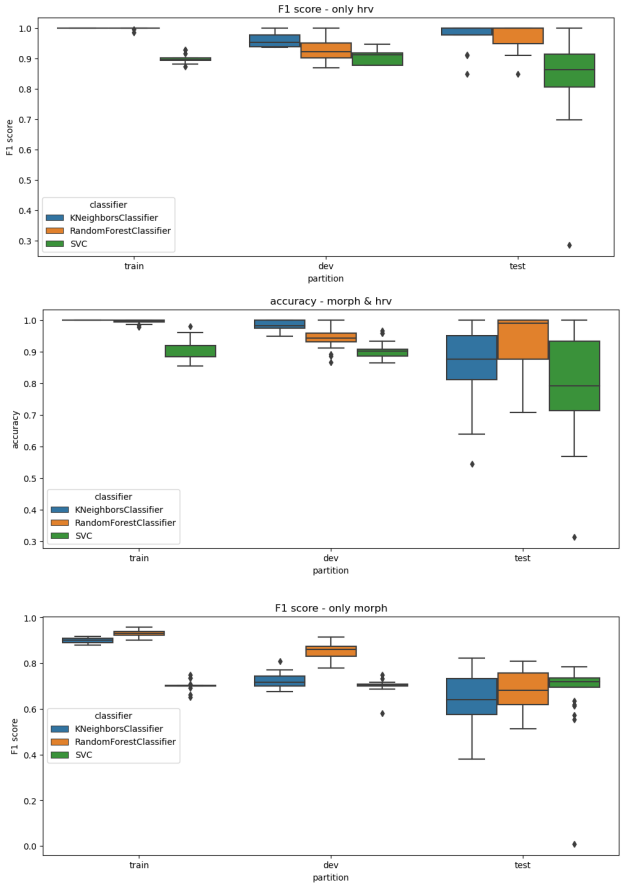

Figure 2. Baseline classical ML models results. The boxplots are taken over multiple train/test subject splits

Notably, the LSTM model yields notably improved outcomes when solely utilizing the morphological configuration. Conversely, the results obtained from the sole hrv configuration exhibit a similar performance level, while the morphological and hrv configuration yields inferior results in comparison.

5. Discussion

The results obtained from our evaluation of different feature set configurations for AF detection using PPG data reveal valuable insights into the performance and effectiveness of the models. By comparing the baseline performance with the LSTM-based approach, we can identify key considerations for improving AF detection accuracy.

In the baseline analysis (refer to 4), the evaluation of diverse feature set configurations demonstrates substantial variations in test performance. Specifically, when considering only the hrv features, both the KNN and random forest algorithms exhibit promising results, with a proximity to 1. However, when incorporating morphological features

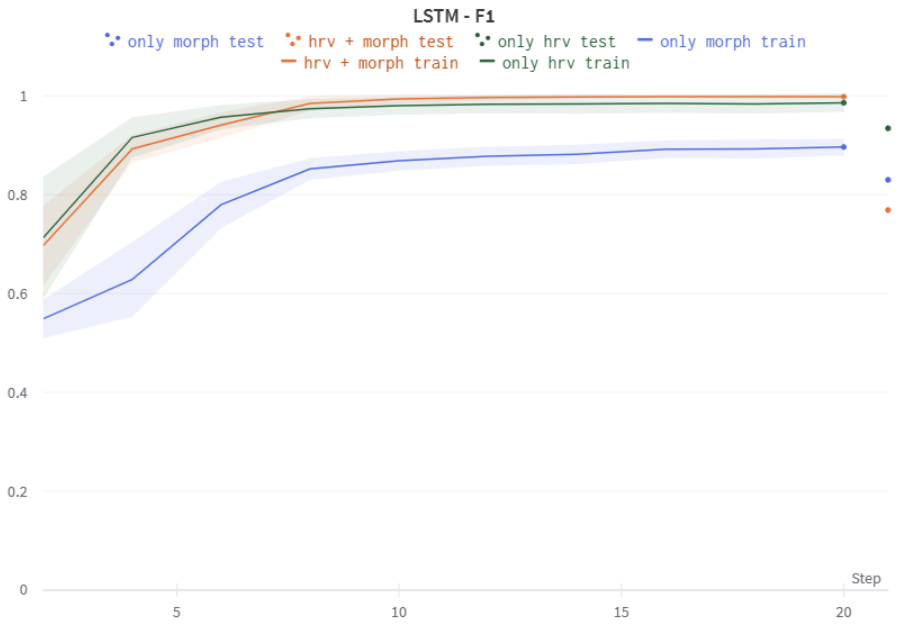

Figure 3. LSTM $F1$ score for all 3 feature set configurations 4 The margin is taken over multiple train/test subject splits. We included only the train and test results, the validation results are similar to the train results for all feature set configurations.

alongside hrv, the test results deteriorate, accompanied by a higher degree of variance.

Interestingly, the LSTM model yields notably improved outcomes when exclusively utilizing the morphological feature set configuration. This finding suggests that extracting cycle level features and leveraging their serial structure can lead to more effective and informative full signal representations. The results obtained from the *only hrv* configuration exhibit a high performance level, suggesting that the hrv features alone can contribute significantly to AF detection. However, the *only hrv* configuration yields inferior results in comparison, indicating that combining both feature sets may introduce complexities that hinder the model's ability to accurately identify AF.

References

[1] M. Aboy et al. "An automatic beat detection algorithm for pressure signals". In: *IEEE Transactions on Biomedical Engineering* 52.10 (2005), pp. 1662–1670. DOI: 10.1109/TBME.2005.855725.

[2] Kirstin Aschbacher et al. "Atrial fibrillation detection from raw photoplethysmography waveforms: A deep learning application". In: *Heart Rhythm O2* 1 (2020), pp. 3–9. ISSN: 2666-5018. DOI: https://doi.org/10.1016/j.hroo.2020.02.002. URL: https://www.sciencedirect.com/science/article/pii/S2666501820300040.

[3] Syed Khairul Bashar. "Atrial Fibrillation annotations of electrocardiogram from MIMIC III matched subset". In: (Apr. 2020). DOI: 10.6084/m9.figshare.12149091.v1. URL: https://figshare.com/articles/dataset/Atrial_Fibrillation_annotations_of_electrocardiogram_from_MIMIC_III_matched_subset/12149091.

[4] Syed Khairul Bashar et al. "Noise Detection in Electrocardiogram Signals for Intensive Care Unit Patients". en. In: *IEEE Access* 7 (July 2019), pp. 88357–88368.

[5] Alan S Go et al. "Heart disease and stroke statistics–2014 update: a report from the American Heart Association". en. In: *Circulation* 129.3 (Dec. 2013), e28–e292.

[6] Marton Aron GODA. *pyPPG toolbox*. URL: https://github.com/godamartonaron/GODA_pyPPG/tree/main.

Appendix

A. Code

All the code used for this work is located in this GitHub repository. Contact this address for any questions or bug reports

B. HRV feature list

Note that a lot of these features use the RR interval which is associated with ECG data. We instead plug in the PP interval (interval between two consecutive PPG peaks)

B.1. Time domain features

- mean_nni: The mean of PP-intervals.
- sdnn : The standard deviation of the time interval between successive normal heart beats (i.e. the PP-intervals).
- sdsd: The standard deviation of differences between adjacent PP-intervals
- rmssd: The square root of the mean of the sum of the squares of differences between adjacent NN-intervals. Reflects high frequency (fast or parasympathetic) influences on hrV (i.e., those influencing larger changes from one beat to the next).
- median_nni: Median Absolute values of the successive differences between the PP-intervals.
- nni_50: Number of interval differences of successive PP-intervals greater than 50 ms.
- pnni_50: The proportion derived by dividing nni_50 (The number of interval differences of successive PP-intervals greater than 50 ms) by the total number of PP-intervals.
- nni_20: Number of interval differences of successive PP-intervals greater than 20 ms.
- pnni_20: The proportion derived by dividing nni_20 (The number of interval differences of successive PP-intervals greater than 20 ms) by the total number of PP-intervals.
- range_nni: difference between the maximum and minimum nn_interval.
- cvsd: Coefficient of variation of successive differences equal to the rmssd divided by mean_nni.
- cvnni: Coefficient of variation equal to the ratio of sdnn divided by mean_nni.
- mean_hr: The mean Heart Rate.
- max_hr: Max heart rate.
- min_hr: Min heart rate.
- std_hr: Standard deviation of heart rate.

B.2. Frequency domain features

- total_power : Total power density spectral
- vlf : variance (= power) in HRV in the Very low Frequency (.003 to .04 Hz by default). Reflect an intrinsic rhythm produced by the heart which is modulated primarily by sympathetic activity.
- lf : variance (= power) in HRV in the low Frequency (.04 to .15 Hz). Reflects a mixture of sympathetic and parasympathetic activity, but in long-term recordings, it reflects sympathetic activity and can be reduced by the beta-adrenergic antagonist propanolol.
- hf: variance (= power) in HRV in the High Frequency (.15 to .40 Hz by default). Reflects fast changes in beat-to-beat variability due to parasympathetic (vagal) activity. Sometimes called the respiratory band because it coPPesponds to HRV changes related to the respiratory cycle and can be increased by slow, deep breathing (about 6 or 7 breaths per minute) and decreased by anticholinergic drugs or vagal blockade.
- lf_hf_ratio : lf/hf ratio is sometimes used by some investigators as a quantitative miPPor of the sympatho/vagal balance.
- lfnu : normalized lf power.
- hfnu : normalized hf power.

B.3. Poincare plot features

- sd1 : The standard deviation of projection of the Poincare plot on the line perpendicular to the line of identity.
- sd2 : SD2 is defined as the standard deviation of the projection of the Poincare plot on the line of identity (y=x).
- ratio_sd2_sd1 : Ratio between SD2 and SD1.

C. Morphological feature list

- CP: Cardiac Period, the time between two consecutive systolic peaks
- SUT: Systolic Upslope Time, the time between left onset and the systolic
- DT: Diastolic Time, the time between the systolic peak and right onset
- SW10: Systolic Width, width at 10% of the pulse height from systolic part
- SW25: Systolic Width, width at 25% of the pulse height from systolic part
- SW33: Systolic Width, width at 33% of the pulse height from systolic part
- SW50: Systolic Width, width at 50% of the pulse height from systolic part
- SW66: Systolic Width, width at 66% of the pulse height from systolic part
- SW75: Systolic Width, width at 75% of the pulse height from systolic part
- SW90: Systolic Width, width at 90% of the pulse height from systolic part

- DW10: Diastolic Width, width at 10% of the pulse height from diastolic part
- DW25: Diastolic Width, width at 25% of the pulse height from diastolic part
- DW33: Diastolic Width, width at 33% of the pulse height from diastolic part
- DW50: Diastolic Width, width at 50% of the pulse height from diastolic part
- DW66: Diastolic Width, width at 66% of the pulse height from diastolic part
- DW75: Diastolic Width, width at 75% of the pulse height from diastolic part
- DW90: Diastolic Width, width at 90% of the pulse height from diastolic part
- SW10+DW10: Sum of Systolic and Diastolic Width at 10% width
- SW25+DW25: Sum of Systolic and Diastolic Width at 25% width
- SW33+DW33: Sum of Systolic and Diastolic Width at 33% width
- SW50+DW50: Sum of Systolic and Diastolic Width at 50% width
- SW66+DW66: Sum of Systolic and Diastolic Width at 66% width
- SW75+DW75: Sum of Systolic and Diastolic Width at 75% width
- SW90+DW90: Sum of Systolic and Diastolic Width at 90% width
- STT: Slope Transit Time, which based on geometrical considerations of the PPG pulse wave to account for simultaneous
- AUCPPG: The area under the curve, a good indicator of change in vascular
- PIR: PPG Intensity Ratio, the ratio of Systolic Peak intensity and PPG valley intensity, reflects on the arterial diameter changes during one cardiac cycle from systole to diastole
- SA: Systolic Peak Amplitude
- SPT: Systolic Peak Time
- tpi: The time between the two onsets of the PPG systolic peak
- SOC: Systolic Peak Output Curve, Ratio between SUT and SA