Deep Learning Classifier for Advancing Video Monitoring of Atrial Fibrillation

Kamil Bukum, Celal Savur and Gill R. Tsouri
Rochester Institute of Technology,
Rochester, NY, USA
{kb4372, cs1323, grteee}@rit.edu

Abstract

Video-based non-contact monitoring of cardiac conditions offers an attractive alternative to contact-based monitoring using sensors attached to the skin. Specifically, video monitoring can significantly improve the monitoring of atrial fibrillation; a prevalent and growing cardiac disease affecting millions around the world. We propose and investigate the performance of a deep learning classifier for the detection of atrial fibrillation. We compare the performance of the proposed classifier with a benchmark of five existing classifiers based on traditional signal processing and machine learning. In addition, we compare performance across various sensing modalities, including a high-end camera, a webcam, an earlobe oximeter, and an electrocardiogram holter. To this end, we conduct a clinical study with 55 atrial fibrillation patients in a hospital setting. Results show that the proposed classifier outperforms the benchmark, especially when using a low-cost webcam, and provides consistently accurate detection when applied to an electrocardiogram, a photo plethysmography sensor, and two video camera sensors, thereby placing video monitoring on par with its contract-based counterparts.

1. Introduction

Atrial Fibrillation (AF) is a prevalent and spreading heart rhythm disorder, characterized by rapid and irregular atrial activation [12, 17]. Structural heart diseases such as rheumatic heart disease, hypertension, and heart failure are important risk factors for the development of AF. Furthermore, AF can cause systemic embolism, hemodynamic disorder, tachycardia-related myopathy as well as symptoms such as heart palpitations, lightheadedness, extreme fatigue, shortness of breath, and chest pain that reduce the patient's quality of life [1,5,41]. The Global Burden of Disease Study 2017 showed that 37.57 million prevalent cases and 3.05 million incident cases were caused by AF globally followed by 287,241 death cases in 2017 [10].

Sensors capturing cardiac signals present the most important tools for medical diagnosis and detection of AF. The gold standard in cardiac sensing is the Electrocardiogram (ECG). ECG is widely used to analyze heart functionality and detect heart disease [19]. It captures electrical signals coming from the heart that provide information on cardiac activity, but requires accurate placement of multiple electrodes on the skin [3].

An alternative to ECG sensing is Photoplethysmography (PPG) [25]. PPG is based on an optical sensor that captures the pulsatile flow of blood propagating through the body [7,27,33,38].It provides coarse information on cardiac activity compared to ECG [35]. However, a PPG sensor can be easily placed on the patient's finger or earlobe [25].

Over the past decade, Video plethysmography (VPG) [13, 18, 21, 39, 40], aka remote-PPG (rPPG), was developed to measure cardiac activity using a camera. VPG is similar to PPG, since it captures the pulsatile signal to and from the subject's face. It typically provides a weaker signal compared to PPG but doesn't require contact with the patient's skin and can be implemented using commercially available cameras.

In recent years, we experience an accelerating proliferation of cameras in our daily life, e.g., security cameras, front facing cameras on smartphones, embedded cameras in laptops and webcams. VPG offers the potential to significantly improve the treatment of AF patients, by turning every camera we encounter to a cardiac monitoring device, thereby providing more frequent measurements and extending monitoring beyond healthcare facilities and into our homes.

Traditional automated detection of AF is typically based on assessing Heart Rate Variability (HRV) measures [6,32]. Commonly used HRV features include Root Mean Square of Successive Differences (RMSSD), Standard Deviation of NN-intervals (SDNN) and The proportion of the number of pairs of successive beat to beat intervals that differ by more than 50 milliseconds (pNN50) [4, 8, 9, 36]. A transitional binary classifier uses a thresholding approach to infer AF. For example, if RMSSD is higher than 100 milliseconds, the subject could be in AF. We collectively refer to such classifiers as Binary Threshold Classifiers (BTC). More recently, classifiers were developed based on machine learning techniques, where data is used to train a classification structure to infer the presence of AF. See [13, 20] and references therein for examples.

In the past, classification performance depended heavily on the sensing modality being used, where ECG-based classifiers provided the best performance [20, 23, 26, 29], followed by PPG and VPG based classifiers. This resulted in sensor-dependent performance and a justified tendency to rely solely on ECG for AF detection. However, monitoring patients using ECG is limiting, since it requires a professional to place the electrodes and the patient to carry the device on the body [20]. For these reasons, ECG monitoring is typically performed infrequently in a healthcare facility. Occasionally, a patient would be given a holter or ECG patch to be carried for a period of 1 day to 2 weeks at the most. There are ECG sensors providing reliable monitoring when performed by the patient, such as the KardiaMobile ECG sensor (AliveCorr, Mountain View, CA), but they are expensive and require maintenance. It is clear that it would be beneficial to have a classifier that provides accurate AF detection across the ECG, PPG and VPG sensing modalities. Such a classifier would enable extended monitoring beyond ECG in healthcare facilities. For example, one could supplement infrequent ECG monitoring at a hospital with less cumbersome PPG monitoring during routine checkups and augment them much further using effortless noncontact VPG monitoring using cameras at home.

In this contribution, we propose, train and test a deep learning classifier designed to provide sensor-agnostic and consistently reliable AF detection for ECG, PPG and VPG sensors. We compare performance of the classifier to a set of 5 benchmark classifiers: 3 traditional classifiers and 2 machine learning classifiers. To this end, we use data collected in a large clinical study performed in a hospital with 55 patients diagnosed with AF before and after undergoing a cardioversion procedure. The collected data consist of synchronized capture from ECG electrodes, PPG earlobe sensor, VPG using a simple webcam and VPG using a high-end camera. Our results show that the proposed classifier outperforms the benchmark classifiers, especially when using a low cost webcam, and also provides consistently accurate detection across all sensors. This work could help advance the acceptance of video monitoring as an alternative to contact based ECG and PPG monitoring.

2. Proposed Deep Learning Classifier

Fig. 1 presents a flow diagram of the proposed Deep Learning Approach based on a Convolutional Neural Network (CNN). The raw data consists of one of the following sensors' 25 seconds output: ECG, PPG, VPG 180Hz and VPG 30Hz signals. The raw data is split to overlapping 25 sec intervals. The number of samples per sensor per interval depends on that sensor's sampling frequency. Namely, 4500, 7500, 4500 and 750 data points for ECG, PPG, VPG 180Hz and VPG 30Hz respectively. Therefore, we apply different signal preprocessing per sensor as is explained next. The preprocessing output is used as input to the network. This can be seen as a passive "Input Layer" receiving 1D data. following, a CNN-1D layer creates a convolution kernel that is convolved with the layer input over a single spatial (or temporal) dimension to produce a tensor of outputs as was described in [19]. A Dropout layer is then used to avoid overfitting, and a Max-Pooling layer is used to calculate the maximum value for each CNN-1D layer's output to create new input for the next layer. This CNN-1D with Dropout and Max-Pooling structure is repeated once more. The result is flattened from 2D data to 1D data by using a Flatten Layer [34] [19]. One last dropout layer is then applied to avoid overfitting. The flattened output is processed by a Fully Connected Layer (FCL) with a dropout layer whose output is the binary AF classification decision following binary cross entropy loss function. See Table 1 for the detailed structure description.

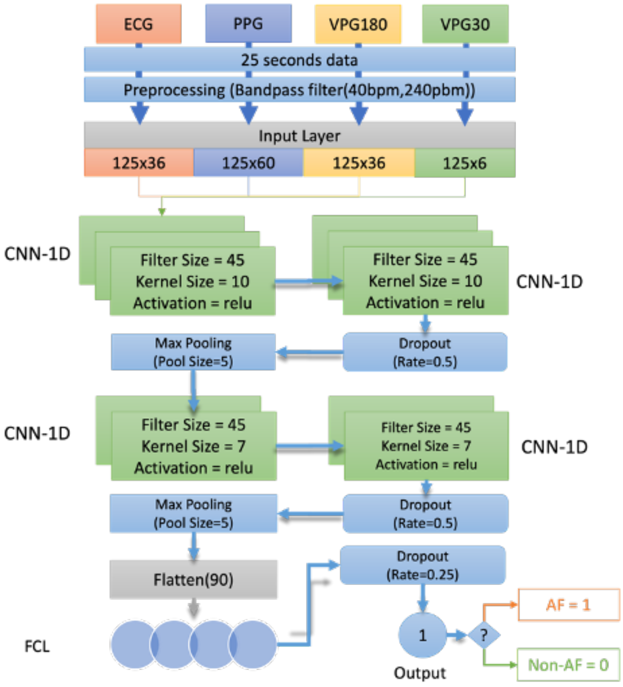

Figure 1. Proposed classifier's structure

2.1. Preprocessing and Feature Extraction

To improve performance of the proposed classifier, we initially filter the signals captured by each sensor to the range of frequencies where cardiac activity is expressed. Namely, [0.67 4] Hz representing a pulse rate range of [40 240] beats per minute. To this end, we filter the sensor's

Table 1. Summary of configuration parameters.

Component	**Description**
25 seconds input data block	Consists of 4500(ECG), 7500(PPG), 4500(VPG 180Hz), or 750(VPG 30Hz) data points
Preprocessing	Moving average with window size(fs/2) applied to the data for PPG, VPG 180Hz, VPG 30 Hz followed by a 4th order Butterworth filter with bandpass of [0.67,4]Hz and normalization to 0-1 range.
Input Layer	Dimension of the input layer is 125x36, 125x60, 125x36, or 125x6 depending on data type, ECG, PPG, VPG 180Hz, VPG 30Hz respectively.
CNN-1D	The filters size is 45, kernel Size is 10, kernel initializer is ‘random_normal’, padding is ‘same’, and the activation function is ‘relu’
CNN-1D	The Filters size is 45, kernel Size is 10, padding is ‘same’, and activation function is ‘relu’.
Dropout	The dropout rate is 0.5
MaxPooling	The pool size is 5
CNN-1D	The filters size is 45, kernel Size is 7, padding is ‘same’, and activation function is ‘relu’
CNN-1D	The filters size is 45, kernel Size is 7, padding is ‘same’ , and activation function is ‘relu’
Dropout	Dropout rate is 0.5
MaxPooling	Pool size is 10
Flatten	Flatten is a function used to transform 2D data to 1D.
Fully Connected Layer	45 units with activation function ‘relu’
Dropout	Dropout rate is 0.25
Output	The Adam algorithm is used for optimization with parameters ’learning learning_rate=0.0001’, ‘beta_1=0.9’, ‘beta_2=0.999’, ‘amsgrad=False’. ‘binary_crossentropy’ loss function is used to binarize the probability to 0 or 1. Metrics are found by using accuracy
Training	Batch size set to 30 and 300 epochs are used for training with cross-validation data.

output signal using a 4th order Butterworth bandpass filter designed to match each sensor’s sampling rate. As is commonly done, the filtered signals are all normalized to have no units and reside within a range of [0 1].

For ECG, no further preprocessing is performed. For PPG, VPG-30Hz and VPG-180Hz, prior to bandpass filtering, a moving average filter with length of the sampling frequency divided by 2 is applied to smooth the signal.

Table 2. Parameters of 2-D conversion for all sensors.

Signal Type	**Sec.**	**fs**	**Length**	**CNN Tensor (2D)**
ECG	25	180	4500	(sec) x (fs/5) = 125 x 36
PPG	25	300	7500	(sec) x (fs/5) = 125 x 60
VPG (180 hz.)	25	180	4500	(sec) x (fs/5) = 125 x 36
VPG (30 hz.)	25	30	750	(sec) x (fs/5) = 125 x 6

Each filtered signal is then converted to a 2-D tensor input to the CNN as depicted in Fig. 1. Each processed signal encompasses 25 seconds and each sensor has its own sampling frequency (fs), the corresponding conversions to a 2-D tensor are performed based on the parameters in Table 2.

2.2. Classification

Following training of the CNN classifier using the features extracted from data, its output is a probability for the measurement to be labelled as AF. This output is compared to a predefined threshold. When the probability is above the threshold, it implies irregular rhythm and the measurement is classified as AF. Conversely, when the probability is below the threshold, it implies Sinus Rhythm (SR) and the measurement is classified as SR.

3. Benchmark Classifiers

Most common arrhythmia classifiers are based on extracting features from the VPG/PPG IBI’s or the ECG RR-intervals. IBI/RR detection are based on peak detection algorithms designed to identify local peaks of the the signal. It is common practice to perform preprocessing to clean the signals prior to peak detection in order to improve IBI/RR estimation. Preprocessing depends on the signal at hand.

3.1. ECG Preprocessing

We start by applying zero-score normalization to the signal. We then equalize the filtered output by replacing ECG samples that reside within 2-4 standard deviations from the mean ECG signal with their square root values. This reduces the scaling difference across QRS complexes. Following equalization, we apply the Pan Tompkins algorithm for filtering and identifying the ECG QRS complexes [31]. Extracting RR interval boils down to finding the time difference between consecutive QRS complexes.

3.2. PPG and VPG Preprocessing

We start by applying the same 4th order Butterworth bandpass filter described above. We then apply a moving average window with length of 50, 30, 5, to PPG, VPG-180Hz and VPG-30Hz respectively to account for their different sampling frequencies. The outputs of the moving average filter are further detrended by applying 15th order polynomial detrending, as was commonly done in previous work [18, 21, 24, 40]. The detrended signal is then normalized using zero-score normalization. Peaks are identified in the normalized signal using adaptive maxima identification with threshold of 0.2 and distances of 75, 45 and 7.5 for PPG, VPG-180Hz and VPG-30Hz respectively. The functions used to implement processing can be found in [30].

Extracting IBIs is done by finding the time difference between consecutive peaks. IBI’s that are lower than a time corresponding to a heart beat of 240 beats per minute are removed since a human heart cannot beat this fast making such IBIs the likely result of erroneous peak detection [2, 15].

3.3. Feature Extraction

We use the IBI’s and RR’s to derive the RMSSD, SDNN and PNN50 parameters to be used as features. We further derive the parameters: mean, median (MD), standard deviation (STD) and median absolute deviation (MAD). These parameters are then grouped into categories summarized in Table 3.

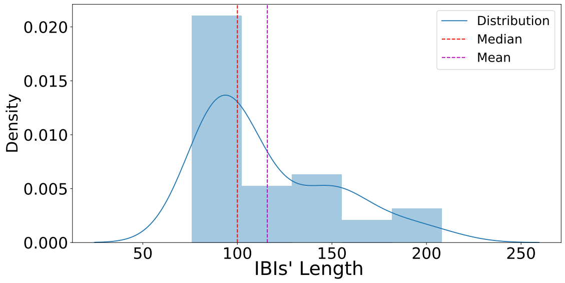

Figure 2. ECG Sample of IBI's Distribution

Table 3. The category of the RR and IBIs for future extraction.

Categories	IBI range	
	Lower Limit	Upper Limit
Median Center (MDC)	MD - 2 * MAD	MD+2*MAD
Median Left (MDL)	MD - 2 * MAD	MD+0.5*MAD
Median Right (MDR)	MD - 0.5*MAD	MD+2*MAD
Mean Center (MEC)	MEAN - 2 * STD	STD + 2 * MAD
Mean Left (MEL)	MEAN-2*STD	STD+ MAD/2
Mean Right (MER)	MEAN – STD/2	STD + 2 * MAD

These categories are helpful in characterizing the heart rate variability and by that arrhythmia while damping the impact of outliers the result from erroneous peak detection. An example of IBI distribution along with the aforementioned categories is depicted in Fig. 2.

Additional features defined from the IBIs and the above categories: Standard Deviation (STD), Interquartile Range (IQR), Singular Value Decomposition Entropy (SVD).

3.4. Classification

We define three benchmark classifiers based on a Binary Threshold Classifier (BTC): BTC-RMSSD, BTC-SDNN and BTC-PNN50, by using the corresponding parameters RMSSD, SDNN and PNN50. Namely, we perform straight forward thresholding of these parameters with a predefined threshold. When the parameter is above the threshold, it implies irregular rhythm and the measurement is classified as AF. Conversely, when the parameter is below the threshold, it implies Sinus Rhythm (SR) and the measurement is classified as SR. Note that there other BTCs based on other derivatives of RRs and IBIs, such as the one in [11], where a thresholding approach was applied to a parameter defined over the spectra, domain.

To represent the family of Machine Learning (ML) classifiers we define a fourth benchmark classifier based on a Support Vector Machine (SVM) classifier: SVM-Classic, by training an SVM classifier using the three parameters: RMSSD, SDNN, and PNN50 as features. Classification is then performed by comparing the probability output of a measurement to a threshold as explained above. Note that other ML classifiers could have been used as well, such as Random Forest.

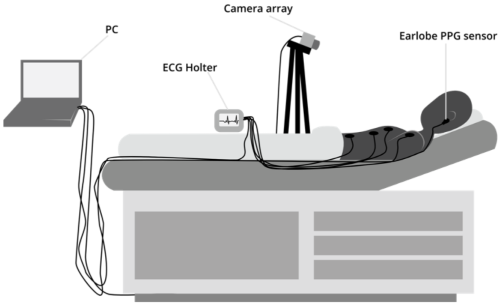

Figure 3. Schematic description of experimental setup.

We further define a fifth benchmark classifier based on handcrafted features: SVM-Handcrafted, by training an SVM classifier using RMSSD, SDNN, PNN50, STD, IQR and SVD as features. Classification is then performed similarly to SVM-Classic.

4. Clinical Study

Fifty-five patients diagnosed with AF and scheduled for cardioversion procedure were enrolled in a clinical study approved by the Internal Review Committees for Protecting Human Subjects at both the University of Rochester Medical Center (URMC, Rochester, NY) and the Rochester Institute of Technology (RIT, Rochester, NY).

A sensors-synchronized measurement of 5 minutes and 30 seconds of ECG, PPG and VPG sensing data was performed before and after cardioversion. The measurement setup is depicted in Fig. 3. The experiment was performed in one of the URMC hospital's standard patient rooms illuminated by fluorescent ceiling lights. The ECG signals were recorded using 10 ECG electrodes placed on the subject's body by a trained nurse. The PPG signals were recorded using an earlobe oximeter sensor placed on the left ear of the subject. Two cameras (high-quality and low-quality) were placed in front of the subject approximately 1 meter away from the subject's head.

An H12+ (Mortara Instruments, Milwaukee, MN) ECG Holter was used to capture the ECG signals with a 180Hz sampling rate. A HeartSensor HRS-07UE PPG sensor (Binar, Poulsbo, WA) was used to obtain the PPG signal with a 300Hz sampling rate. A Logitech Quickcam Pro 9000 (simple webcam, relative low-quality) camera was used to capture a video with a 30Hz sampling rate, and a Basler ACE 1920-155uc camera (relative high-quality) was used to capture a second video with a 180Hz sampling rate [16]. In a typical VPG application, face detection is used to define a Region of Interest (RoI) from which to extract the cardiac signal. In our study, the subjects were still and in a supine position, so the RoI was manually defined and did not change during the recoding.

All subjects were diagnosed with AF immediately prior to cardioversion. This means that the first synchronized measurement represents AF data. Except for 4 subjects, all subjects reverted from AF to SR after cardioversion. This means that the second synchronized measurement from subjects that reverted to SR represents non-AF data. Data from another 9 subjects was found to be corrupted (ECG electrode missplacement, PPG sensor malfunction, etc.). The data captured from those 9 subjects were discarded.

Note that subjects were asked to be still and were sedated before cardioversion, therefore the captured data was largely not corrupted by motion and changing ambient lighting conditions (shadowing, flickering, etc.). An example of prepossessed data captured from a subject in the study before and after cardioversion is presented in Fig. 4.

While AF data and SR data per subject are captured in a continuous 5.5 minutes measurements, in practical applications it would be beneficial to detect AF on shorter time intervals. This is because as the interval becomes longer, the subject is more likely to move room and lighting conditions are more likely to change, thereby corrupting signal capture. However, shortening the time interval too much would prevent efficient AF detection because less beats would be available for the classifier to make a decision. We experimented with training data to resolve this tradeoff. We found that a measurement of 25 sec is long enough to detect rhythm irregularities associated with AF. This result is in line with prior published studies, where an interval of 10 sec to 30 sec was used [28].

To generate a large enough dataset per subject in the study and improve training, we converted each 5.5 minutes measurement to approximately 31 measurements of 25 seconds each with a 60% overlap between consecutive measurements (in some cases the recording was slightly lower than 5.5 minutes). This improves training and is common practice in such applications [14]. It means that each measurement is correlated to its immediately preceding and succeeding measurement. Note that overlapping is done per subject and since a subject's data is either exclusively in the training dataset or in the testing dataset, training and testing data are uncorrelated.

5. Comparative Analysis

The total data collected from the 46 subjects with non corrupted data capture is comprised of 1425 25 seconds measurements labelled as AF (before cardioversion) and 1280 25 seconds measurements labelled as SR (after cardioversion). We divided the data to training and testing datasets. The training dataset comprised all data from 36 subjects; 1094 labelled as AF and 1001 labelled as SR. The testing dataset comprised all data from the remaining 6 subjects; 310 labelled as AF and 279 labelled as SR.

Training of the SVM-Classic, SVM-Handcrafted and proposed CNN classifiers was done using the training dataset. Testing of these classifiers was done by applying the trained classifiers to the testing dataset. The output probability per measurement was then used to generate Receiver Operating Characteristics (ROC) curves by varying the classifier's threshold from 0 to 1 with 0.01 increments and calculating the True Positive Rate (TPR) and False Positive Rate (FPR) per threshold. Note that TPR and FPR are similar to Sensitivity and 1-Specificity. ROC curves were also generated for the BTC-RMSSD, BTC-SDNN and BTC-PNN50 benchmark classifier in a similar manner by replacing the output probability with the classifier's underlying parameter (RMSSD, SDNN and PNN50).

Using the ROC curves, we selected the threshold or parameter that maximizes the Accuracy (ACC) of the classifier. We then calculated the following performance parameters for each classifier applied to each sensor: Precision (PR), Recall (RE), F1 Score(F1) and Accuracy (ACC) [22].

6. Results

Fig. 5 presents ROC curves obtained from ECG data. Each curve is drawn by varying the threshold associated with the underlying classifier and evaluating the True Positive Rate (TPR) and the False Positive Rate (FPR), aka Sensitivity and 1-Specificity respectively for each value of the threshold. For all ROC curves, we mention the value of the threshold ("th" on the figure legend) providing the maximum accuracy. This threshold is also marked by a vertical line corresponding to its respective ROC curve.

For all ROC curves, we observe the typical knee-shaped behavior, where the tradeoff between FPR and TPR can be resolved by selecting the appropriate threshold for maximum accuracy. The poorest performance is obtained for the BTC classifiers BTC-SDNN and BTC-RMSSD. However, BTC-PNN50 performs significantly better and is on par with the SVM-Classic classifier. The SVM-Handcrafted classifier offers the best performance within the benchmark approaches, corresponding to a TPR of 1 and FPR of 0.014. This means that all AF instances are detected, while falsely reporting 1.4% of SR instances as AF. The proposed method based on CNN provides the same perfect TPR of 1, but a significantly lower FPR of 0.007. This means that the proposed approach detects all AF instances, while falsely reporting 0.7% of sinus rhythm instances as AF. It follows that the proposed approach provides a 2-fold reduction in false positives compared to the best benchmark classifier.

Fig. 6 represents the ROC curves for the same classification approaches shown in Fig.6 for PPG data. For all PPG ROC curves, we observe typical knee-shaped behavior similar to the ECG data for the threshold achieving maximum accuracy. Similar to ECG data, the poorest performance

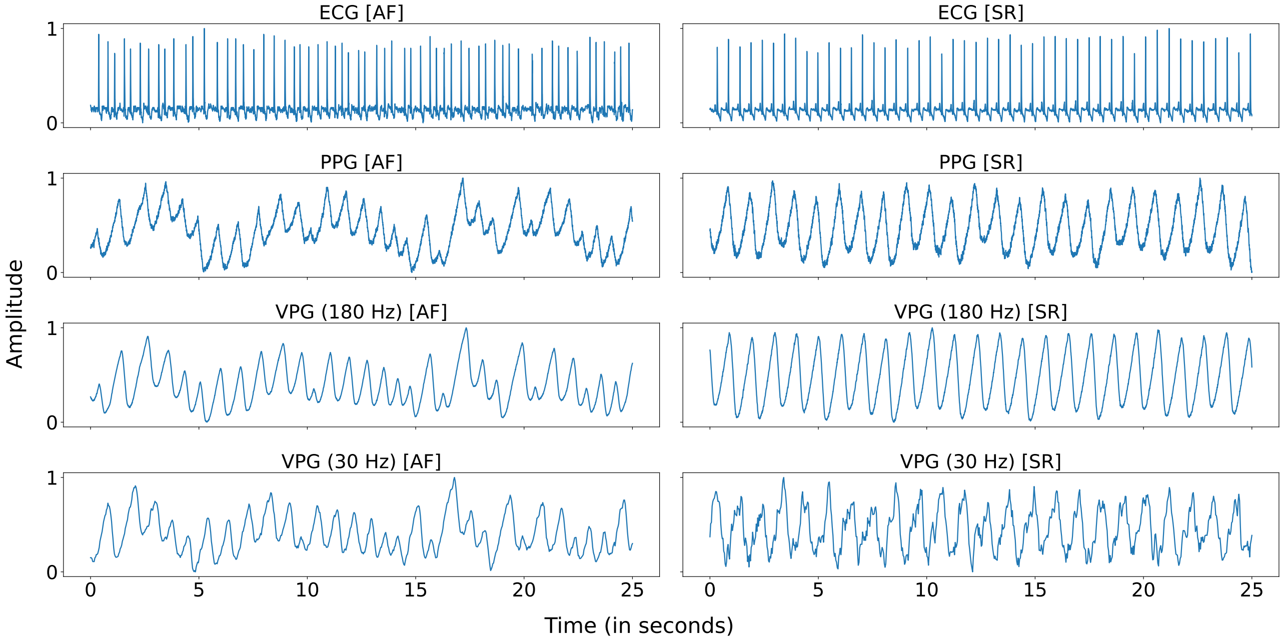

Figure 4. Example of signals captured during study from a subject before cardioversion (left) and after cardioversion (right).

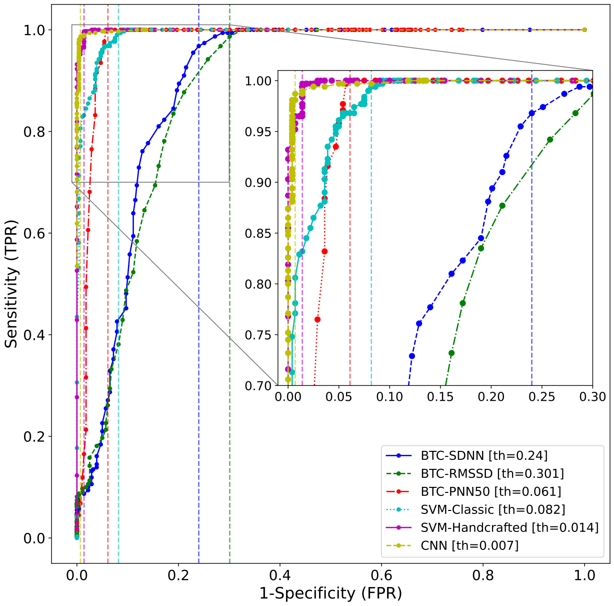

Figure 5. BTC-SDNN, BTC-RMSSD, BTC-PNN50, SVM-Classic, and SVM-Handcrafted ROC curves for ECG.

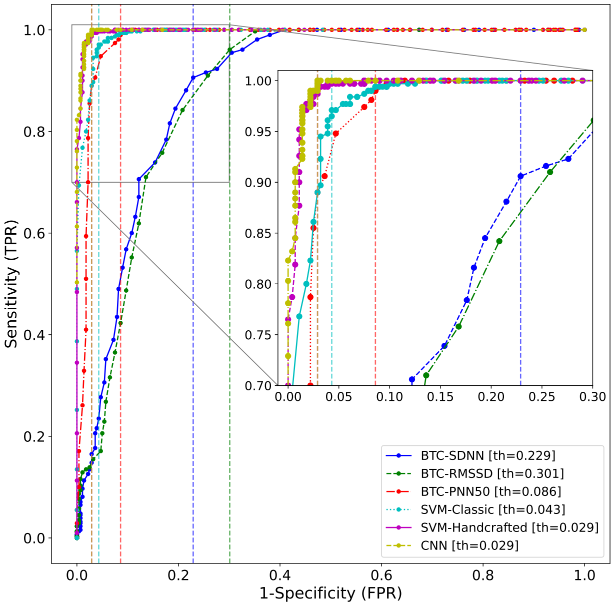

Figure 6. BTC-SDNN, BTC-RMSSD, BTC-PNN50, SVM-Classic, and SVM-Handcrafted ROC curves for PPG.

is obtained for the BTC classifiers based on thresholding of SDNN and RMSSD. PNN50 and SVM-based classifiers show significantly better results. In addition, as for ECG, SVM-Classic performs poorer than the SVM-Handcrafted classifier. Comparing Figs. 6 and 7, it is clear that all classifiers except CNN perform better for ECG compared to PPG. The proposed CNN approach and SVM-Handcrafted performs better than all other classifiers and provides similar high performance for both ECG and PPG. For example, the SVM-Handcrafted classifier falsely reports SR instances as

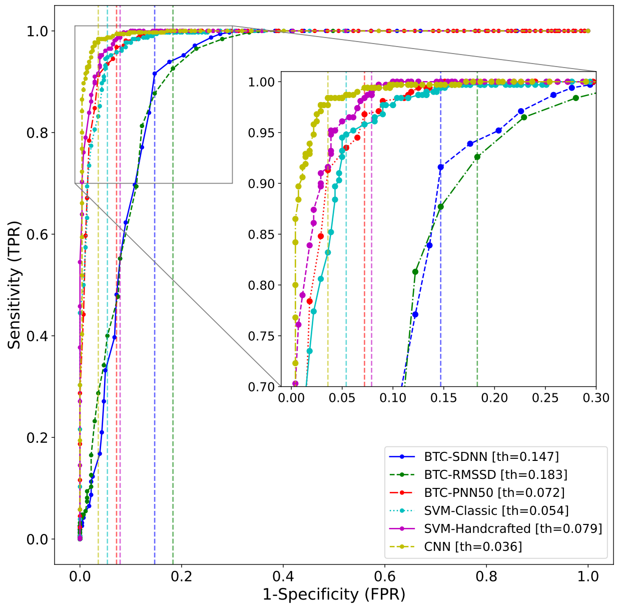

Figure 7. TC-SDNN, BTC-RMSSD, BTC-PNN50, SVM-Classic, and SVM-Handcrafted ROC curves VPG 180Hz.

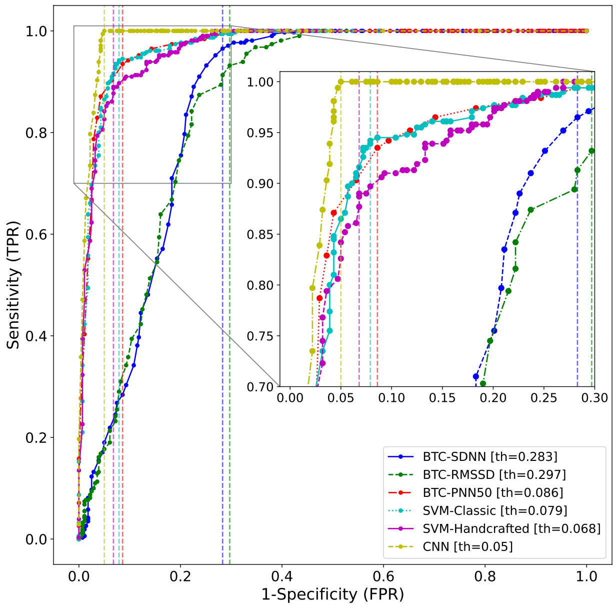

Figure 8. BTC-SDNN, BTC-RMSSD, BTC-PNN50, SVM-Classic, and SVM-Handcrafted ROC curves VPG 30Hz.

AF 1.4% and %2.9 of the time for ECG and PPG respectively for the selected threshold. The proposed CNN approach provides the same results for PPG and a 2 fold improvement for ECG (0.7% of AF instances reported as SR).

Fig. 7 shows the ROC curves for the VPG (180 Hz) data. We observe the same overall trends as in Figs. 6 and 7. However, although BTC-PNN50 performs better than the other BTC classifiers, it performs slightly poorer than the SVM-Classic classifier. We conclude that for VPG (180 Hz) data, the BTC classifiers should be avoided if more sophisticated SVM or the proposed CNN classifier can be implemented. Note that SVM-Handcrafted is the best benchmark classifier. As for ECG and PPG, the proposed CNN approach performs the best. In this case it falsely reports 3.6% of SR instances as AF. This means the proposed approach provides a 2-fold reduction in false positives compared to the best performing alternative in the benchmark.

Fig. 8 shows the ROC curves for VPG (30 Hz). In general, we observe the same trends as in Figs. 6-8. Note that the disparity in performance of the different classifiers is the greatest in this case. BTC-SDNN falsely reports 28.3% of SR instances as AF. BTC-RMSSD and BTC-PNN50 falsely report 29.7% and 8.6% of SR instances as AF respectively. Both provide poorer results than the SVM classifiers. SVM-Handcrafted provides the better result in the benchmark (FPR of 6.8%). Finally, the proposed CNN approach provides better result than benchmark classifiers with 5% classification of SR vs AF instances.

Table 4 presents PR, RE, F1 and ACC for ECG, PPG, VPG (180 Hz) and VPG (30 Hz) for all the aforementioned classifiers. At first glance, we can see that the PR, RE, F1, and ACC are increasing from the top to bottom of Table 4. This shows that using more advanced classification techniques improves performance. Note that using an advanced technique like SVM-Classic has a negative impact on the classification of AF. For example, using the BTC-PNN50 compared to SVM-Classic has a negative effect on the Recall which means that it causes to miss detect some of the AF instances. Also note that Precision is reduced resulting in miss-classification of SR measurements as AF. A possible explanation for this is that BTC-PNN50 performs much better than its BTC counterparts that combining all three via SVM-Classic deteriorates its superior performance.

As expected from the ROC curves, SVM-Handcrafted performs the best out of all benchmark classifiers across all parameters and sensing modalities. We attribute to the extra features used in SVM-Handcrafted. Namely, SVD Entropy and IQR. In addition, we find that CNN performs the same or better in all cases compared to SVM-Handcrafted. Note for example F1 score where Precision and Recall are combined to show CNN is better.

The CNN classifier provides the most consistent and highest performance across all sensors. For instance, the accuracy of detecting AF by applying BTC based on SDNN, RMSSD, and PNN50 is approximately 80%, 82%, and 90% respectively across all sensors. Accuracy of SVM classifiers varies greatly within 5%-10%. While RMSSD resulted in 84% percent accuracy for the ECG data, it resulted in 77%

Table 4. Performance of classifiers

	ECG				PPG				VPG (180 hz)				VPG (30 hz)			
Methodology	**PR.**	**RE.**	**F1**	**ACC**	**PR.**	**RE.**	**F1**	**ACC**	**PR.**	**RE.**	**F1**	**ACC**	**PR.**	**RE.**	**F1**	**ACC**
BTC-SDNN	0.82	0.97	0.89	0.87	0.81	0.91	0.86	0.84	0.87	0.92	0.89	0.89	0.79	0.96	0.87	0.85
BTC-RMSSD	0.78	0.99	0.87	0.85	0.78	0.96	0.86	0.84	0.85	0.93	0.89	0.87	0.78	0.93	0.85	0.82
BTC-PNN50	0.95	1.00	0.97	0.97	0.93	0.99	0.96	0.95	0.94	0.97	0.95	0.95	0.92	0.94	0.93	0.93
SVM-Classic	0.93	0.99	0.96	0.96	0.96	0.97	0.97	0.96	0.95	0.95	0.95	0.95	0.93	0.94	0.94	0.93
SVM-Handcrafted	0.99	1.00	0.99	0.99	0.97	0.99	0.98	0.98	0.93	0.99	0.96	0.96	0.94	0.89	0.91	0.91
CNN	0.99	0.99	0.99	0.99	0.97	1.00	0.99	0.99	0.97	0.98	0.98	0.97	0.96	1.00	0.98	0.98

percent for VPG 180 Hz. We can see that the classical BTC approach does not provide a stable and high accuracy for all data types. For example, the difference in the accuracy between ECG and VPG is around 8% percent. Similarly, note that while SVM-Handcrafted significantly increase the accuracy; 96%, 94%, 94%, and %93 respectively for ECG, PPG, VPG (180 Hz), VPG (30 Hz), CNN yields a more stable result of 98-99% accuracy for all kinds of data.

7. Limitations and Future Research

While the presented results show great promise, it should be noted that the clinical study was performed in a rather controlled environment of a hospital room. For the most part, subjects kept still and the lighting conditions were static. This means that our results represent a real-world scenario of monitoring AF patients admitted to healthcare facilities. It is well known that video monitoring can be impaired when the subject is moving and when the lighting conditions change due to shadowing and flickering lights. These limitations are expected to reduce detection accuracy when monitoring is performed in an uncontrolled environment such as residential areas, offices or while being left unsupervised in a hospital room. Our ongoing research include assessing the proposed classifier in such environments, where we develop algorithms for tracking motion and lighting conditions to circumvent their effect. To this end, we are currently finalizing the collection of data in a second large clinical study involving 250 AF patients being monitored in their homes using personal smart devices.

In our work, the classifier was trained across multiple subjects, implying that one classifier fits all. However, it is expected that training the classifier per subject would result in a more robust and accurate performance. This is true for the benchmark classifiers as well. We were unable to perform such analysis in this work due to the limited data collected per subject. Future work where more data is collected per subject could support such analysis.

Note that in our data collection setup the camera was 1 meter away from the subject. while this setup represents use of personal devices (laptops, tablets, smartphones, office environment) as well as telemedicine applications, it does not address scenarios where the camera is expected to be farther from the subject such as emergency rooms, airports, etc. Future research could address these scenarios.

Our work focused on feature extraction from time domain beat to beat parameters. Additional features can be used to improve classification based on spectral domain. It is also possible to add features relating to motion and lighting conditions when addressing an uncontrolled environment to add robustness to classification. In future research, we plan to expand the feature space to address these issues.

In future work, we would expand the benchmark classifiers to include emerging techniques that extend beyond classical signal processing and ML, such as those described in [37]. In addition, classification generality can be improved further, e.g., using k-fold CV or LOSO validation.

8. Conclusion

In this contribution, we proposed, designed and tested a deep learning classifier for improving non-contact video-based detection of AF. We compared performance of the proposed classifier with a benchmark of 5 existing classifiers using data collected in a clinical study encompassing 55 patients diagnosed with AF before and after receiving cardioversion procedure. Results show that the proposed classifier is equivalent or improves performance against all benchmark classifiers when using any of the major sensors: ECG, PPG and VPG. Most notable is a significant improvement when using a low cost webcam sensor, thereby promoting the use of video monitoring on widely distributed low cost platforms (smartphones, tablets, laptops, etc.). Most importantly, results show that the proposed classifier provides consistently accurate classification across all sensors, bringing VPG on par with ECG and PPG and thereby promoting potential acceptance of video monitoring as a reliable alternative to contact based sensors.

Acknowledgment

The clinical study was funded by Verily Life Sciences. The authors would like to thank Jean Xia and Jean-Philippe Couderc for their help in conducting the study.

References

[1] Atrial fibrillation," centers for disease control and prevention, 27-sep-2021. Available:. 1

[2] Heart rate response to exercise stress testing in asymptomatic women: The st. james women take heart project. *Circulation*, 122:130–137, 7 2010. 3

[3] M. Alghatrif and J. Lindsay. A brief review: history to understand fundamentals of electrocardiography. *Journal of Community Hospital Internal Medicine Perspectives*, 2(1):14383,. 1

[4] A. Amin, A. Houmsse, A. Ishola, J. Tyler, and M. Houmsse. The current approach of atrial fibrillation management. *Avicenna J Med*. 1

[5] W.S. Aronow and M. Banach. Atrial fibrillation: The new epidemic of the ageing world. *Journal of Atrial Fibrillation*, 1(6). 1

[6] D. B. Discriminant analysis of heart rate variability after electrical cardioversion predicts atrial fibrillation recurrence. *International Journal of Clinical Cardiology*, 1(2). 1

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

[8] G.E. Billman. Heart rate variability - a historical perspective. *Frontiers in Physiology*, 2. 1

[9] H. ChuDuc, K. NguyenPhan, and D. NguyenViet. A review of heart rate variability and its applications. *APCBEE Procedia*, 7:80–85,. 1

[10] S.S. Chugh, G.A. Roth, R.F. Gillum, and G.A. Mensah. Global burden of atrial fibrillation in developed and developing nations. *Global Heart*, 9(1):113,. 1

[11] Jean-Philippe Couderc, Survi Kyal, Lalit Mestha, Beilei Xu, Derick Peterson, Xiaojuan Xia, and Burr Hall. Detection of atrial fibrillation using contactless facial video monitoring. *Heart rhythm : the official journal of the Heart Rhythm Society*, 12, 08 2014. 4

[12] H. Dai, Q. Zhang, A.A. Much, E. Maor, A. Segev, R. Beinart, S. Adawi, Y. Lu, N.L. Bragazzi, and J. Wu. Global, regional, and national prevalence, incidence, mortality, and risk factors for atrial fibrillation, 1990–2017: Results from the global burden of disease study 2017. *European Heart Journal - Quality of Care and Clinical Outcomes*, 7(6):574–582,. 1

[13] Cigdem Polat Dautov, Ruslan Dautov, Jean-Philippe Couderc, and Gill R Tsouri. Machine learning approach to detection of atrial fibrillation using high quality facial videos. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4, 2021. 1, 2

[14] A. Dehghani, O. Sarbishei, T. Glatard, and E. Shihab. A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors (Basel, Switzerland*, 19(22):5026. 5

[15] Ronald L Gellish, Brian R Goslin, Ronald E Olson, Audry McDonald, Gary D Russi, and Virinder K Moudgil. Longitudinal modeling of the relationship between age and maximal heart rate. *Medicine and science in sports and exercise*, 39(5):822—829, May 2007. 3

[16] D.M. German, M.M. Kabir, T.A. Dewland, C.A. Henrikson, and L.G. Tereshchenko. Atrial fibrillation predictors: Importance of the electrocardiogram. *Annals of Noninvasive Electrocardiology*, 21(1):20–29,. 4

[17] A.S. Go, E.M. Hylek, K.A. Phillips, Y.C. Chang, L.E. Henault, J.V. Selby, and D.E. Singer. Prevalence of diagnosed atrial fibrillation in adults. *JAMA*, 285(18):2370,. 1

[18] J.H. Guzman, J.-P. Couderc, and G.R. Tsouri. Accurate hemodynamic sensing using video plethysmography with high quality cameras. In *13th International Symposium on Medical Information and Communication Technology (ISMICT*. 1, 3

[19] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, and A.Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69,. 1, 2

[20] S.H. Jambukia, V.K. Dabhi, and H.B. Prajapati. Classification of ecg signals using machine learning techniques: A survey. In *2015 International Conference on Advances in Computer Engineering and Applications*. 2

[21] S. Kyal, L.K. Mestha, B. Xu, and J.-P. Couderc. A method to detect cardiac arrhythmias with a webcam. In *2013 IEEE Signal Processing in Medicine and Biology Symposium (SPMB*. 1, 3

[22] D. Lane. *Introduction to Statistics*. Open textbook library. David Lane, 2003. 5

[23] D. Li, J. Zhang, Q. Zhang, and X. Wei. Classification of ecg signals based on 1d convolution neural network. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom*. 2

[24] S. Liaqat, K. Dashtipour, A. Zahid, K. Assaleh, K. Arshad, and N. Ramzan. Detection of atrial fibrillation using a machine learning approach. *Information*, 11(12):549,. 3

[25] G. Lu, F. Yang, J.A. Taylor, and J.F. Stein. A comparison of photoplethysmography and ecg recording to analyse heart rate variability in healthy subjects. *Journal of Medical Engineering Technology*, 33(8):634–641,. 1

[26] E.J. Luz, W.R. Schwartz, G. Cámara-Chávez, and D. Menotti. Ecg-based heartbeat classification for arrhythmia detection: A survey. *Computer Methods and Programs in Biomedicine*, 127:144–164,. 2

[27] A.V. Moço, S. Stuijk, and G. Haan. New insights into the origin of remote ppg signals in visible light and infrared. *Scientific Reports*, 8(1). 1

[28] M Loretto Munoz, Arie Van Roon, Harriëtte Riese, Chris Thio, Emma Oostenbroek, Iris Westrik, Eco J C De Geus, Ron Gansevoort, Joop Lefrandt, Ilja M Nolte, and Harold Snieder. Validity of (ultra-)short recordings for heart rate variability measurements. 2015. 5

[29] K. Muthuvel, S.H. Veni, L.P. Suresh, and K.B. Kannan. Ecg signal feature extraction and classification using harr wavelet transform and neural network. In *2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014*. 2

[30] L. Hermann. Negri. *Python library PeakUtils*, 2020 [Online]. 3

[31] J. Pan and W.J. Tompkins. A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236,. 3

[32] J. Park, S. Lee, and M. Jeon. Atrial fibrillation detection by heart rate variability in poincare plot. *BioMedical Engineering OnLine*, 8(1). 1

[33] T. Pereira, N. Tran, K. Gadhoumi, M.M. Pelter, D.H. Do, R.J. Lee, R. Colorado, K. Meisel, and X. Hu. Photoplethysmography based atrial fibrillation detection: A review. *npj Digital Medicine*, 3(1). 1

[34] A.H. Ribeiro, M.H. Ribeiro, G.M. Paixão, D.M. Oliveira, P.R. Gomes, J.A. Canazart, M.P. Ferreira, C.R. Andersson, P.W. Macfarlane, W. Meira, T.B. Schön, and A.L. Ribeiro. Author correction: Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature Communications*, 11(1). 2

[35] F. Rundo, S. Conoci, A. Ortis, and S. Battiato. An advanced bio-inspired photoplethysmography (ppg) and ecg pattern recognition system for medical assessment. *Sensors*, 18(2):405,. 1

[36] F. Shaffer and J.P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5. 1

[37] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen, and Guoying Zhao. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2781–2795, 2020. 8

[38] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida. Wearable photoplethysmographic sensors—past and present. *Electronics*, 3(2):282–302,. 1

[39] Alexander Trumpp, Johannes Lohr, Daniel Wedekind, Martin Schmidt, Matthias Burghardt, Axel Heller, Hagen Malberg, and Sebastian Zaunseder. Camera-based photoplethysmography in an intra-operative setting. *BioMedical Engineering OnLine*, 17:33, 03 2018. 1

[40] G.R. Tsouri and Z. Li. On the benefits of alternative color spaces for noncontact heart rate measurements using standard red-green-blue cameras. *Journal of Biomedical Optics*, 20(4):048002,. 1, 3

[41] M. Zoni-Berisso, F. Lercari, T. Carazza, and S. Domenicucci. Epidemiology of atrial fibrillation: European perspective. *Clinical Epidemiology*, pages 213,. 1