TIME SERIES AND MORPHOLOGICAL FEATURE EXTRACTION FOR CLASSIFYING CORONARY ARTERY DISEASE FROM PHOTOPLETHYSMOGRAM

Rohan Banerjee, Sakyajit Bhattacharya, Shahnawaz Alam

Research and Innovation, Tata Consultancy Services

ABSTRACT

In this paper we propose a feature extraction algorithm for classifying Coronary Artery Disease (CAD) patients from Photoplethysmogram (PPG) signals. Several domain-independent features, representing inherent properties of a time series are explored in our study. These are combined with Heart Rate Variability (HRV) and other popularly used morphological PPG features. A statistical feature selection algorithm, based on Maximal Information Coefficient (MIC) is applied on MIMIC II dataset for ranking and choosing the optimum features. A second hospital dataset of different patient demography is used for performance evaluation. Results show that, Support Vector Machine (SVM) classifier, designed using the selected features yields average sensitivity and specificity of more than 0.8 in identifying CAD patients and also outperforms two recent prior art approaches when applied on the test dataset.

Index Terms— Photoplethysmogram, Coronary Artery Disease, Feature selection, Classification

1. INTRODUCTION

Coronary Artery Disease (CAD) is a common heart disease and also a leading cause of death in both developed and developing nations. CAD is formed due to narrowing of coronary arteries over years, which may end up causing heart attack or stroke. An early non-invasive detection/ screening of CAD is an open area of research till date. Prior art techniques reported commending accuracy in identifying CAD by analysing several biomedical signals. ECG is considered as a clinically accepted signal for first level of diagnosis. Heart Rate Variability (HRV), obtained from ECG often shows discriminative markers for CAD patients [1], [2], [3]. However, Recording and analysis of ECG for a prolonged duration for estimation of HRV is not very feasible for a low cost screening system. Analysing heart sound or Phonocardiogram (PCG) signals, captured using a digital stethoscope is also considered as an alternative approach that reports promising accuracy [4], [5], [6]. Fundamental heart sounds (S1, S2) are extracted from raw PCG. Time, frequency and statistical features are computed from the segregated heart sounds for classification. However, ambient noise in audible range often corrupts the signal quality. Hence an accurate segregation of heart sounds for feature extraction becomes a challenging task.

Photoplethysmogram (PPG) has been very popular in recent days in several biomedical applications due to low cost implementation and easy portability. PPG measures the volumetric blood flow in capillaries over time. Application of PPG is commonly found in wearable devices and smart phones for measurement of physiological vitals like heart rate or blood pressure. Literatures suggest that morphological PPG features can be utilized in detection of several cardiac diseases like arrhythmia [7] and atrial fibrillation [8]. Angius *et al.* [9] show that 'relative crest time', derived from PPG is a good indicator for distinguishing cardiovascular patients from healthy subjects. The work by Banerjee *et al.* [10] proposes several morphological PPG features for classifying CAD patients. Their algorithm is successfully validated upon two hospital datasets of different patient demography and sensor device quality. We find that, the prior art approaches are mostly limited to domain specific features related to individual cardiac cycle and HRV. However, clinically proven PPG features for identifying CAD are yet to be fully established. In this paper we propose several domain-independent features to capture the inherent properties of a PPG time series along with the existing features. Further, we present an optimum feature selection technique based on Maximal Information Coefficient (MIC) for creating a stable CAD classifier. Rest of the paper is organised as follows, Section 2 describes our feature set in detail. Then we move on to describe our experimental dataset, feature selection technique and results in Section 3, before concluding our work in Section 4.

2. PROPOSED FEATURE SET

This section describes our feature set in detail, upon which a classifier is trained for identifying CAD and non-CAD subjects. The feature set can be broadly categorized into three groups as discussed subsequently.

2.1. Time Series Features

Time series domain-independent features often contain important information regarding inherent properties of a signal, that can become discriminative markers for disease classifica-

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

tion. In time series analysis, decomposition is a critical step to transform the series into a format for statistical measuring. To obtain a precise and comprehensive calibration, some measures are calculated on the raw time series data Y_t (referring as RAW data) and some on the remaining time series after de-trending and de-seasonalizing Y_t' (referring as Trend and Seasonally Adjusted (TSA) data). A total of twelve measures are extracted from each time series including seven on the RAW data and five on the TSA data (as shown in Fig. 1). Out of the twelve features, four features, trend, seasonality, serial correlation and non-linear autoregressive structure are extracted following the method in [11]. Skewness and kurtosis are extracted using the method of moment technique. We introduce some other features, viz. periodicity, self similarity, Maharaj's distance and number of direction changes, whose extraction procedure is discussed in the following.

2.1.1. Periodicity

Periodicity determines the nature of the cyclic pattern in a time series. Due to irregularities in heart rate, periodicity varies in frequency length over the time periods, unlike seasonality. Algorithm 1 depicts the procedure of measuring periodicity used in our paper. For time series with no seasonal pattern, the period is set to 1.

Algorithm 1 Periodicity Measurement of a Time Series

```
procedure $period(X_t)$
    $X_t^* \leftarrow detrend(X_t)$
    $numlag \leftarrow floor(length(X_t^*)/3)$
    $r_k \leftarrow autocorr(X_t^*, numlag)$
        ▷ autocorrelation for all lags up to 1/3 of series length
    $[p_{loc}, t_{loc}] \leftarrow peakdet(r_k)$
        ▷ $p_{loc}$: array of peak locations
        ▷ $t_{loc}$: array of trough locations
    if $(p_{loc}[0] - t_{loc}[0]) \geq 0.1$ unit then
        $f \leftarrow p_{loc}[0]$
    else
        $f \leftarrow 1$
    end if
    return $f$
end procedure
```

2.1.2. Self similarity

Self similarity measures the rate of decrease in the autocorrelation of a time series with the increase in lag between pair of observations. It is measured only on RAW data. Self similarity of a time series can be measured by Hurst exponent (H) [12]. H can be computed using Fractional Autoregressive Integrated Moving Average (FARIMA) processes, which is generated from Brownian motion. In an ARIMA(p, q, r) process, p is the order of AR, r is the order of MA and q is the degree of differencing which is measured by the number of times the data have had past values subtracted. Generally for stationary time series the parameter is integer. However, if long-range dependence is suspected in the time series, q can be a non-integer, and it results in a FARIMA model. Here, we fit a FARIMA($0, d, 0$) by an approximation of the maximum likelihood method, as shown in [13]. We then estimate the Hurst parameter using the relation $H = d + 0.5$.

2.1.3. Average Maharaj's distance

Maharaj's distance [14] is indicative of a moving average factor along with the number of changes in direction in the time series data. Maharaj's distance can capture a desired similarity metric across spatial entities. An Autoregressive Moving Average (ARMA) time series Y_t with autoregression parameter p and moving average parameter r can be defined according to the following equation:

$$Y_t = \lambda + \sum_{i=1}^{p} \Psi_i Y_{t-i} + \sum_{i=1}^{r} \theta_i \epsilon_{t-i} + \epsilon_t$$

where λ is a constant, ϵ_t is white noise, Ψ_i-s are the autoregression parameters and θ_i-s are the moving average parameters. For such ARMA processes, discrepancy measurement based on hypotheses testing can be used to determine whether or not two time series X_t and Y_t have significantly different generating processes. The output metric of this algorithm is called the Maharaj's distance and can be used to find whether the time series are similar to each other. A p-value is computed which lies between 0 and 1. A p-value close to 1 indicates the two time series are similar, and a p-value close to 0 indicates the two time series are different. For purposes of feature extraction, the average Maharaj's distance (AMD) for the i^{th} time series (related to the PPG signal from i^{th} subject sub_i, both in training and test dataset) is measured as follows:

$$\text{AMD}_i = \sum_{j \neq i}^{n} \text{MD}_{ij}/(n-1)$$

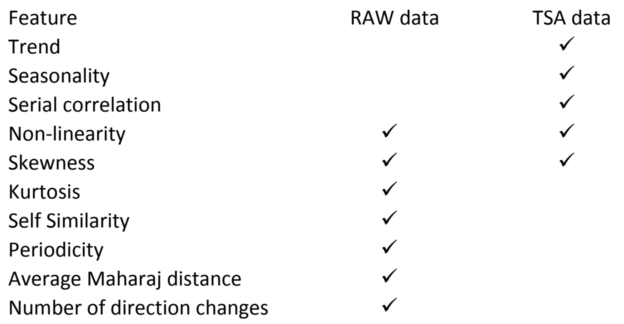
where MD_{ij} is the Maharaj's distance of the time series of sub_i from the time series of the j^{th} CAD subject in the training dataset and n is the total number of CAD subjects in the training set. This measures the average dissimilarity of an unknown test subject from the CAD population in the training set.

2.1.4. Number of direction changes

For a certain subject, a frequent irregularity in the recorded signal is an important indicator of his/her overall pathological conditions. It is observed that the sudden irregularity in the PPG waveform of a CAD subject is typically more frequent than a non-CAD subject whose signal shows a more static periodicity and stability. The number of changes in direction in the time series can therefore be determined as feature. Specifically, for a subject data Y_t, a function δ_t such that $\delta_1 = 0$ can be expressed as

$$\delta_t = \begin{cases} 1, & \text{if } Y_{t-1} > Y_t > Y_{t-2} \quad \text{or} \quad Y_{t-1} < Y_t < Y_{t-2} \\ 0, & \text{otherwise} \end{cases}$$

Let $\Delta = \sum_t \delta_t / length(Y_t)$. Then Δ is the sum of the number of direction changes is taken to be a new feature. This feature can also be used as a heuristic method to catch the noise, because the value will be too high for a noisy signal.



Feature	RAW data	TSA data
Trend		✓
Seasonality		✓
Serial correlation		✓
Non-linearity	✓	✓
Skewness	✓	✓
Kurtosis	✓	
Self Similarity	✓	
Periodicity	✓	
Average Maharaj distance	✓	
Number of direction changes	✓	

Fig. 1. Summary of Time Series Features

2.2. HRV Features

HRV of a CAD patient shows significantly different pattern compared to a non-CAD subject [3], [2]. HRV related features are measured from the successive peak to peak distances (NN intervals) in a signal. Shannon entropy of the NN intervals is found to be an important feature for classification.

$$E_{sh} = -\sum_{m=1}^{N} p_m \log p_m$$

A normalized histogram with N bins is computed for the NN interval distribution. The empirical probability of each bin is denoted by p_m. Here $m \in 1...N$ and $\sum_{m=1}^{N} p_m = 1$. In general, E_{sh} is found to be higher for CAD patients due to irregularities in HRV. Other features include, Root Mean Square of successive NN interval differences, normalized by mean hearts rate ($nRMSSD$), mean absolute deviation (MAD) of NN intervals as well as Kurtosis and skewness of NN intervals.

Frequency domain HRV features are derived from the power spectrum of NN intervals. The normalized spectral power in three frequency regions (VLF, LF and HF) as detailed in [10] are used as features.

2.3. Morphological PPG Features

Several prior art morphological features, detailed in [9] and [10] are also considered. These are mean and standard deviation of pulse interval (T_c), normalized crest time (T_1), normalized diastolic time (T_2) and ratio between crest time and diastolic time ($ratio$) in a measurement.

3. DATA ANALYSIS

This section describes our experimental dataset, feature selection technique and results. Performance of our proposed methodology in comparison with two recent prior art techniques are also reported.

3.1. Experimental Dataset

A PPG dataset, selected from MIMIC II waveform dataset matched subset [15], [16] is used for feature selection. Subsequently the selected features are applied on another hospital dataset of different patient demography and sensor quality for performance evaluation and prior art comparison. A total of 56 CAD and 74 non-CAD patient subjects are selected from MIMIC II depending upon availability of PPG. The disease information can be retrieved from the billing information available in the matched subset against individual patient for annotation. The second dataset is prepared by us from an urban hospital in Kolkata, India under the supervision of a medical practitioner using non-medical grade commercial pulse-oximeter (CMS 50D+) at a sampling rate of 60 Hz. This dataset comprises a total of 99 patient data, including 52 CAD and 47 non-CAD subjects. The data collection drive is approved by the hospital ethics committee. Individual patient consent for collecting his/her data is also in place. The dataset ensures a wide variation in patient demography along with different pathological conditions for non-CAD patients and also varying percentage level of heart blockages for CAD patients. Being collected in an uncontrolled environment (hospital cath labs) using non-medical grade oximeter device, this dataset is noisier than MIMIC II, recorded from ICU patients with restricted body movement using clinical instruments. The signal quality assessment algorithm in [17] is applied on both the datasets to extract two minutes of clean signal from each subject. This duration ensures to preserve the HRV information in the collected signal.

3.2. Feature Selection

Feature selection is often found useful in classification problems to reduce the processing time and also to improve the accuracy by removing noisy features. The feature selection algorithm used in this paper is a combination of both filter and wrapper methods. All features are initially ranked with respect to ground truth labels based on MIC score on MIMIC II dataset. The optimum feature set is selected from the ranked feature list in a cross validation approach.

MIC measures the statistical relationship between a pair of dataset, by forming grids with various sizes to find the largest mutual information between them [18]. For each pair of data (x, y), if I is the mutual information for a grid G, then MIC of a set D of pairwise data with sample size n and grid size (xy) less than $B(n)$ is given by [18]

$$MIC(D) = max_{xy<B(n)}\{M(D)_{x,y}\}$$

where $B(n)$ is a function of sample size (usually $B(n) = n^{0.6}$). For different distributions of G, $M(D)$ is given by

$$M(D)_{x,y} = \frac{max\{I(D|G)\}}{\log min(x,y)}$$

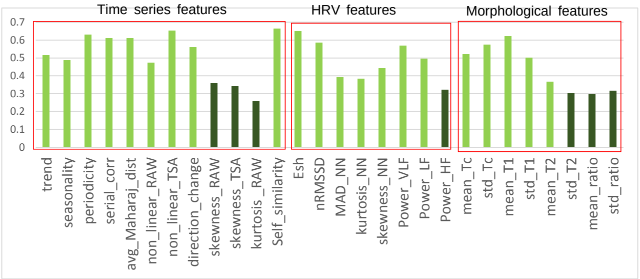


Fig. 2. MIC of All 28 PPG Features w.r.t. Ground Truth

Fig. 2 shows that time series and HRV features are more strongly related to the ground truth labels due to higher MIC values. To obtain the optimum feature set, 5-fold cross validation is applied on MIMIC II in an iterative way, increasing the feature dimension by adding one feature at each iteration from the ranked list. Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel is used for classification. For an unbalanced dataset, the classifier might be biased towards the major class even if a very high accuracy is obtained. Thus in our approach our stopping rule aims to identify the feature list of minimum dimension that produces an optimum and stable sensitivity and specificity of detecting CAD patients.

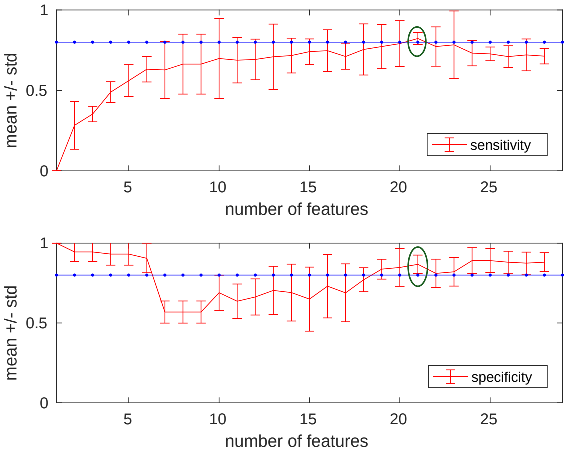


Fig. 3. Effect of Adding Features on MIMIC II

Fig. 3 depicts the effect of adding new features in sensitivity and specificity in terms of $mean \pm std$. It can be observed that, sensitivity starts improving from zero with addition of new features, reaches a maximum value, then starts falling. On the contrary, specificity starts with one, drops to a certain level and eventually gets saturated as more features are added. The circled regions in Fig. 3 where both mean sensitivity and specificity become sufficiently high (> 0.8) with minimum standard deviation, is considered the region of optimum feature set. Thus top 21 features (light shaded bars in Fig. 2), including 9 time series, 7 HRV and 5 morphological features construct the optimum feature set for classification.

3.3. Results

Our experimental results are performed on the in-house dataset of different patient demography and sensor quality than MIMIC II. A 5-fold cross validation approach is used to report the average and minimum values of sensitivity and specificity of detecting CAD patients. SVM with RBF kernel is used for classification. Our methodology is also compared with two recent prior art approaches in [9] and [10] on the same dataset. Table 1 shows that the classifier trained using all 28 features outperforms both the prior arts without applying feature selection. This proves the importance of adding the proposed time series features with morphological and HRV features popularly used in prior arts. However, it can be observed that the classification performance across the folds are quite unstable, resulting in a high difference between average and minimum values of both sensitivity and specificity. Whereas the proposed feature selection technique ensures a stable performance by removing the noisy features. Thus an improved and stable performance can be achieved by incorporating the same. Consistency in performance of our methodology on two datasets during feature selection and final evaluation suggests that the feature selection is independent to patient demography and sensor quality.

Table 1. Performance Analysis on Test Dataset

Methodology	$Sensitivity$		$Specificity$	
	$min.$	$avg.$	$min.$	$avg.$
Prior art [9]	0.46	0.62	0.64	0.77
Prior art [10]	0.67	0.75	0.73	0.80
All 28 features	0.70	0.79	0.74	0.82
Selected 21 features	0.78	0.83	0.82	0.86

4. CONCLUSION

This paper explores several time series, HRV and morphological PPG features for classifying CAD and non-CAD subjects. A statistical feature selection technique is also proposed for stabilizing the classification performance. Results show the utility of the proposed time series features for identifying CAD patients as our methodology outperforms two recent prior art techniques. Our future works include a successful validation of the methodology, on a larger and more diverse test dataset. We are also planning to explore the feasibility of applying deep learning based approaches for possible performance improvement.

5. REFERENCES

[1] Tatyana Mironova, Vladimir Mironov, Vladimir Antufiev, Eleonora Safronova, Michael Mironov, and Evgenia Davydova, "Heart rate variability analysis at coronary artery disease and angina pectoris," *Recent patents on cardiovascular drug discovery*, vol. 4, no. 1, pp. 45–54, 2009.

[2] Rungroj Krittayaphong, Wayne E Cascio, Kathleen C Light, David Sheffield, Robert N Golden, Jerry B Finkel, George Glekas, Gary G Koch, and David S Sheps, "Heart rate variability in patients with coronary artery disease: differences in patients with higher and lower depression scores," *Psychosomatic Medicine*, vol. 59, no. 3, pp. 231–235, 1997.

[3] H. V. HUIKURI, "Heart rate variability in coronary artery disease," *Journal of Internal Medicine*, vol. 237, no. 4, pp. 349–357, 1995.

[4] Dominique Gauthier, Yasemin M Akay, Robert G Paden, William Pavlicek, F David Fortuin, John K Sweeney, Richard W Lee, and Metin Akay, "Spectral analysis of heart sounds associated with coronary occlusions," in *Information Technology Applications in Biomedicine, 2007. ITAB 2007. 6th International Special Topic Conference on*. IEEE, 2007, pp. 49–52.

[5] Samuel E Schmidt, John Hansen, Henrik Zimmermann, Dorte Hammershøi, Egon Toft, and Johannes J Struijk, "Coronary artery disease and low frequency heart sound signatures," in *Computing in Cardiology, 2011*. IEEE, 2011, pp. 481–484.

[6] R. Banerjee, A. Dutta Choudhury, P. Deshpande, S. Bhattacharya, A. Pal, and K. M. Mandana, "A robust dataset-agnostic heart disease classifier from phonocardiogram," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2017, pp. 4582–4585.

[7] L. F. Polana, L. K. Mestha, D. T. Huang, and J. P. Couderc, "Method for classifying cardiac arrhythmias using photoplethysmography," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015.

[8] J. Lee, B. A. Reyes, D. D. McManus, O. Mathias, and K. H. Chon, "Atrial fibrillation detection using a smart phone," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2012, pp. 1177–1180.

[9] G. Angius, D. Barcellona, E. Cauli, L. Meloni, and L. Raffo, "Myocardial infarction and antiphospholipid syndrome: A first study on finger ppg waveforms effects," in *2012 Computing in Cardiology*, Sept 2012, pp. 517–520.

[10] Rohan Banerjee, Ramu Vempada, K. M. Mandana, Anirban Dutta Choudhury, and Arpan Pal, "Identifying coronary artery disease from photoplethysmogram," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 2016, UbiComp '16, pp. 1084–1088, ACM.

[11] Xiaozhe Wang, Kate Smith, and Rob Hyndman, "Characteristic-based clustering for time series data," *Data Mining and Knowledge Discovery*, vol. 13, pp. 335–364, 2006.

[12] Bo Qian and Khaled Rasheed, "Hurst exponent and financial market predictability," in *IASTED conference on Financial Engineering and Applications*, 2004, pp. 203–209.

[13] John Haslett and Adrian E. Raftery, "Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource," *Applied Statistics*, vol. 38, no. 1, 1989.

[14] E.A. Maharaj, "Clusters of time series," *Journal of Classification*, vol. 17, pp. 297–314, 2000.

[15] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L. Lehman, G.B. Moody, T. Heldt, T.H. Kyaw, B.E. Moody, and R.G. Mark, "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): A public-access intensive care unit database," *Critical Care Medicine*, vol. 39, no. 5, pp. 952–960, 2011.

[16] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[17] Shahnawaz Alam, Shreyasi Datta, Anirban Dutta Choudhury, and Arpan Pal, "Sensor agnostic photoplethysmogram signal quality assessment using morphological analysis," in *Proceedings of the 14th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2017.

[18] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.