CAD Patient Classification Using MIMIC-II

Swarnava Dey[1(✉)], Swagata Biswas[1], Arpan Pal[1], Arijit Mukherjee[1], Utpal Garain[2], and Kayapanda Mandana[3]

[1] Innovation Labs, Tata Consultancy Services Ltd., Kolkata, India
{swarnava.dey,swagata.biswas,arpan.pal,mukherjee.arijit}@tcs.com
[2] Indian Statistical Institute, Kolkata, India
utpal.garain@gmail.com
[3] Cardiothoracic and Vascular Surgery Department,
Fortis Healthcare Limited, Kolkata, India
kmmandana@gmail.com
http://www.tcs.com, http://www.isical.co.in

Abstract. With availability of large volume of collected data from healthcare centers and significant improvement in computation power, evidence based learning is helping in building robust disease diagnostic models.

In this work MIMIC-II database, consisting of physiologic waveforms and clinical Information about ICU patients, is used for patient classification, taking Coronary Artery Disease (CAD) as a use case.

A learning algorithm (wavelet transform + SVM) is trained and evaluated for CAD patient segregation with 89% accuracy on ICD-9 labeled MIMIC-II Photoplethysmogram (PPG) signals. Due to the noisy nature of machine collected MIMIC-II ICU data, the same SVM model was validated on a local hospital dataset containing doctor labeled PPG signals resulting a 5% accuracy gain.

This work is the first attempt of CAD patient classification on MIMIC-II, using heart rates from easily obtainable PPG signal suitable in mobile/wearable setting.

Keywords: Heart rates · HRV · Wavelet transform · SVM · ICD-9 · PPG · Photoplethysmogram · CAD · Coronary Artery Disease · Bigdata

1 Introduction

MIMIC-II [9] is largescale database of medical records for 32000 patients. Among these patients, waveform (physiologic signals sampled at 125 Hz) data for around *2800* patients are matched with clinical records (information about patients including their diseases). The longitudinal physiologic readings when accompanied by a disease related ground truth, form the bases of evidence based learning of disease based patient classification.

For diagnosis/prognosis on MIMIC-II data, ground truth is sourced from hospital billing codes to perform model training using the physiologic signals and

K. Giokas et al. (Eds.): eHealth 2016, LNICST 181, pp. 370–375, 2017.
DOI: 10.1007/978-3-319-49655-9_44

the patient's disease class. The billing codes stored in MIMIC-II are International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9 -CM) codes. A more accurate field *Diagnosis on admission* is being added to the new MIMIC-III database [5].

Coronary Artery Disease (CAD) [3] can be defined as the blockage of arteries that supply blood to the heart, resulting from atherosclerosis (an accumulation of fatty materials on the inner linings of arteries). MIMIC-II has 600+ CAD patient's waveforms in the matched dataset forming a rich resource for people working on CAD diagnosis. As of now *coronary angiography* [4] is the only reliable way to diagnose CAD. The World Health Organization has marked CAD as a *modern epidemic* and there are not enough facilities available for timely diagnosis, especially in India due to a high patient to doctor ratio. In our Lab, more than one teams are working together to build non-invasive methods using easily available physiologic signals, to help clinicians screen CAD patients and ensure that the access to angiography is available to patients who really need it.

State of the art (SoA) survey revealed that earlier work attempted CAD diagnosis using ECG signal that not realizable in a mobile or wearable setting. A photoplethysmogram (PPG) signal, easily obtainable from fingertip using pulse oximeter or mobile camera, can denote the cardiac cycle [2]. In the current work feature extraction and machine learning methods were applied on the heart rate timeseries (HR) of 611 patients (around 850 Gigabytes waveform) in MIMIC-II dataset. Though MIMIC-II physiological signals were cleaned before using, the noisy nature of machine collected MIMIC-II ICU data was suspected to be the reason behind relatively low classification accuracy. For a second level validation the same model was evaluated on HR signal from a controlled proprietary PPG dataset and there was significant improvement of results.

This work is first attempt of CAD/non CAD patient classification on public dataset like MIMIC-II, using HR signal from easily obtainable PPG signal, using a new set of wavelet transform based features. This paper reports the design and evaluation of CAD patient diagnosis and is structured as follows: Sect. 2 gives a brief overview of SoA, Sect. 3 discusses the methodology, Sect. 4 outlines the classification results and finally the conclusion is drawn in Sect. 5.

2 Related Work

The SoA survey for the current work was undertaken in three areas namely: (1) earlier works on diagnosis of CAD using PPG signals, (2) CAD diagnosis on a large patient dataset and (3) disease diagnosis on large datasets, specifically MIMIC-II using ICD codes.

Though there is no earlier attempt of using PPG signals for diagnosis of CAD, several earlier works attempt non invasive detection of CAD using ECG and HRV from ECG. In [1] authors applied statistical and signal processing to extract features from HRV (from ECG) and applied SVM for classification. The two datasets comprised of 20 CAD, 20 normal subjects and 6 CAD, 6 normal subjects, respectively. Results indicate accurate classification of the subjects.

In [7] Wavelet Package Transform(WPT) is used to analyze HRV signals. Performance evaluation is done using least square support vector machine(LS-SVM) classification algorithm. An average of 90% accuracy is achieved on the test dataset using db4 as the wavelet function. In [6] HR signal is decomposed into frequency sub-bands using wavelet transform and dimensionality reduction are applied on the coefficients to get top features. Selected features are fed into different classifiers. For 10 CAD subjects and 15 normal subjects, accuracy of 96.8%, sensitivity of 100% and specificity of 93.7% is achieved using a combination of Independent Component Analysis-Gaussian Mixture Model.

This work reports the best result for CAD detection, however dataset size on which the experiments were conducted are small. Also, use of ECG as the physiologic signal source is not realizable for mobile or wearable setting. For the same reason we ignore several CAD detection methods using costly technique like Stress Cardiovascular Magnetic Resonance(CMR), Single Photon Emission Computed Tomography(SPECT) etc. In the area of disease diagnosis on large datasets, no attempt was found to use MIMIC-II database for diagnosis of CAD. However, MIMIC-II dataset is used for several other disease diagnosis works.

3 Methodology

The end to end system is represented in Fig. 1 and detailed in later subsections. Total 611 patients in MIMIC-II dataset are having PPG data (CAD:267 and non CAD:344), segregated using SQL queries on the MIMIC-II clinical database.

3.1 Data Acquisition

PPG signal for MIMIC-II, sampled at 125 Hz, is obtained using WFDB tools [10].

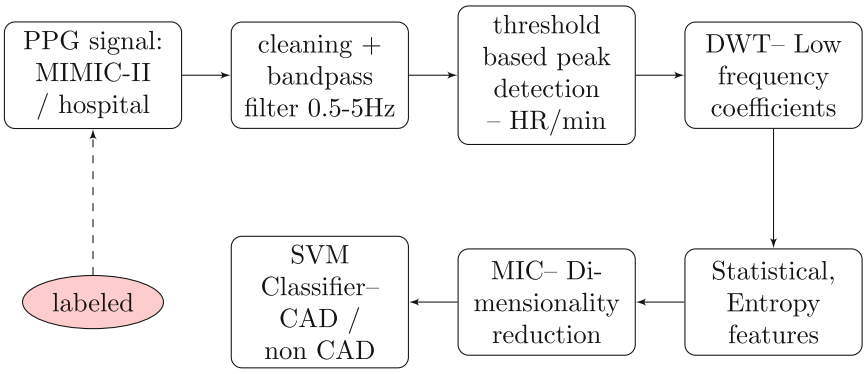

Fig. 1. Processing chain

ICD-9 code 414.01 is taken as CAD and all other patients except patients with any circulatory disease are taken as non CAD, as per doctor's advise.

The PPG waveform is obtained from the patients in a local hospital using a standard pulse oximeter, sampled at 60 Hz. This doctor verified PPG dataset is comprised of files from 14 CAD and 15 non CAD patients. Detected cardiac cycles [11] from PPG waveform, are used to get interpolated $(60 * \frac{1}{RR})$ HR bits/minute timeseries.

3.2 Feature Extraction

CAD or non CAD patients, having associated PPG waveform can be represented as vectors $V_1, V_2, ..., V_n$ for n number of patients. The vector $V_i \in R^{t_i}$ contains PPG signal sample at time t_i. It is observed from literature that patients with CAD have different heart rate variations from normal people, noticed in low frequency ranges [6]. Wavelet analysis helps both time and frequency localization using different sized windows at desired frequencies [11]. The wavelet transform of V_i is given as: $[W_\psi Vi](a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \overline{\psi\left(\frac{t-b}{a}\right)} Vi(t)dt$, where ψ(t) is the transforming function called the mother wavelet. The scaling and translation parameters are a and b respectively. For HR signal, daubachies wavelet db3 [12] is found to be most suitable as mother wavelet after observing the similarities with the HR signals, both visually and using energy and entropy measures. The HR signal from PPG of V_i are wavelet transformed to find coefficients. Statistical and entropy related features like mean, variance, maximum and minimum of energy, amplitude, frequency are computed from each coefficient. Total number of features extracted is dependent on the number of coefficients generated, i.e., level of DWT decomposition and for the current work the level of decomposition used is *four*.

3.3 Feature Selection and Classification

In order to avoid *curse of dimensionality* [13] only the top contributing features (usually 5–10) were selected given by MIC [8] strength. Sample MIC strength of top five parameters for MIMIC-II data are *0.56,0.54, 0.49,0.39, 0.39* and for collected data are *0.72,0.58,0.54,0.49, 0.39*.

The selected features are used to train an SVM classifier [14] and evaluated using 10-fold cross validation. The idea behind SVM is to construct a hyperplane or a set of hyperplanes in a high dimensional space such that the given input gets classified into different classes. For the current work linear SVM did not give good results as it was not possible to construct a maximum-margin hyperplane that divides the training set into two classes. Different kernels were evaluated to map the input space to a high dimensional feature space using non linear transformation and *radial* kernel with tuning gave best results.

4 Experiments and Results

The various accuracy measures used in current context can be expressed as follows: (1) True positive (TP): CAD correctly identified as CAD (2) False positive (FP): Non CAD incorrectly identified as CAD (3) True negative (TN): Non CAD correctly identified as Non CAD (4) False negative (FN): CAD incorrectly identified as Non CAD

$Sensitivity = \frac{TP}{TP+FN}$ i.e., correctly identified CAD out of total CADs
$Specificity = \frac{TN}{TN+FP}$ i.e., correctly identified non CAD out of total non CADs
$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$ i.e., all correct identifications out of all patients.

The experimental results of two class classification for both MIMIC-II and local hospital datasets (described in Sect. 3.1) is presented in Table 1.

Table 1. Comparative classification results in local hospital and MIMIC-II patients

Dataset	Patient & class labels	Accuracy	Sensitivity	Specificity
MIMIC-II	CAD:267 NCAD:344 Total:611	89%	86%	90%
Local hospital	CAD:14 NCAD:15 Total:29	93%	92%	94%

In summary, for MIMIC-II 14% of CAD patients and 10% of non CAD patients were predicted wrongly and it is suspected that machine noises, motion artifacts etc. in MIMIC-II data (collected automatically at ICU setup), is the reason behind this relatively low classification accuracy Table 1. For the current work signal files of 611 patients were cleaned by visual inspection and some erroneous data may have remained. Thus, for a second level validation, the same model was validated on a controlled dataset of 29 patients collected from a local hospital (Sect. 3.1) and the increase in both sensitivity and specificity was 6% and 4% respectively. Joint analysis with consultant doctor revealed that the two misclassified patient out of 29, were diabetic, which possibly prevented the 100% correct classification.

The efficacy of the feature extraction and learning algorithm developed in the current work proved to be high as the model was trained using ICU data collected at USA and gave good results when evaluated on a hospital dataset in India.

5 Conclusion

In this work the issue of evidence based disease diagnosis is addressed on a large patient database taking CAD as an example disease.

On a set of patients labeled CAD/nonCAD, each having associated PPG waveform, wavelet transform was applied to generate wavelet coefficients at frequency ranges where peaks are often observed for cardiac patients. Statistical

and entropy related features were extracted from coefficients and top contributing features were selected using MIC dimensionality reduction. Selected features were used to train SVM classifier and evaluate the classifier Performance. With no prior benchmark of CAD patient classification on MIMIC-II dataset, the above scheme classified MIMIC-II PPG signals labeled by ICD-9 code with 89% accuracy. Due to noisy nature of MIMIC-II data the scheme was re-evaluated on a doctor labeled hospital dataset resulting in 5% accuracy gain. The major contribution of the work is CAD diagnosis using easily obtainable PPG signal suitable in mobile/ wearable setting. The local hospital data does not have many samples and this is probably the scenario for many datasets and therefore, these datasets might not be suitable for training and validating machine learning algorithms. In this context, our present experiment shows that use of a pre-trained network (e.g., trained on MIMIC-II dataset) could be useful for the problem in hand. As significant time was spent in manual feature extraction, use of unsupervised feature extraction and transfer learning from MIMIC-II model will be explored for CAD diagnosis in future endeavors.

References

1. Kampouraki, A., et al.: Heartbeat time series classification with support vector Machines. IEEE Trans. Inf. Technol. Biomed. **13**(4), 512 (2009)
2. Mohamed, E.: On the analysis of fingertip photoplethysmogram signals. Curr. Cardiol. Rev. (2012). doi:10.2174/157340312801215782
3. Coronary artery disease. https://en.wikipedia.org/wiki/Coronary_artery_disease
4. Cardiac catheterisation and coronary angiography. http://www.nhs.uk/conditions/CoronaryAngiography/Pages/Introduction.aspx
5. MIMIC-II vs MIMIC-III. https://mimic.physionet.org/mimicdata/whatsnew/
6. Giri, D., et al.: Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA discrete wavelet transform. Knowl. Based Syst. **37**, 274 (2013)
7. Kheder, G., et al.: HRV analysis using wavelet package transform and Least Square Support Vector Machine. Int. J. Circ. Syst. Sig. Process. **2**(1) (2008)
8. Reshef, D.V., et al.: Detecting novel associations in large datasets. Science. **334**(6062), 1518–1524 (2011)
9. Saeed et al.: Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database **39**, 952–960 (2011). doi:10.1097/CCM.0b013e31820a92c6
10. The WFDB software package. https://www.physionet.org/physiotools/wfdb.shtml
11. Camm, et al.: Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Circulation **93**, 1043–1065 (1996)
12. Daubechies, I.: Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia (1992)
13. Verleysen, M., François, D.: The curse of dimensionality in data mining and time series prediction. In: Cabestany, J., Prieto, A., Sandoval, F. (eds.) IWANN 2005. LNCS, vol. 3512, pp. 758–770. Springer, Heidelberg (2005). doi:10.1007/11494669_93
14. Cortes, C., Vapnik, V.: Networks, support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). doi:10.1023/A:1022627411411