PREDICTION OF CORONARY ARTERY DISEASE USING ELECTROCARDIOGRAPHY: A MACHINE LEARNING APPROACH

GAUTAM PHADKE[1]**, MOHAMMED REZA RAJATI**[1]**, LEENA PHADKE**[2]

[1]Department of Computer Science, University of Southern California, Los Angeles, USA
[2]Department of Physiology, SKN Medical College and General Hospital, Pune, India
E-MAIL: gphadke@usc.edu, rajati@usc.edu, leena_phadke@hotmail.com

Abstract:

Coronary Artery Disease (CAD) is a leading cause of cardiovascular morbidity and mortality globally. There has been an indication of association between Electrocardiography (ECG), a measurement for electrical activity in the heart, and CAD, which makes ECG a promising screening tool. Consequently, Machine Learning techniques can detect patterns of ECG that are able to screen CAD cases. We developed a machine learning tool that extracts RR interval features from ECG signals, and used different statistical learning algorithms to detect CAD based on these features. Our results indicate that patterns in ECG signals and attributes of patients such as age and gender can predict CAD in diverse clinical scenarios in real life with a performance superior to the available screening and diagnostic tests.

Keywords:

Coronary Artery Disease; Feature Extraction; Extreme Gradient Boosting Classifier; Statistical Machine Learning

1. Introduction

According to World Health Organization reports [1], approximately 3.8 million men and 3.4 million women worldwide die each year due to Coronary Artery Disease (CAD), a leading cause of cardiovascular mortality globally. Two third of patients die even before reaching the hospital. Clinically, CAD may either remain asymptomatic for long period, or it may manifest as life threatening Acute Coronary Syndrome (ACS). Coronary Angiography (CAG), a gold standard investigation for definitive diagnosis of CAD, requires well equipped labs, and is a costly procedure [2]. Also, it is highly exacting to identify genuine patients for CAG referrals at the primary point of contact, when the patient load is high and quick decision making is required. CAG referrals are based on scores of ACS triage tool at the emergency departments, whereas, screening tests like exercise stress test, stress echocardiography, etc. are used at secondary care centers. Unfortunately, these tests have known limitations, contraindications and moderate sensitivities [3]. Some of the limitations include high cost of device used for stress test, and limited availability of professionals that have expertise in interpretation of results of these tests [4]. It is therefore important to develop a screening tool for prediction of CAD, that not only overcomes these limitations, but would be cost effective and easy to use even at first point of contact.

There has been an indication of association between Electrocardiography (ECG), a measurement of electrical activity in the heart, and CAD [5],[6]. Being non-invasive in nature, and a low cost procedure with wide operational availability, use of ECG to develop a promising screening tool merits research. However, wide inter and intra individual variations in ECG records of patients make it a challenging task to effectively use ECG data to screen CAD patients.

With the upsurge of application of statistical machine learning and deep learning in healthcare over the past few years [7],[8],[9], multiple studies have attempted to predict CAD with various bio-markers such as Heart Rate Variability (HRV), genetic variations, etc [10],[11],[12]. The aim of the present study is to explore the utility of ECG as a screening tool for CAD. The rest of this paper is organised as follows: Section 2 introduces viewers to previous work on CAD prediction and detection. Section 3 provides our data collection process. Section 4 presents methodology of our study. We divide section 4 into two parts: Feature Extraction/Selection and application of machine learning algorithms for CAD prediction using these features. Section 5 lists the results obtained using algorithms developed in section 4. In section 6, we discuss our conclusions.

2. Related Work

With advancement in application of Machine Learning in health sciences over the past few years, multiple studies, albeit in restricted clinical scenarios, have attempted to predict CAD using various bio-markers. In [10], Random Projections with K-Nearest Neighbours were used to predict CAD based on a feature set comprising genetic variations at base-pair level. In [13] several machine learning algorithms such as neural networks, Support Vector Machines, Decision Trees and Bayesian Classifier are applied to data that consist of combination of HRV and image of Carotid artery. In [14] a novel variable selection technique was developed using Random Forests to select the best HRV features for predictive analysis. Principal Component Analysis was applied in [15] to HRV indices for reducing the dimensionality of the feature set, followed by a multilayer perceptron algorithm to predict the presence of CAD.

3. Cohort

Our study is conducted in compliance with the guidelines given by Taskforce of European Society for Cardiology [2]. It is a prospective observational cohort study and was performed in Sinhagad Kashibai Navale Medical College and General Hospital (SKNMC & GH), Pune, India, between April 2018 and January 2019. All consecutive consenting patients referred for CAG were included in the study. CAG referrals were based on Frammingham risk score [16], patients with ACS provisionally diagnosed based on clinical features, resting ECG troponin test [17], and echocardiography impression.

In the stipulated duration, a total of 235 patients between 30 to 75 years of both genders were enrolled. All patients received medical treatment as per the diagnosis and treatment protocols. Patients with cardiac arrhythmia, ectopic on resting ECG, contraindications for angiography, and previously diagnosed with CAD were excluded from the study. After exclusion, data of 152 patients was used for analysis. Table 1 depicts demographic profile and clinical risk factors for the patients. The study was approved by institutional ethics committee of SKNMC & GH.

TABLE 1. Demographic profile of patients

Parameters	CAD Present	CAD Absent
Count	108	44
Age	59 ± 9 years	53 ± 8 years
Gender	31 F, 77 M	22 F, 22 M
Hypertension	76	22
Smoking	23	3
Dyslipidaemia	17	6
Obeisity	23	9

4. Methodology

4.1. Feature Extraction

30 minutes high sampled two lead ECG (1KHz) was recorded using Chronovisor ambulatory data acquisition system in supine position prior to angiography. We selected multiple time windows of 100 seconds for every patient as our data points. Figure 1 displays a snapshot of ECG over a six second time interval. The bounding box represents one single unit of PQRST complex [18]. This complex depicts the electrical impulses inside the heart, and is divided into 3 parts: The P wave that indicates atrial depolarization, the QRS complex that is responsible for ventricular depolarization and the T wave that represents ventricular re-polarization [19]. We use the Pan-Tomkins algorithm to detect QRS complex in ECG signal [20]. This algorithm applies a series of filters on ECG signal and removes the background noise. Then, it squares the signal to amplify QRS complex. Finally, it applies adaptive thresholds to detect peaks of the filtered signal. We used the HeartPy package [21] for process of R peak detection. Figure 2 displays detected R peaks (marked by green points).

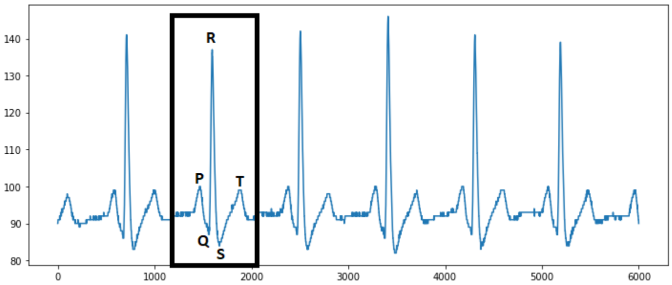

FIGURE 1. ECG Signal

4.2. Feature Selection

For a given patient, 30 consecutive RR intervals from 100 second ECG window were selected as feature sets. The RR interval corresponds to time elapsed between two successive R waves of the QRS complex. It is a function of intrinsic properties of the sinus node as well as autonomic influences [22]. Thus, it is reasonable to use RR intervals as feature sets.

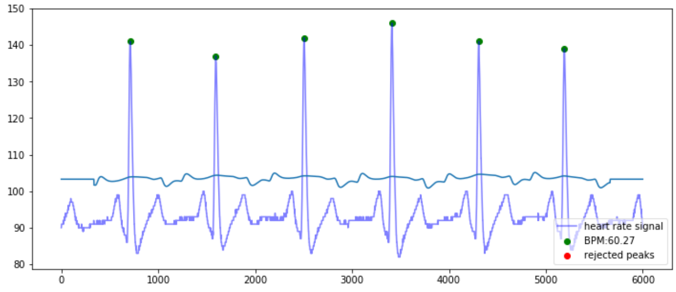

FIGURE 2. Peak Detection

The original dataset contained 108 patients (71%) with CAD Present and 44 patients (29%) with CAD Absent, indicating a class imbalance. Thus, we developed a custom pre-processsing technique to balance the class distribution, as well as to augment the data. Our technique is inspired from bootstrapping [23], which is a technique that uses random sampling with replacement. For patients with CAD Present, 9 blocks of 100 seconds, whereas for patients with CAD Absent, 18 blocks of 100 seconds ECG window were sampled, and the feature set was populated with 30 consecutive RR intervals derived from each of these blocks. 1743 samples were obtained after overasmpling, out of which 954 (53.5%) belong to CAD Present, and 789 (46.5%) belong to CAD Absent. Figure 3 shows an example of 5 such RR intervals.

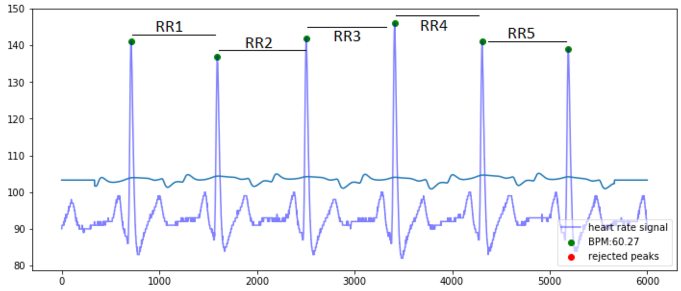

FIGURE 3. RR Interval

We prefer the above technique over other oversampling techniques like Synthetic Minority Oversampling (SMOTE) [24] for two reasons. First, unlike SMOTE, we don't generate artificial patients for oversampling our data. Second, we use the inherent assumption that a given 100 second ECG block equally represent the heart signal activity, as any other 100 second ECG block would represent. This assumption is based on the findings that shorter span ECG data (about more than 1 minute) can be reliably used for heart rate variability analysis [25]. Additionally, along with the 30 RR intervals, we use Age and Gender as our features as well.

4.3. Problem Formulation

Our dataset D is defined as $D = (X, y)$, where $X^T = [x_1, x_2, ..x_n]$ and $y^T = [y_1, y_2, ..y_n]$. Every $x_i \in R^m$ (an m dimensional feature vector), $y_i \in \{0, 1\}$ and $\mid D \mid = n$. The dataset consists of 1743 samples (n = 1743) and 32 features (m = 32). This is a binary classification problem where the label y can either be 0 (represents patients with CAD absent) or 1 (represents patients with CAD present). A stratified split (based on label values) is performed on the dataset so as to obtain 75% training samples (X_{train}) and 25% test samples (X_{test}). This ensures that ratio of subjects with CAD absent, to subjects with CAD present remains same in both training set and test set. We train different learning algorithms, such as Random Forests, Logistic Regression, Support Vector Machines, Naive Bayes, and Gradient Boosting on our training set. We use a stratified K-Fold cross-validation to obtain the best model, as well as the parameters for out best model (A process called as Hyperparameter optimization).

4.4. Learning Algorithms

We use multiple statistical machine learning algorithms, such as Random Forest [26], Logistic Regression, Naïve Bayes [27], Extreme Gradient Boosting (XGBoost) [28], etc. to train on our dataset. We follow the recommendations of [29] and use 5 fold grid-search cross-validation to determine optimal hyperparameters for each of the learning algorithm. The approach methodically builds and evaluates a model for each combination of estimator parameters specified in the grid. We briefly discuss some of these algorithms below.

4.4.1 Naïve Bayes Classifier

Naive Bayes algorithm applies Bayes' theorem with the naive assumption of conditional independence between every pair of features conditioned on value of class variable. Let **x** = $[x_1, x_2, ..., x_n]$ be the features, and let y be the class label. According to Bayes theorem:

$$P(y|x_1, x_2, ..., x_n) = \frac{P(y)P(x_1, x_2, ..., x_n|y)}{P(x_1, x_2, ..., x_n)} \tag{1}$$

Using conditional independence assumption on (1), we obtain the following classification rule:

$$\hat{y} = \underset{y}{\text{argmax}}\, P(y) \prod_{i=1}^{n} P(x_i|y) \quad (2)$$

where, $\hat{y}$ is the predicted label. We use both Gaussian and Multinomial versions of Naive Bayes algorithm.

4.4.2 Logistic Regression

It is a linear model used for classification, that models conditional probability of label y given features $\mathbf{x}$ using a logistic (sigmoid) function. In the case of binary classification, logistic function can be written as:

$$h_\theta(\mathbf{x}) = P(y = 1|\mathbf{x}; \theta) = \frac{1}{1 + \exp{(-\theta^T \mathbf{x})}} \quad (3)$$

where, θ is a set of learnable parameters. Optimal parameters are the one that maximize the log likelihood function, given by:

$$\hat{\theta} = \max_{\theta} \sum_{i=1}^{n} (y^i log(h_\theta(x^i)) + (1 - y^i) log(1 - h_\theta(x^i))) \quad (4)$$

Our implementation uses regularized version of logistic regression, namely l_1 penalized logistic regression and l_2 penalized logistic regression, in order to avoid overfitting.

4.4.3 XGBoost Classifier

Extreme Gradient Boosting Classifier (XGBoost) [28] is an implementation of Gradient Boosting Machines (GBM's) that are used for supervised learning. As per the author of XGBoost algorithm, XGBoost uses a more regularized model formalization to control overfitting, which gives it better performance compared to other GBM's. An exhaustive study conducted by [29] on Penn Machine Learning Benchmark (PMLB) dataset shows that XGBoost outperforms every other traditional Machine Learning algorithm present in the literature. Our implementation uses decision tree ensembles as a modelling choice for XGBoost.

4.5. Evaluation Metrics

As our problem falls under the category of binary classification, we use the metrics of Sensitivity, Specificity and F-1 Score to evaluate the performance of all our models. Let TP, TN, FP, FN denote number of True Positives, True Negatives, False Positives and False Negatives respectively, obtained after performing classification on test dataset. Thus, our evaluation metrics are defined as follows:

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (7)$$

5. Results

Table 2 shows the optimal hyperparameters obtained after performing 5 fold grid-search cross-validation on each of the mentioned classification algorithm. We use these sets of hyperparameters to evaluate the test data (X_{test}). Table 3 shows the performance of each of the algorithms (abbreviations shown in the table) on X_{test}. As we can notice, XGBoost Classifier gives the best performance, with F1 score of 89.1%, Specificity of 88.79% and Sensitivity of 87.17%.

TABLE 2. Optimal Hyperparameters

Algorithm	Hyperparameters

XGBoost Classifier (XBG)	learning_rate = 0.07
	n_estimators = 100
	max_depth = 3
	min_child_weight = 2

Random Forest Classifier (RF)	criterion = ”Gini”
	n_estimators = 200
	max_depth = 3
	max_features = 0.25

L1 penalized Logistic Regression (LR1)	C = 2.0
	penalty = ”l1”
	fit_intercept = True

L2 penalized Logistic Regression (LR2)	C = 1.75
	penalty = ”l2”

Gaussian Naïve Bayes (GNB)	var_smoothing = 1e-09

Multinomial Naive Bayes (MNB)	alpha = 1.0
	fit_prior = True

K Nearest Neighbour Classifier (KNN)	n_neighbors = 7
	weights = ”uniform”

Support Vector Classifier (SVC)	C = 1.5
	kernel = ”poly”
	degree = 4

TABLE 3. Results

Algorithm	Specificity	Sensitivity	F1 Score
XGB	88.79%	87.17%	89.1%
RF	73.91%	87%	84.2%
LR1	70.11%	71.37%	74.33%
LR2	69.32%	70.2%	73.33%
GNB	68.10%	69.32%	72.20%
MNB	64.2%	71.37%	68.84 %
KNN	57.60%	67.12%	61.02%
SVC	14%	56%	47.2%

6. Discussion and Conclusions

In comparison to other studies, our results have a wider clinical applicability. In a study conducted by [30], the best predictor model was based on Random Forest Classifier using HRV measured from 24 hour ECG in hypertensive patients, which showed sensitivity and specificity of 71.4% and 87.8% respectively. However, recording a 24 hour ECG is an impractical choice if screening for CAG has to be done at first point of contact. In [23], the authors developed a multi-parametric measure for classifying normal subjects (20 cases) from patients suffering from two types of CAD (64 cases), i.e. Angina Pectoris (AP) and Acute Coronary Syndrome (ACS). The measure was constructed using multiple discriminant analysis of several linear and non-linear parameters extracted from ECG signal. The authors obtained accuracies of 75.0%, 72.5% and 84.6% for control, AP, and ACS groups respectively. In [15], the authors used multilayer perceptron on HRV indices and got highest classification accuracy of 89.5% for prediction of CAD. In all the experiments mentioned above, small sample size (varying from minimum of 10 in each group to maximum of 90), and controlled recording conditions were considered and therefore, the real life utility of these algorithms is not clear. Moreover, a majority of these studies have reported accuracy, rather than sensitivity and specificity of the test, weakening the confidence in their clinical applicability.

We obtain a specificity and sensitivity of 88.79% and 87.17% using XGBoost Classifier. These results are superior to the avaliable screening tests [3] such as stress test (Specificity range is 70% - 80% and sensitivity range is 60% - 70%). ECG based machine learning solution is a highly suitable versatile screening tool. Compared to other screening and triage tests, obtaining ECG is simple, non-invasive and cost-effective, and therefore can be performed even at first point of contact.

Acknowledgements

This paper is supported by Department of Physiology, Central Research Laboratory, SKNMC & GH, Pune.

References

[1] https://www.who.int/health-topics/cardiovascular-diseases/

[2] Task Force Members, et al. ”2013 ESC guidelines on the management of stable coronary artery disease: the Task Force on the management of stable coronary artery disease of the European Society of Cardiology.” European heart journal 34.38 (2013): 2949-3003.

[3] Skelly, Andrea C., et al. ”Noninvasive testing for coronary artery disease.” (2016).

[4] REDWOOD, DAVID R., and STEPHEN E. EPSTEIN. ”Uses and limitations of stress testing in the evaluation of ischemic heart disease.” Circulation 46.6 (1972): 1115-1131.

[5] Mahmoodzadeh, Solmaz, et al. ”Diagnostic performance of electrocardiography in the assessment of significant coronary artery disease and its anatomical size in comparison with coronary angiography.” Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences 16.6 (2011): 750.

[6] Auer, Reto, et al. ”Association of major and minor ECG abnormalities with coronary heart disease events.” Jama 307.14 (2012): 1497-1505.

[7] Lu, Diyuan, et al. ”Towards Early Diagnosis of Epilepsy from EEG Data.” arXiv preprint arXiv:2006.06675 (2020).

[8] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. ”U-net: Convolutional networks for biomedical image segmentation.” International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[9] Beer, Tom, et al. ”Using Deep Networks for Scientific Discovery in Physiological Signals.” arXiv preprint arXiv:2008.10936 (2020).

[10] Duan, Hubert Haoyang. ”Applying supervised learning algorithms and a new feature selection method to predict coronary artery disease.” arXiv preprint arXiv:1402.0459 (2014).

[11] Kim, W-S., et al. ”A study on development of multiparametric measure of heart rate variability diagnosing cardiovascular disease.” World Congress on Medical Physics and Biomedical Engineering 2006. Springer, Berlin, Heidelberg, 2007.

[12] Lee, Heon Gyu, et al. ”Coronary artery disease prediction method using linear and nonlinear feature of heart rate variability in three recumbent postures.” Information Systems Frontiers 11.4 (2009): 419-431.

[13] Kora, Padmavathi, Ajith Abraham, and K. Meenakshi. ”Heart disease detection using hybrid of bacterial foraging and particle swarm optimization.” Evolving Systems 11.1 (2020): 15-28.

[14] Manini, Alex F., et al. ”Adverse cardiac events in emergency department patients with chest pain six months after a negative inpatient evaluation for acute coronary syndrome.” Academic Emergency Medicine 9.9 (2002): 896-902.

[15] Dua, Sumeet, et al. ”Novel classification of coronary artery disease using heart rate variability analysis.” Journal of Mechanics in Medicine and Biology 12.04 (2012): 1240017.

[16] Touboul, Pierre-Jean, et al. ”Carotid intima-media thickness, plaques, and Framingham risk score as independent determinants of stroke risk.” Stroke 36.8 (2005): 1741-1745.

[17] Scharnhorst, Volkher, et al. ”Rapid detection of myocardial infarction with a sensitive troponin test.” American journal of clinical pathology 135.3 (2011): 424-428.

[18] Mukhopadhyay, Sayantan, et al. ”Wavelet based QRS complex detection of ECG signal.” arXiv preprint arXiv:1209.1563 (2012).

[19] Maršánová, Lucie, et al. ”ECG features and methods for automatic classification of ventricular premature and ischemic heartbeats: A comprehensive experimental study.” Scientific reports 7.1 (2017): 1-11.

[20] Pan, Jiapu, and Willis J. Tompkins. ”A real-time QRS detection algorithm.” IEEE transactions on biomedical engineering 3 (1985): 230-236.

[21] van Gent, Paul, et al. ”HeartPy: A novel heart rate algorithm for the analysis of noisy signals.” Transportation research part F: traffic psychology and behaviour 66 (2019): 368-378.

[22] Berne, Robert M. ”Cardiovascular physiology.” Annual Review of Physiology 43.1 (1981): 357-358.

[23] Sharma, Pratyush N., and Kevin H. Kim. ”A comparison of PLS and ML bootstrapping techniques in SEM: A Monte Carlo study.” New perspectives in partial least squares and related methods. Springer, New York, NY, 2013. 201-208.

[24] Chawla, Nitesh V., et al. ”SMOTE: synthetic minority over-sampling technique.” Journal of artificial intelligence research 16 (2002): 321-357.

[25] Takahashi, Naomi, et al. ”Validity of spectral analysis based on heart rate variability from 1-minute or less ECG recordings.” Pacing and Clinical Electrophysiology 40.9 (2017): 1004-1009.

[26] Breiman, Leo. ”Random forests.” Machine learning 45.1 (2001): 5-32.

[27] Rish, Irina. ”An empirical study of the naive Bayes classifier.” IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[28] Chen, Tianqi, and Carlos Guestrin. ”Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[29] Olson, Randal S., et al. ”Data-driven advice for applying machine learning to bioinformatics problems.” arXiv preprint arXiv:1708.05070 (2017).

[30] Melillo, Paolo, et al. ”Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis.” PloS one 10.3 (2015): e0118504.