Arrhythmia Detection - A Machine Learning based Comparative Analysis with MIT-BIH ECG Data

Conference Paper · March 2019

DOI: 10.1109/I2CT45611.2019.9033665

4 authors:

Vishavpreet Singh
Academy of Scientific and Innovative Research

Suman Tewary
National Metallurgical Laboratory

Viren Sardana
Institute of Genomics and Integrative Biology

Hk Sardana
Central Scientific Instruments Organization

2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

Arrhythmia Detection - A Machine Learning based Comparative Analysis with MIT-BIH ECG Data

Vishavpreet Singh[1,2], Suman Tewary* [2] -*IEEE Graduate Member*, Viren Sardana[2] and H K Sardana[1,2]

[1]Academy of Scientific and Innovative Research (AcSIR-CSIO), Chandigarh, India
[2]CSIR-Central Scientific Instruments Organisation, Chandigarh, India
[*]Corresponding author email – suman.tewary@csio.res.in

***Abstract*— Cardiovascular arrhythmias are most common cardiac problem in the world. This work while focusing on development of automated detection and classification of arrhythmia using MIT-BIH database compared five machine learning algorithms with three different features. Pre-processing followed by beat detection is applied on one channel to get individual beats having PQRST window. Three features (viz. simple amplitude feature of 300 samples, area feature with non-overlapped sliding window, and area feature with overlapped sliding window) were used for five classifiers used in this study. It was observed that Artificial Neural Network with amplitude features gave best result with 99.59% accuracy which is comparable to state-of-the art methods.**

Keywords—ANN, Arrhythmia, Decision Tree, ECG, MIT-BIH, Machine Learning, Signal Processing, Random Forest, SVM

I. Introduction

Cardiovascular arrhythmias are very common form of disease which might cause cardiac arrest or even death [1]. Occurrences of some arrhythmias are very infrequent, hence the patient has to be monitored for a long time to identify the arrhythmia [2]. Manual diagnostic of arrhythmia from ECG is time consuming and often vary with the expertise of cardiologists. Various techniques for auto-detection of arrhythmia have been developed across the world to diagnose the patients [1], [3]–[6]. There is a need for robust detection of arrhythmias for prevention of further loss of life.

In this work, we have developed an auto-detection and classification approach for different types of arrhythmias. Normal and four different types of arrhythmias from benchmark dataset MIT-BIH are used for the development of approach. Beat detection followed by classification using three different features (viz. amplitude, area with non-overlapped sliding window and area with overlapped sliding window) were used to compare five different classifiers (viz. Support Vector Machine, Decision Tree, Random Forest, Naïve Bayes and Artificial Neural Network). It is demonstrated that the Artificial Neural Network with amplitude feature is performing the best amongst all with 99.59% accuracy.

II. State-of-the-art

During the acquisition of ECG data, various noises and physiological artifacts affect the signal [5]-[7]. Therefore, a pre-processing algorithm is required to fine-tune the data. Previous work includes various ECG features such as R-time domain, frequency and morphological features etc. [1]. Thereafter, various machine learning algorithms in recent works trained to identify arrhythmia or its types. Most of the work used support vector machine (SVM) [8], artificial neural network (ANN) [9], Decision Trees [10] and Random Forest classifiers [11] etc. Deep learning methodology is applied to differential different classes of arrhythmia [3], [4]. Recently artificial neural network (ANN) and deep learning architecture are becoming prominent in decisive systems. A CNN based 34-layered deep learning framework trained on patient's data to achieve cardiologist level performance for arrhythmia detection to classify 14 classes in real time [3]. To automate the task of arrhythmia detection, another work to classify normal (N), right bundle branch block (RBBB) and paced beat by transfer learning approach using AlexNet [4]. In another work, ventricular arrhythmia and non-ventricular arrhythmia classes were diagnosed with the help of personalized two and three features with support vector machine [5]. A system based on nonlinear analysis of variational modes of ECG was presented with two novel features viz. variational mode sample entropy and variational mode distribution entropy, followed by multiclass support vector machine classifier with radial basis function [6]. Multinomial logistic regression for detecting arrhythmia achieves 93.13 % using R-R interval based features [1].

III. Materials and Methodology

In this section, the description of database, selection of types of arrhythmias, beat detection and classification approach is discussed.

A. Benchmark Dataset

MIT-BIH dataset which is publicly available at *PhysioNet* [12] is used in this work. The dataset which was prepared in five years consists of 48 records of 2 channel ECG digitized at a rate of 360 samples per second. 25 male subjects aged from 32 to 89 years and 22 female subjects aged from 23 to 89 years were involved in the development of this database. 60% of the total subjects were inpatient. The beats which could be identified as QRS were annotated and are about 109,000 in numbers. The six records in dataset contain 33 beats which remain unclassified because of inability to reach agreement on beat types [13].

This work is focused on classification of five classes of arrhythmia as Normal (N), Paced Beat (/), Right Bundle Branch Block Beat (R), Left Bundle Branch Block (L) and

978-1-5386-8075-9/19/$31.00 ©2019 IEEE

Premature Ventricular Beat (V). From the two channel ECG data, we considered only first channel of ECG for this work. ECG may be affected from noise, so we involved pre-processing steps to refine signal data. Then we segmented the signal into 300 sample points for further feature computation units as explained in the next section.

B. Data Pre-processing

The ECG signal may contain various noises due to baseline wandering noises and respiratory muscle noise etc. A sliding window of 300 sample size to compute mean of ECG and then subtracted from each of these samples to shift their mean to zero in that window. After application of mean shift, an average filter with kernel size 10 was used to smooth signal within that sliding window of size 300. This step generated the pre-processed signal for further preparation of data used for beat detection followed by classification.

C. Data Preparation

For peak detection, Pan-Tompkins's algorithm is applied [14]. After detecting peaks, the closest matching annotated peaks were taken into consideration for our studies. Based on the detected peaks, a window of 300 samples around the peak (P-149 to P+150 samples) was segmented. This window of 300 samples was assumed as amplitude features. From the too many beat types, we considered only those beats who were enough in numbers i.e. more than 5000 for this work. We took 5 types of beats as they were more than 5000 in numbers. We prepared data by taking equal numbers of beats of each type to balance the dataset for performing comparisons.

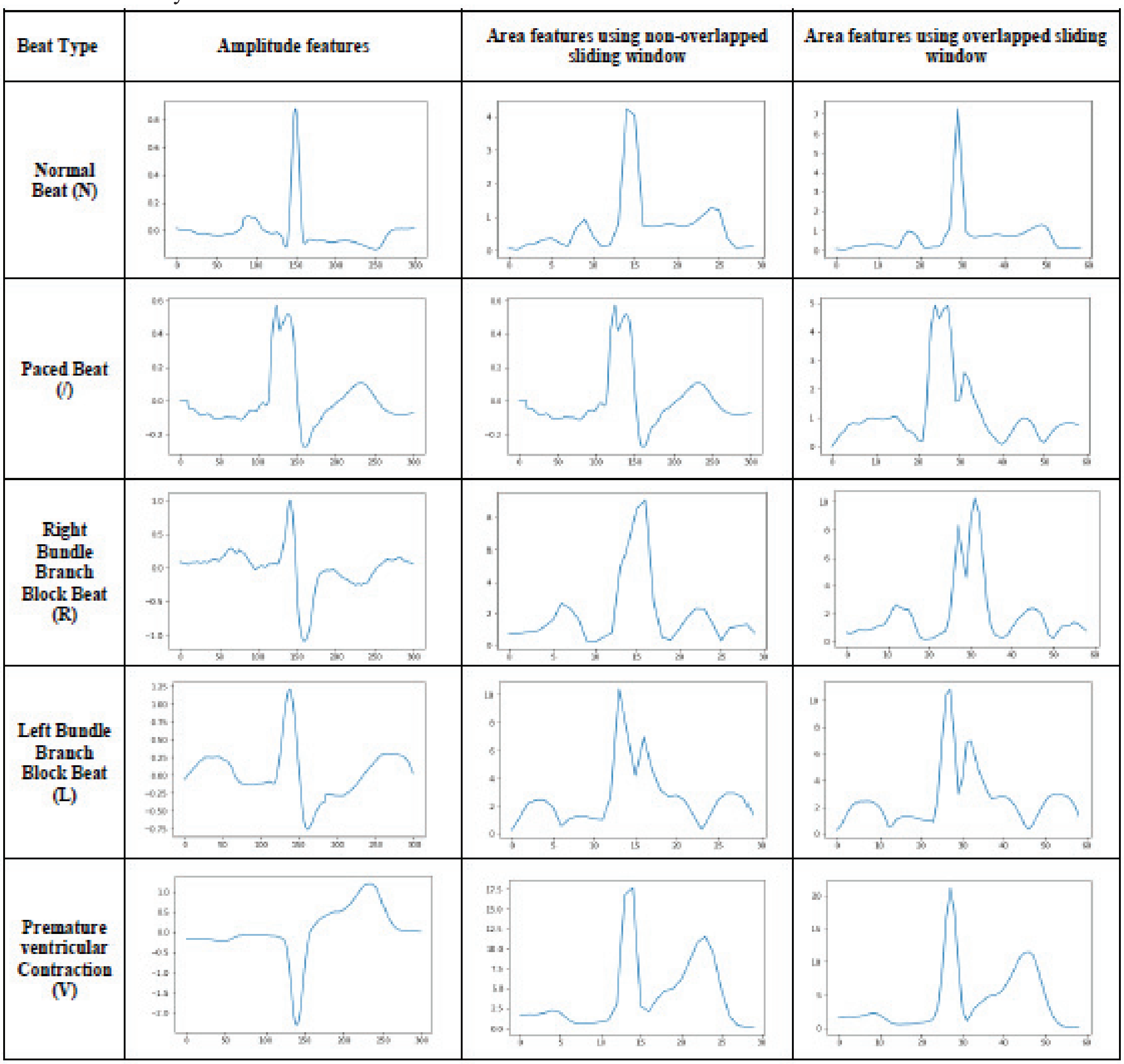

Fig. 1: Normal beat, 4 types of abnormal and their corresponding features

IV. BEAT FEATURE EXTRACTION

We extracted the same feature as two ways by considering two different types of sliding window in ECG Signal. We considered area features with an intuition that QRS curve encompasses more area and a pattern of QRS will alter in different beats and area under the curve will be changed accordingly. We considered 3/8 Simpson rule as it gives promising results while calculating the area under the curve.

$$\int_a^b f(x)dx = \frac{3h}{8}\left[f(a) + 3\sum_{i\neq 3k}^{n-1} f(x_i) + 2\sum_{j=1}^{\frac{n}{3}-1} f(x_{3j}) + f(x_n)\right]$$

Where $k \in N_0$

A. Non-overlapped sliding window

A sliding non-overlapped window of size 10 to calculate the area for 300 data samples each prepared in previous section C. We applied Simpson rule on absolute values over each sliding window of 10 samples. Thus, we formed a feature vector of 30 features for each sample. Figure 1 in second column represents area under the curve for non-overlapped sliding window.

B. Overlapped sliding window

An overlapped sliding window of size 50 and stride 25 was applied to calculate the area of 300 data samples prepared in previous section III.C. Again, we applied Simpson rule on absolute values over an overlapped window of 50 samples. Thus, a feature vector of 11 features prepared for training shows graphical representation for area under the curve by considering overlapped sliding window.

TABLE I: CONFUSION MATRIX, CLASSIFICATION REPORTS AND OVERALL RESULT OF FIVE CLASSIFIERS

A. Best confusion matrix with ANN model with amplitude features

Predicted/Actual	N	/	R	L	V
N	1463	0	0	1	1
/	2	1556	0	0	1
R	0	0	1519	0	9
L	0	0	1	1506	8
V	0	2	1	5	1425

B. Worst confusion matrix with Naïve Bayes classifier model with overlapped area features

Predicted/Actual	N	/	R	L	V
N	1411	2	100	3	0
/	620	830	4	23	42
R	21	3	1373	121	15
L	391	22	186	792	98
V	10	309	39	215	870

C. Classification Report for ANN model with amplitude features

	PRECISION	RECALL	F1-SCORE	SUPPORT
N	1.00	1.00	1.00	1465
/	1.00	1.00	1.00	1559
R	1.00	0.99	1.00	1528
L	1.00	0.99	1.00	1515
V	0.99	0.99	0.99	1433
AVG/ TOTAL	1	1	1	7500

D. Classification Report for Naïve Bayes classifier model with overlapped area features

	PRECISION	RECALL	F1-SCORE	SUPPORT
N	0.58	0.93	0.71	1516
/	0.71	0.55	0.62	1519
R	0.81	0.9	0.85	1533
L	0.69	0.53	0.6	1489
V	0.85	0.6	0.71	1443
AVG/ TOTAL	0.72	0.7	0.7	7500

E. Overall comparison for five classifiers and different features in terms of accuracies

S. No.	Classifiers	Amplitude features	Area features with Non-overlapped sliding window	Area features with overlapped sliding window
1	SVM with RBF Kernel	98.05%	**98.97%**	85.52%
2	Decision Tree	97.61%	96.15%	95.68%
3	Random Forest	99.08%	98.43%	**97.73%**
4	Naïve Bayes	83.61%	73.49%	70.35%
5	Artificial Neural Network	**99.59%**	97.91%	95.80%

V. Classification of Beats

On a total of 25000 beat samples (5000 for each class), we split data for training and testing i.e. 70% for training data and 30% for testing. Five different models for our study are described below:

Support vector machine (SVM): SVM is a class of supervised machine learning algorithms [8]. We applied four different variants of SVM i.e. Linear SVM, SVM with RBF kernel, SVM with polynomial kernel of degree 3 and Sigmoid SVM. It was noticed that SVM with RBF gives satisfactory results among others.

Decision tree classifier: A decision tree is classification algorithm which contains nodes that form a directed tree with a node called root that has no incoming edge [10].

Random forest classifier: Random forest classifier, also known as random decision forest, is a set of various decision trees and its capacity can be arbitrarily increased or decreased to improve accuracy for both training and unseen data [11]. We implemented Random forest with 10 trees which gave significant accuracy.

Naïve Bayes classifier: This classifier is a simple probabilistic classifier based on Bayes theorem. There are many variants of Naïve Bayes classifiers viz. Gaussian naïve Bayes, Multinomial naïve Bayes etc. In this study Gaussian naïve Bayes function is used.

Artificial neural network (ANN): Basic element of processing of neural networks are called neurons which learn by adjusting weights in accordance to the data to be learnt [9]. We used ANN with input layer of 300 nodes, 4 hidden layers with 152, 300, 300 and 152 nodes, and one output node with 5 nodes.

VI. Results and Discussions

The algorithm is developed and tested on a HP Z6 Workstation with 32 GB RAM and 8 GB *NVIDIA* P4000 graphics card. *Python* libraries *sklearn* and *keras* were used for the development to make it an open-source package.

The performance of classifiers is compared and the results shown in Table 1. It suggests that ANN with pre-amplitude features of 300 samples gives best result with 99.59% accuracy. With area features with non-overlapping sliding window, SVM with RBF performed best with 98.97%. Likewise, area features with overlapped sliding window when fed to Random Forest with 10 trees gives third best result with 97.733%.

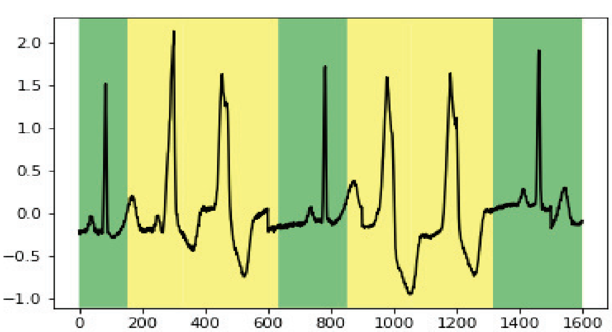

Fig. 1: Detection of Normal and Premature Ventricular beat with artificial neural network using amplitude features

The generated models are tested on the data for final validation. A sample data is used with normal and abnormal beats for the testing of model. **Fig. 1** shows segmentation and classification of normal and Premature Ventricular beat with artificial neural network using amplitude features. Different colour representation is used to highlight the normal and abnormal beats. In this example green and yellow colours represent normal and Premature Ventricular beats respectively.

An automated beat detection and classification system for the quantification of types of arrhythmia has been shown here. Various signal features could be targeted for further development.

VII. Conclusion

In this work, we have demonstrated an automated approach for detection of beats followed by classification. Five different classifiers were compared for five classes of data (normal and four types of arrhythmias) with three different features. It is observed that Artificial Neural Network with amplitude feature gives the best results with 99.59% accuracy. In future, we will target to develop a comprehensive module for mobile platform.

Acknowledgment

Authors wish to acknowledge funding received by CSIR-CSIO under CSIR mission on *Intelligent Systems (IS) – Intelligent Technologies and Solutions*. Vishavpreet Singh is thankful to the financial Support by CSIR-Senior Research Fellowship program.

References

[1] O. Behadada, M. Trovati, M. A. Chikh, N. Bessis, and Y. Korkontzelos, “Logistic regression multinomial for arrhythmia detection,” in *Foundations and Applications of Self* Systems, IEEE International Workshops on*, 2016, pp. 133–137.

[2] R. A. Sanders, T. A. Kurosawa, and M. D. Sist, “Ambulatory electrocardiographic evaluation of the occurrence of arrhythmias in healthy Salukis,” *J. Am. Vet. Med. Assoc.*, vol. 252, no. 8, pp. 966–969, 2018.

[3] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, “Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks,” *arxiv*, 2017.

[4] A. Isin and S. Ozdalili, “Cardiac arrhythmia detection using deep learning,” in *Procedia Computer Science*, 2017.

[5] P. Cheng and X. Dong, “Life-threatening ventricular arrhythmia detection with personalized features,” *IEEE Access*, 2017.

[6] A. Chetan, R. K. Tripathy, and S. Dandapat, “A Diagnostic System for Detection of Atrial and Ventricular Arrhythmia Episodes from Electrocardiogram,” *J. Med. Biol. Eng.*, 2018.

[7] J. Park, M. Kang, J. Gao, Y. Kim, and K. Kang, “Cascade Classification with Adaptive Feature Extraction for Arrhythmia Detection,” *J. Med. Syst.*, 2017.

[8] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intell. Syst. their Appl.*, vol. 13, no. 4, pp. 18–28, 1998.

[9] A. Abraham, “Artificial neural networks,” *Handb. Meas. Syst. Des.*, 2005.

[10] L. Rokach and O. Maimon, “Top-down induction of decision trees classifiers-a survey,” *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 35, no. 4, pp. 476–487, 2005.

[11] T. K. Ho, “Random decision forests,” in *Document analysis and recognition, 1995., proceedings of the third international conference on*,

1995, vol. 1, pp. 278–282.
[12] G. B. Moody, R. G. Mark, and A. L. Goldberger, "PhysioNet: a web-based resource for the study of physiologic signals," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 70–75, 2001.
[13] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, 2001.
[14] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng*, vol. 32, no. 3, pp. 230–236, 1985.