Infinite Feature Selection Based Arrhythmia Classification Using Semantic Features and Random Forest

Vedavathi Gauribidanur Rangappa[1]*   **Sahani Venkata Appala Varaprasad Prasad[1]**   **Alok Agarwal[1]**

[1]Lingaya's Vidyapeeth, India
* Corresponding author's Email: vedavathi.gr@gmail.com

Abstract: In this research study, arrhythmia classification was assessed by using Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) arrhythmia database. After signal acquisition, semantic feature extraction (combination of statistical, entropy, linear and non-linear features) was used for extracting the feature values from acquired electrocardiogram (ECG) signals. Here, twenty-six features were combined to perform feature extraction on acquired ECG signals. Then, the infinite feature selection was used for rejecting the irrelevant feature vectors or selecting the optimal feature subsets. After selecting the optimal feature vectors, a binary classifier: random forest was used to classify the eight arrhythmia heartbeat conditions: Normal sinus rhythm (N), Sinoauricular Heart Block (SHB), Premature Ventricular Contraction (PVC), Auricular Fibrillation (AF), Left Bundle Branch Block (LBBB), Premature Atrial Contraction (PAC), Supra-Ventricular-Tachycardia (SVT) and Right Bundle Branch Block (RBBB). This research work includes numerous advantages; earlier detection of arrhythmia diseases, assists clinicians during surgery, and cost efficient compared to the existing machine learning methodologies. The experimental outcome shows that the proposed methodology effectively distinguishes the heartbeat conditions by means of sensitivity, specificity, Negative Predictive Value (NPV), Positive Predictive Value (PPV) and accuracy. The proposed methodology enhances the classification accuracy up to 1.4-1.5% compared to the existing methodologies.

Keywords: Arrhythmia classification, Electrocardiogram, Infinite feature selection, Random forest, Semantic feature extraction.

1. Introduction

ECG is an inexpensive and non-invasive diagnostic tool that is extensively utilized in numerous applications. ECG recognizes the changes in electrical activity of the heart and also it extracts necessary physiological information, which is utilized for analysing the heart function [1-3]. The ECG is a periodic signal that is composed of a wave sequence (P, Q, R, S, and T waves), which repeat periodically in time [4, 5]. Cardiac arrhythmia occurs intermittently in early stages of heart disease, which is difficult in early diagnosis. The arrhythmic events lead to sudden dead, occurring less than an hour after symptom onset [6, 7]. Thus, a large part of the biomedical research spectrum is directed towards developing ECG diagnostic equipment to contribute early diagnosis of heart disease treatment at the early stages [8]. Numerous algorithms are developed for automatic detection of ECG beats by different researchers. In existing papers, the researchers utilized dissimilar classification approaches and features for classifying the different arrhythmia heartbeat conditions. The major issues in the existing studies are; requires manual investigation, the possibility of missing value is high, and time consuming/tedious for physicians [9, 10]. To overcome these concerns, automatic arrhythmia classification and detection systems are required. The automatic arrhythmia detection systems deliver accurate and reliable recognition approach for a wide range of arrhythmia types.

The major contribution of this research work is to accomplish arrhythmia classification using a superior approach, which enables the management

of large ECG datasets, for which a new system was developed. At first, ECG signals were collected from the dataset: MIT-BIH arrhythmia. The unwanted artifacts in the collected ECG signals were eliminated by employing normalization and sixth order zero phase Butterworth filter. After pre-processing, Pan-Tomkins and windowing techniques were used for peak detection. Then, semantic feature extraction was applied at the peak detected ECG signals for extracting the feature values. **Novelty:** Semantic feature extraction was the procedure of obtaining feature subsets from the set of data inputs by the rejection of redundant and irrelevant features. After obtaining the feature information, the infinite feature selection was utilized to select the optimal feature subsets. Output of infinite feature selection specifies, which features of the ECG signals were essential in describing the dataset signals. These optimal feature values were given as the input for random forest to classify the normality and abnormality of arrhythmia. While working with non-uniform data, random forest is a good choice for classification.

This research paper is composed as follows. Section 2 presents a broad survey of recent papers in arrhythmia classification. In section 3, an effective supervised system is developed (semantic feature extraction is presented with infinite feature selections and random forest classifier) for arrhythmia classification. In section 4, quantitative and comparative analysis of proposed and existing methodologies is presented. The conclusion is made in section the section 5.

2. Literature review

Numerous research methodologies are suggested by the researchers in ECG arrhythmia classification. The brief evaluation of a few essential contributions to the existing literatures is presented in this section.

G. Sannino, and G. De Pietro, [11] developed a new deep learning system for ECG beat classification. In this literature, Deep Neural Network (DNN) was developed using deep learning library from Google and tensor flow framework. The developed deep learning system comprises of seven hidden layers with 5, 10, 30, 50, 30, 10 and 5 neurons. This experiment was conducted on a publicly available dataset; MIT–BIH arrhythmia and compared the experimental outcomes with the recent scientific literature. The final outcome shows that the developed system was more effective by means of accuracy, sensitivity and specificity. The gradient in DNN was a major issue, which was unstable and tends to either vanish or explode in earlier layers.

R.R. Linhares, [12] presented an automatic system for arrhythmia classification, which was the combination of smooth detrended fluctuation analysis (SDFA) and principal of wavelet shrinkage scaling analysis to represent the time series correlation properties. This research was carried out on the online databases; MIT-BIH arrhythmia dataset and MIT-BIH Normal Sinus Rhythm dataset. The extensive experiments were conducted and the efficiency of the developed methodology was verified by means of mean square error. The developed system fails to achieve better classification, when the acquired dataset size was high.

M. Mohanty, S. Sahoo, P. Biswal, and S. Sabut, [13] developed a new system for classifying the ventricular fibrillation arrhythmias and ventricular tachycardia using statistical, spectral, and temporal features. In this literature, MIT-BIH malignant ventricular ectopy and CU ventricular tachyarrhythmia datasets were used for evaluating the performance of the developed system. The extracted features were given as the input for C4.5 classification methodology for classifying the two heartbeat conditions. The experimental section confirmed that the developed system was more effective than the existing methodologies by means of accuracy, sensitivity and specificity. Here, the computational complexity of the developed system was high, while performing with "n" number of features.

A.F. Khalaf, M.I. Owis, and I.A. Yassine, [14] developed a new CAD system to classify five heartbeat conditions; N, PVC, PAC, RBBB, LBBB. At first, statistical features followed by Principal Component Analysis (PCA) and fisher score were used for feature space reduction. The extracted feature vectors were given as the input for support vector machine with linear kernel for classifying the five heartbeat conditions. The classifier performance was determined by using classification accuracy sensitivity, specificity, accuracy, PPV and NPV. The experimental phase confirmed that the developed system delivered better classification accuracy compared to the existing classification approaches. The developed system was only suitable for uniform pattern datasets not for non-uniform pattern datasets.

J.S. Wang, W.C. Chiang, Y.L. Hsu, and Y.T.C. Yang, [15] presented an effective ECG arrhythmia classification system for classifying eight different heartbeat conditions. Initially, linear discriminant analysis and PCA were used to identify the essential features from ECG heart beats and also enhances the classification accuracy of the classifier. With the selected features, the Probabilistic Neural Network

(PNN) classifier is used to classify the eight different heartbeat conditions; N, PVC, paced beat, RBBB, LBBB, atrial premature contraction, ventricular flutter wave, and ventricular escape beat. The average classification accuracy of developed scheme was better than the other existing methodologies. Still, the developed system requires an effective feature extraction methodology for further improving the classification accuracy.

A. Gutiérrez-Gnecchi, R. Morfin-Magana, D. Lorias-Espinoza, A. del Carmen Tellez-Anguiano, E. Reyes-Archundia, A. Méndez-Patiño, and R. Castañeda-Miranda, [16] developed an effective arrhythmia classification system to classify eight heartbeat conditions such as, N, AF, PAC, LBBB, RBBB, PVC, SHB, and SVT. This system used wavelet transform on the basis of quadratic wavelets for obtaining a fiducial marker array and also to identify individual ECG waves. In this research study, the classification was done by employing PNN classifier for classifying the eight heart beast conditions. This experiment was carried out on an online database (i.e., MIT-BIH arrhythmia dataset) to validate its result by means of classification accuracy. Here, the developed system did not concentrate on the signal pre-processing, which was considered as one of the major drawback.

H. Ebrahimnezhad, and S. Khoshnoud, [17] developed a new system for classifying four heartbeat conditions: N, APC, RBBB and LBBB. Initially, linear predictive coefficients of QRS peaks were determined by using multi-resolution wavelet analysis. These coefficients were given as the input for PNN classifier in order to classify the four heartbeat conditions. In this research study, the effectiveness and efficiency of the developed system was evaluated by employing MIT-BIH arrhythmia dataset. A major issue in the developed system was high computational time related to the other conventional systems.

To overcome the above mentioned drawbacks, a supervised methodology (semantic features with infinite feature selection and random forest classifier) is implemented for enhancing the performance of arrhythmia classification.

3. Proposed methodology

In this research study, a new CAD system is developed for enhancing the effectiveness of cardiac arrhythmia classification. The proposed method for classifying eight arrhythmia heartbeat conditions from the acquired ECG signals contains six major

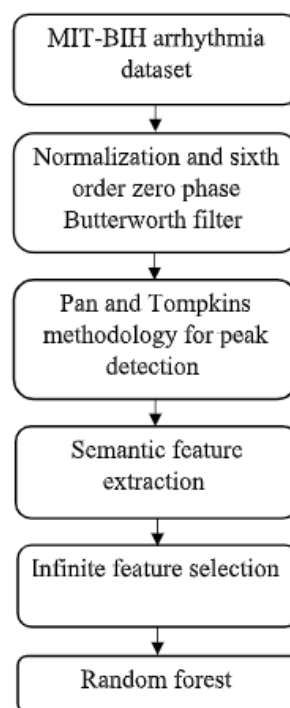phases: signal collection, signal pre-processing,

Figure. 1 Block diagram of proposed methodology

peak detection, feature extraction, feature selection and classification. The proposed system block diagram is denoted in the Fig. 1. The brief description about the proposed methodology is detailed below.

3.1 Data collection

In the initial phase of arrhythmia classification, ECG signals are collected from a standard benchmark dataset; MIT-BIH arrhythmia dataset. It contains forty-eight half-hour excerpts of two-channel ambulatory ECG recordings. In this dataset, the ECG signals are recorded from forty-eight subjects from BIH Arrhythmia Laboratory between 1975 and 1979. Randomly, twenty-three recordings are selected from a set of 4000 records in that 40% are outpatients and 60% are inpatients at Boston's Beth Israel Hospital. The remaining twenty-five recordings are taken from the same dataset in a small random sample. The respective recordings are digitized at 11-bit resolution over a 10 mV range. After collecting the ECG data, an important step in arrhythmia classification is pre-processing of acquired ECG signals. The sample normal and abnormal signal of MIT-BIH arrhythmia dataset is denoted in the Fig. 2.
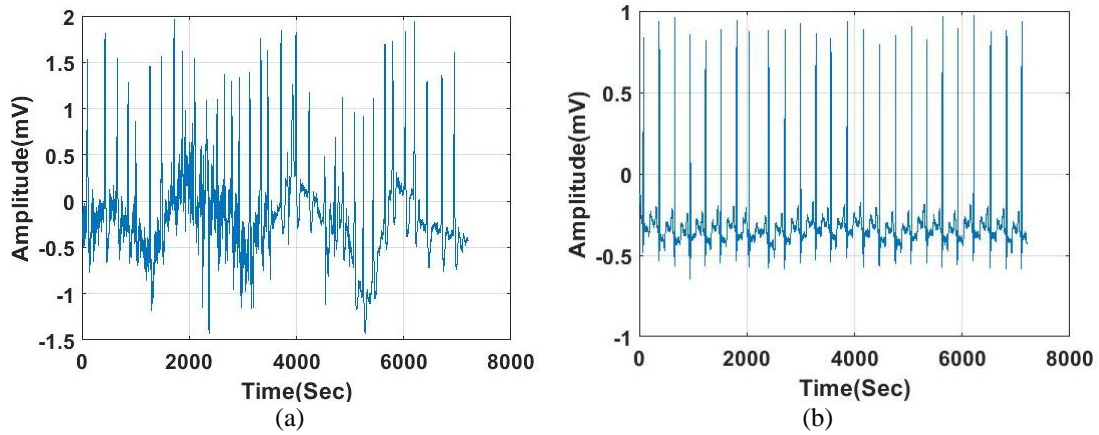
Figure. 2 Sample signal of MIT-BIH arrhythmia dataset: (a) sample normal ECG signal and (b) sample abnormal ECG signal
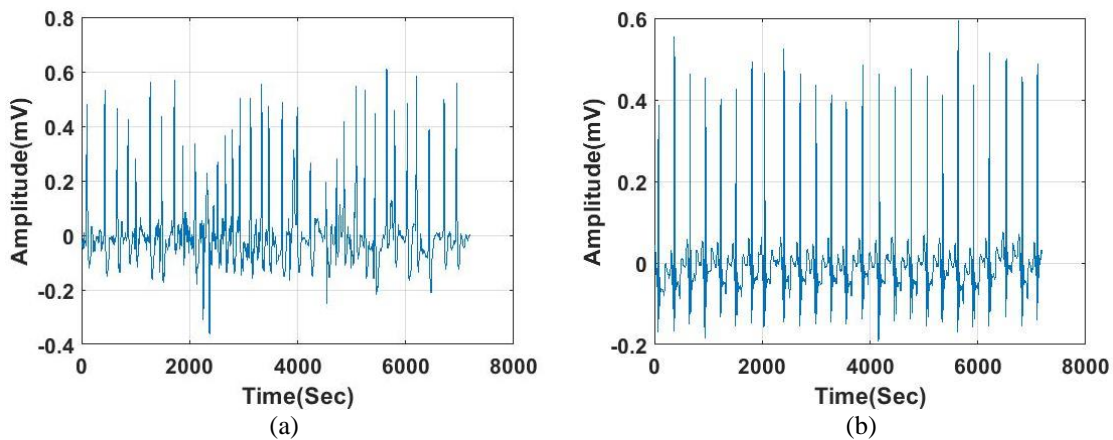

Figure. 3 Filtered MIT-BIH arrhythmia dataset sample signals: (a) sample pre-processed normal ECG signal and (b) sample pre-processed abnormal ECG signal

3.2 Pre-processing of ECG signals

In this sub-section, normalization and 6th order butter worth filter are used for signal pre-processing. Normalization is a statistical operator that is used to scale the heterogeneous set of ECG data. It is accomplished to set the maximum amplitude to 1 or -1 and the baseline to zero. After normalizing the ECG signal, 6th order butter worth filter is used for de-noising the signal. Usually, the ECG signals are captured from the ECG equipment's, so it majorly consists of baseline wander noise. The 6th order butter worth filter is very effective to remove the baseline wander noise (frequency range over 0.5 Hz), and also significantly enhance the quality of normalized ECG signal. The general formula of 6th order butter-worth filter is given in the Eq. (1),

$$G^2(W) = |H(jw)|^2 = \frac{G_0^2}{1+\left(\frac{jw}{jw_C}\right)^{2n}}, n = 6 \qquad (1)$$

Where, W_C is stated as the cut off frequency, G_0 is represented as the DC gain and n is denoted as the filter order. Then, the preprocessed ECG signals are used for peak detection using Pan and Tompkins algorithm and windowing approach. The filtered MIT-BIH arrhythmia dataset sample signals are represented in the Fig. 3.

3.3 Peak detection using Pan-Tompkins and windowing approach

In signal processing, peak detection is a very essential phase, because it produces faster results for users and also allows automatic signal processing. The specific bands are obtained by moderating the low spectral frequencies of ECG signals. These bands have improved quality of frequency ranges compared to raw data. Here, Pan and Tompkins algorithm finds the peaks using filtering and pre-processing techniques. The Pan and Tompkins algorithm comprise of four stages such as, filtering, derivative, squaring and R-Peak for peak detection.
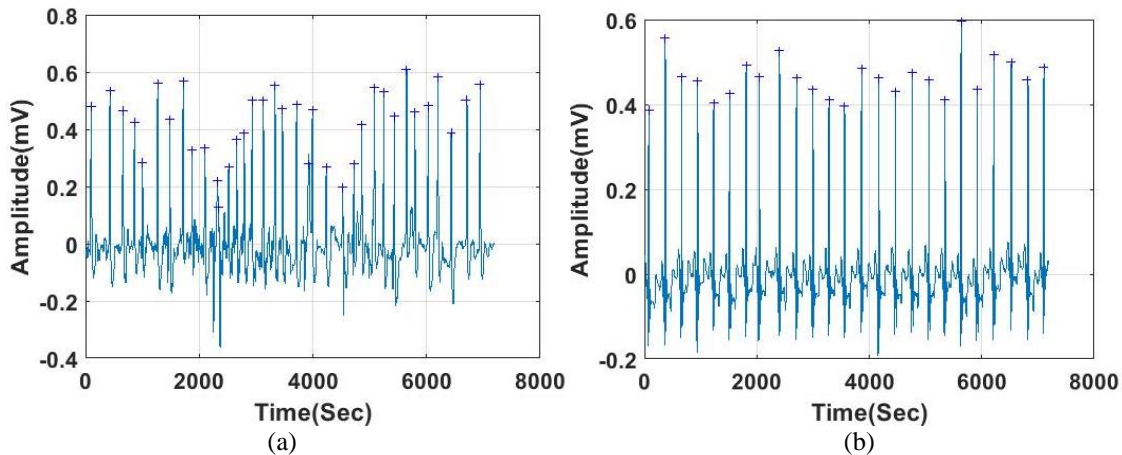
Figure. 4 Normal and abnormal signal after applying Pan and Tompkins algorithm: (a) peak detection of sample normal ECG signal and (b) peak detection of sample abnormal ECG signal

Filtering: Initially, band-pass filter is used for reducing the influence of baseline wander, muscle noise, 60Hz interference, and interference in T-wave. The desirable range of pass-band 4-15 Hz is used for maximizing the QRS complex energy. The low-pass and high-pass filters are cascaded to form Band-pass filter, where the low-frequency interference and higher frequency artifacts are removed automatically.

Derivative: In derivative block, smoothing technique is used to acquire the slope of QRS complex noise-free ECG signal. Here, the smoothing technique determines the position of wave appearance, amplitude of the peak, and width of each peak with the particular interval of time.

Squaring function: After derivative block, the output ECG signals are squared sequentially. The derivative ECG output is amplified non-linearly, and the QRS complex are emphasized by using the Eq. (2).

$$y(nT) = [x(nT)]^2 \qquad (2)$$

Where, $x(nT)$ is represented as input ECG signal and $y(nT)$ is denoted as squared input signal.

Peak detection: Steps involved in peak detection using Pan and Tompkins algorithm are listed below,

Step 1: Select the value and store it to temp local maximum.

Step 2: If newer point value is greater than temp local maximum, then store it to temp local maximum.

Step 3: If newer point value is smaller than half of temp local max, a peak is identified and detected. Otherwise, go to step 2. The normal and abnormal signal after applying Pan and Tompkins algorithm is given in the Fig. 4.

Additionally, windowing approach is applied in the Pan and Tompkins algorithm output to further improve the accuracy of arrhythmia classification. The windowing approach extracts the features of R peak location as the primary peaks by moving some samples to left side and right side of the signal. After performing windowing approach, semantic feature extraction is carried-out on the peak detected ECG signal. Fig. 5 represents the peaks of eight heartbeat conditions.

3.4 Semantic feature extraction

Generally, feature extraction is defined as the action of mapping the ECG data from data space to the feature space and it also transforms the large redundant data into a reduced data representation. This process helps to decrease the complexity of the system. In this research study, feature extraction is performed on the basis of statistical, energy, entropy, linear and non-linear features.

The statistical features comprise of mean, standard-deviation, skewness, minimum, maximum, variance, moments, kurtosis, and joint moment time frequency distribution. Entropy features contain Renyi entropy, Shannon entropy, permutation entropy, differential entropy, threshold entropy, tsallis entropy, sure entropy, log energy entropy, and sample entropy. In addition, the linear and non-linear features comprise of higher order spectra features (cumset 1, 2, and 3), Yule-Walker, burg's, teager's energy operator, root sum squared value and Peak-magnitude to root mean square ratio are also used for extracting the features from ECG signals. The advantage of using several features helps to recognize the data despite significant clutter and occlusion. Among these twenty-six features, ten features are selected as an optimal feature using infinite feature selection. Those ten features are

skewness, minimum, teager's energy operator, cumset 2, sure entropy, log energy entropy, tsallis entropy, Yule-Walker, differential entropy, and standard-deviation. A brief description about infinite feature selection is given in the section 3.5.
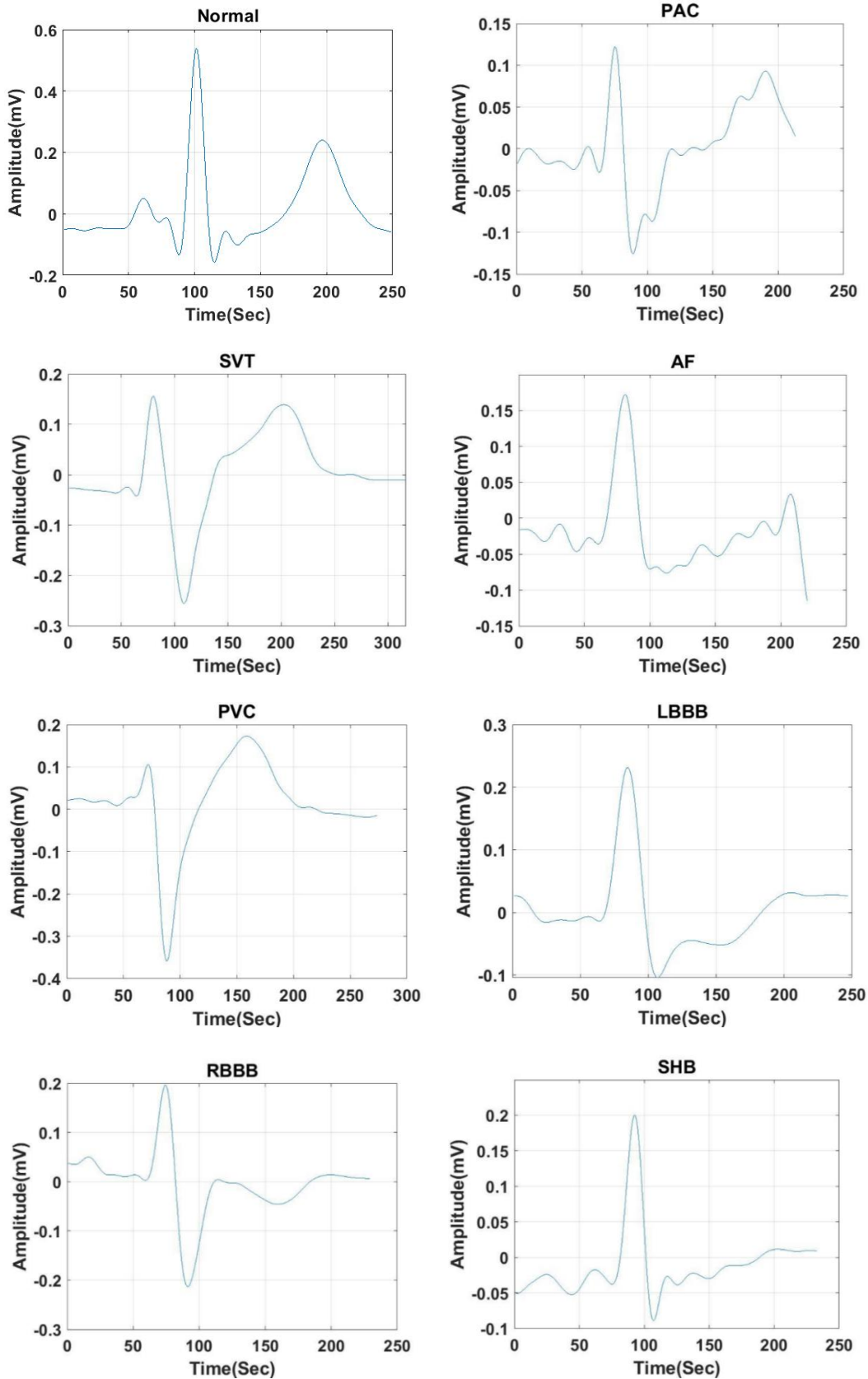

Figure. 5 Peaks of eight heartbeat conditions

3.5 Infinite feature selection

Feature selection is a high-level process that identifies the optimal subsets of ECG data on the basis of a particular criterion. In feature selection, mutual information between the features are calculated for identifying the optimal features that helps to decrease the computational effort. In this research, infinite feature selection is used for selecting the optimal feature vectors. An initial step of infinite feature selection is selecting an appropriate length l and energy scores $e_l(i)$ for each feature vectors, which is given in the Eq. (3).

$$e_l(i) = \sum_{j \in v} \sum_{p \in p_{i,j}^l} \prod_{k=0}^{l-1} a_{vk,vk+1} = \sum_{j \in v} A^l(i,j) \quad (3)$$

Where, $P_{i,j}^l$ is represented as the set of all paths of length l between the nodes j and i, A^l is defined as the power iteration of matrix, A and v is denoted as vertices of the feature vectors. By extending the path length to infinity, the probability of feature vectors is normalized. Therefore, a new energy score for each feature f_i considers all path lengths including infinity, which is mathematically defined in the Eq. (4).

$$e(i) = \left[\left(\sum_{l=0}^{\infty} A^l \right) - I \right] \bar{1} \quad (4)$$

Where, I is represented as identity matrix, and $\bar{1}$ denoted as a column vector of ones.

In matrix algebra, $\sum_{k=0}^{\infty} X^l$ is denoted as the geometric series of matrix X. This series converges to $(I - X)^{-1}$, if $p(X) < 1$, where $p(X)$ is denoted as the maximum magnitude of the eigenvalue of X. Using this property, the regularized energy score for each feature vector is determined using the Eq. (5).

$$e'(i) \left[\left(\left(\sum_{l=o}^{\infty} r^l A^l \right) - I \right) \bar{1} \right]_i = [((I - rA)^{-1} - I) \bar{1}]_i \quad (5)$$

The computation of power iterations of matrix in the Eq. (3) is reduced by computing $((I - rA)^{-1} - I)$. The obtained feature values are given as the input for multi-objective classifier; random forest.

3.6 Classification using random forest

Random forest is one of the emerging classification methods with the advantage of limited usage of resources, while training the large scale ECG data. Random forest is a non-parametric pattern classifier and also one of the most extensively utilized un-supervised machine learning classification approach that decreases the problem of probability density complexity. In random forest classifier, each tree is observed as an individual classifier and the classification outcome is chosen by all the decision trees. In this research study, the tree length is fixed at five. To develop a random forest classification approach, the growth rules of each tree are determined as follows.

If N is the number of training set, then randomly sample the N data from the training set M is the dimension of input features. If $m (m < M)$, select the sub-features from the original feature vectors. Then, m feature variables are chosen randomly out of the M features and the best split on these $m-$ dimensional features are utilized to split the node. Each tree keeps growing until all these training samples are totally separated without pruning. As revealed, the forest error rate depends on two aspects.

- The correlation between the two trees in the forest: More correlation leads to higher error rate and less correlation leads to smaller error rate.
- The strength of each individual tree in the forest: Increases of tree strength reduces the forest error rate.

4. Experimental result and discussion

In this experimental investigation, the proposed methodology was experimented using MATLAB (version 2018a) with 3.0 GHZ Intel i7 processor, 1TB hard disc and 16 GB RAM. For evaluating the effectiveness of proposed methodology, the performance of proposed system was compared with the existing approaches (wavelet transform + PNN, and linear predictive coefficients + PNN) on the reputed dataset: MIT-BIH arrhythmia database. The proposed methodology performance was evaluated by means of sensitivity, NPV, accuracy, PPV and specificity.

4.1 Performance metrics

Generally, performance measure is defined as the regular measurement of outcomes and results that develops a reliable information about the efficiency and effectiveness of proposed system. The relationship between the input and output values of proposed system understood by using suitable performance metrics like sensitivity, NPV, accuracy, PPV and specificity. The general formula

Table 1. Performance evaluation of proposed methodology using sensitivity and specificity

Classifier	Classes	TP	FP	FN	TN	Sensitivity (%)	Specificity (%)
Random forest	N	10	0	0	70	100	100
	AF	9	3	1	67	95.71	90
	PAC	10	1	0	69	98.57	100
	LBBB	10	0	0	70	100	100
	RBBB	9	0	1	70	100	90
	PVC	8	3	2	67	80	95.71
	SHB	9	0	1	70	100	90
	SVT	10	0	0	70	100	100

Table 2. Performance evaluation of proposed methodology using PPV and NPV

Classifier	Classes	TP	FP	FN	TN	PPV (%)	NPV (%)
Random forest	N	10	1	0	69	100	90.91
	AF	10	0	0	70	100	100
	PAC	10	0	0	70	100	100
	LBBB	10	0	0	70	100	100
	RBBB	7	0	3	70	95.89	100
	PVC	9	2	1	68	98.55	81.82
	SHB	10	0	0	70	100	100
	SVT	10	1	0	69	100	90.91

for calculating the NPV, sensitivity, specificity and PPV are given in the Eq. (6), (7), (8) and (9).

$$NPV = \frac{TN}{TN+FN} \times 100 \qquad (6)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100 \qquad (7)$$

$$Specificity = \frac{TN}{TN+FP} \times 100 \qquad (8)$$

$$PPV = \frac{TP}{TP+FP} \times 1 \qquad (9)$$

Additionally, accuracy is one of the effective evaluation metrics used for finding the effectiveness of the proposed methodology in arrhythmia classification. The general formula of accuracy is given in the Eq. (10).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \qquad (10)$$

Where, FP is represented as false positive, TN is denoted as true negative, TP is indicated as true positive, and FN is stated as false negative.

4.2 Quantitative analysis using MIT-BIH arrhythmia dataset

In this sub-section, MIT-BIH arrhythmia dataset is assessed for evaluating the performance of the proposed methodology. In this research study, arrhythmia classification methodology implemented on a digital signal processing platform to classify eight heartbeat conditions: N, AF, PAC, LBBB,

RBBB, PVC, SHB, and SVT. In Table 1, the proposed methodology performance is validated by means of sensitivity, and specificity for a random iteration. Here, the performance evaluation is validated for 400 random ECG signals with 80% of training and 20% of testing. The average sensitivity and specificity of proposed methodology is 96.78% and 95.71%. The graphical representation of proposed methodology using sensitivity and specificity is denoted in the Fig. 6.

In Table 2, the proposed methodology performance is validated by means of PPV and NPV for a random iteration. The average PPV and NPV of proposed methodology is 99.305% and 95.45%. From the experimental analysis, the lower scores are obtained for PVC, RBBB, and SVT (81.82% of NPV, 95.89% of PPV, and 90.91% of NPV, respectively). However, the proposed methodology
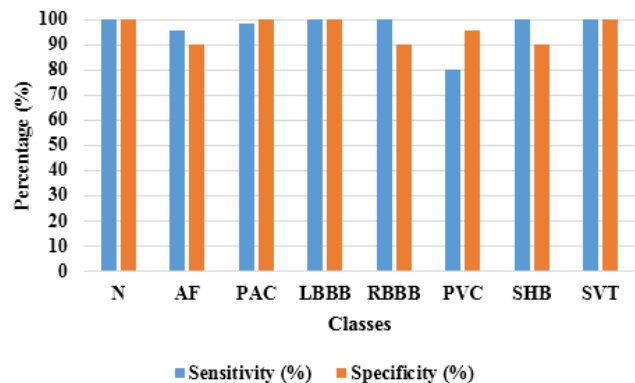

Figure. 6 Graphical representation of proposed methodology by means of sensitivity and specificity

Table 3. Accuracy evaluation of proposed methodology using with and without feature selection

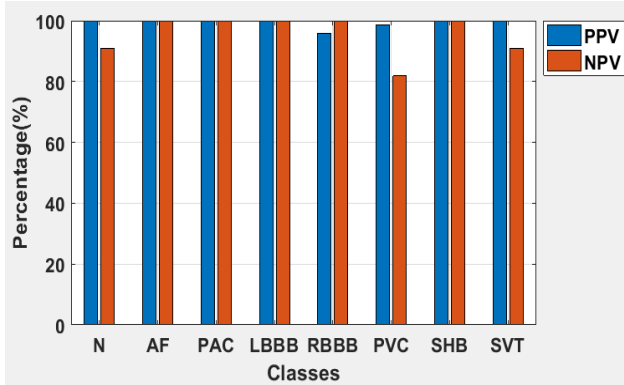| Signal pre-processing	Semantic feature extraction	Feature selection	Classifier	Accuracy (%)
Normalization and sixth order zero phase Butterworth filter	Semantic features (combination statistical features, entropy features, linear and non-linear features)	Without infinite feature selection	Random forest	93.29
		With infinite feature selection		94.32

Figure. 7 Graphical representation of proposed methodology by means of PPV and NPV

Table 4. Cross validation of proposed methodology

Parameters	Validation
Dataset	MIT-BIH arrhythmia dataset
Pre-processing	Normalization and sixth order zero phase Butterworth filter
Feature extraction	Semantic features
Feature selection	Infinite feature selection
Classification method	Random forest
Total number of samples	400 samples
Training samples	80%
Testing samples	20%
Accuracy (%)	94.32%
Sensitivity (%)	96.78%
Specificity (%)	95.71%
PPV (%)	99.305%
NPV (%)	95.45%

yielded 100% of PPV and NPV for the remaining heartbeat conditions. The Tables 1 and 2 confirmed that the proposed methodology performs effectively on MIT-BIH arrhythmia database. The graphical representation of proposed methodology using PPV and NPV is denoted in the Fig. 7.

Table 3 denotes the performance of proposed methodology with infinite feature selection and without infinite feature selection. In with infinite feature selection, the random forest classifier improves the accuracy in arrhythmia classification up to 1.03% compared to with-out infinite features selection. In this research study, the semantic features determine the non-linear and linear properties of ECG data and also preserves the quantitative relationships between the high and low level features. The evaluation metrics confirms that the proposed methodology performs significantly in arrhythmia classification compared to the previous approaches.

Table 4 represents the cross validation of proposed methodology, which is applied to the database: MIT-BIH arrhythmia dataset. Random forest classifier delivers an accuracy of 94.32% with specificity of 95.71% and sensitivity of 96.78%. PPV and NPV value for arrhythmia classification is 99.305% and 95.45%, respectively. The outcome of random forest classifier shows that the proposed methodology performs well and also delivers a significant contribution in computer-aided health monitoring system.

4.3 Comparative analysis

The comparative study of existing and proposed methodology is denoted in the Table 5. A. Gutiérrez-Gnecchi, R. Morfin-Magana, D. Lorias-Espinoza, A. del Carmen Tellez-Anguiano, E. Reyes-Archundia, A. Méndez-Patiño, and R. Castañeda-Miranda, [16] developed a new arrhythmia classification system for classifying eight heartbeat conditions such as, N, AF, PAC, LBBB, RBBB, PVC, SHB, and SVT. This research work was performed on an online database (i.e., MIT-BIH arrhythmia dataset). The developed system almost achieved 92.75% of accuracy in ECG signal classification. In addition, H. Ebrahimnezhad, and S. Khoshnoud, [17] developed a new automated system to classify four heartbeat conditions: N, APC, RBBB and LBBB. Here, the developed system achieved 92.90% of accuracy in ECG signal classification. Compared to these existing papers, the proposed work achieved 94.32% of classification accuracy that was higher than the existing papers. 8. In this research study, semantic feature extraction and feature selection is an integral part of arrhythmia classification. The ECG signals comprises of numerous features and high data-space

Table 5. Comparative analysis between proposed and existing methodology

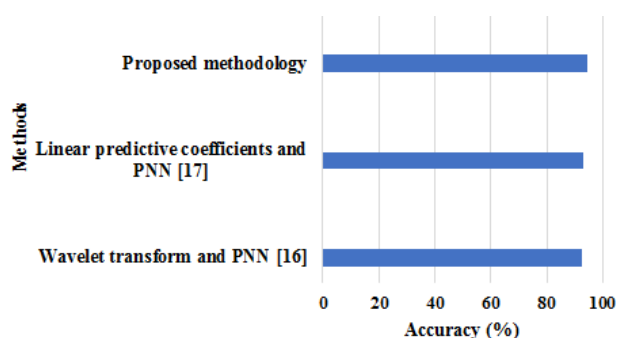| Methodology	Database	Feature classification	Accuracy (%)
Wavelet transform and PNN [16]	MIT-BIH arrhythmia dataset	Eight heart-beat conditions	92.75
Linear predictive coefficients and PNN [17]	MIT-BIH arrhythmia dataset	Four heart-beat conditions	92.90
Proposed methodology	MIT-BIH arrhythmia dataset	Eight heart-beat conditions	94.32

Figure. 8 Graphical representation of comparative analysis

volume that leads to "curse of dimensionality" problem. So, feature selection is essential for optimizing the features, which is fit for better classification. In addition, semantic feature extraction evaluates the both non-linear and linear properties of ECG data and also preserves the quantitative relationships between the high and low level features. The effectiveness of feature selection is denoted in the Table 3. The graphical image of comparative analysis is given in the Fig.8.

5 Conclusion

In health monitoring system, ECG signal based arrhythmia classification is one of the most active research areas. The main aim of this research is to obtain the effective features for classifying the eight heartbeat conditions using MIT-BIH arrhythmia dataset. In this research study, semantic feature extraction is utilized to extract the feature vectors from acquired ECG signals. Then, a superior feature selection methodology: infinite feature selection is used to select the optimal feature subsets or rejects the irrelevant feature vectors. This optimal feature information is given as the input for random forest classifier for classifying the ECG signals. This action helps the doctors/clinicians to diagnose the arrhythmia disease easily. Compared to other

existing methods in arrhythmia classification, the proposed methodology delivered an effective performance by means of accuracy and shows 1.4-1.5% of improvement in classification accuracy. In future work, a new supervised classification methodology is implemented with descriptor level features for further improving the classification rate of arrhythmia and also more number of arrhythmia heartbeat conditions are undertaken for experimental analysis.

References

[1] P. Melin, J. Amezcua, F. Valdez, and O. Castillo, "A new neural network model based on the LVQ algorithm for multi-class classification of arrhythmias", *Information Sciences*, Vol.279, pp.483-497, 2014.

[2] R.J. Martis, U.R. Acharya, H. Prasad, C.K. Chua, C.M. Lim, and J.S. Suri, "Application of higher order statistics for atrial arrhythmia classification", *Biomedical Signal Processing and Control*, Vol.8, No.6, pp.888-900, 2013.

[3] Y. Kutlu and D. Kuntalp, "A multi-stage automatic arrhythmia recognition and classification system", *Computers in Biology and Medicine*, Vol.41, No.1, pp.37-45, 2011.

[4] E.J.D.S. Luz, T.M. Nunes, V.H.C. De Albuquerque, J.P. Papa, and D. Menotti, "ECG arrhythmia classification based on optimum-path forest", *Expert Systems with Applications*, Vol.40, No.9, pp.3561-3573, 2013.

[5] O. Castillo, P. Melin, E. Ramírez, and J. Soria, "Hybrid intelligent system for cardiac arrhythmia classification with Fuzzy K-Nearest Neighbors and neural networks combined with a fuzzy system", *Expert Systems with Applications*, Vol.39, No.3, pp.2947-2955, 2012.

[6] M. Javadi, S.A.A.A. Arani, A. Sajedin, and R. Ebrahimpour, "Classification of ECG arrhythmia by a modular neural network based on mixture of experts and negatively correlated learning", *Biomedical Signal Processing and Control*, Vol.8, No.3, pp.289-296, 2013.

[7] W.H. Jung and S.G. Lee, "An Arrhythmia Classification Method in Utilizing the Weighted KNN and the Fitness Rule", *IRBM*, Vol.38, No.3, pp.138-148, 2017.

[8] M. Mitra and R.K. Samanta, "Cardiac arrhythmia classification using neural networks with selected features", *Procedia Technology*, Vol.10, pp.76-84, 2013.

[9] T.W. Chua and W.W. Tan, "Non-singleton genetic fuzzy logic system for arrhythmias classification", *Engineering Applications of*

Artificial Intelligence, Vol.24, No.2, pp.251-259, 2011.

[10] R.G. Afkhami, G. Azarnia, and M.A. Tinati, "Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals", *Pattern Recognition Letters*, Vol.70, pp.45-51, 2016.

[11] G. Sannino and G. De Pietro, "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection", *Future Generation Computer Systems*, Vol.86, pp.446-455, 2018.

[12] R.R. Linhares, "Arrhythmia detection from heart rate variability by SDFA method", *International journal of cardiology*, Vol.224, pp.27-32, 2016.

[13] M. Mohanty, S. Sahoo, P. Biswal, and S. Sabut, "Efficient classification of ventricular arrhythmias using feature selection and C4. 5 classifier", *Biomedical Signal Processing and Control*, Vol.44, pp.200-208, 2018.

[14] A.F. Khalaf, M.I. Owis, and I.A. Yassine, "A novel technique for cardiac arrhythmia classification using spectral correlation and support vector machines", *Expert Systems with Applications*, Vol.42, No.21, pp.8361-8368, 2015.

[15] J.S. Wang, W.C. Chiang, Y.L. Hsu, and Y.T.C. Yang, "ECG arrhythmia classification using a probabilistic neural network with a feature reduction method", *Neurocomputing*, Vol.116, pp.38-45, 2013.

[16] J.A. Gutiérrez-Gnecchi, R. Morfin-Magana, D. Lorias-Espinoza, A. del Carmen Tellez-Anguiano, E. Reyes-Archundia, A. Méndez-Patiño, and R. Castañeda-Miranda, "DSP-based arrhythmia classification using wavelet transform and probabilistic neural network", *Biomedical Signal Processing and Control*, Vol.32, pp.44-56, 2017.

[17] H. Ebrahimnezhad and S. Khoshnoud, "Classification of arrhythmias using linear predictive coefficients and probabilistic neural network", *Applied Medical Informatics*, Vol.33, No.3, pp.55-62, 2013.