Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Fusion of Distribution Diversity Measures to Optimize Cross-media Features for Arrhythmia Prediction by Ensemble Classification

S. Aarathi[1], S. Vasundra[2]

[*1]Reasearch Scholar, Department of Computer Science and Engineering, College of Engineering, JNTUA, Anantapuramu, Andhra Pradesh, India

aarthi.cse@gprec.ac.in

[2]Professor, Department of Computer Science and Engineering, College of Engineering,JNTUA, Anantapuramu, Andhra Pradesh, India

vasundras.cse@jntua.ac.in

*****Correspondence: aarthi.cse@gprec.ac.in

Article Info

Page Number: 6597 - 6630

Publication Issue:

Vol 71 No. 4 (2022)

Abstract

Arrhythmia is a common cause of death in people affected by Cardiovascular diseases (CVD). In clinical practice, computer-aided arrhythmia prediction using electrocardiograms is critical, and it has the potential to minimize mortality caused by untrained clinicians. To predict arrhythmia on electrocardiograms (ECGs), machine learning models have been designed based on the ECG signal features architecture, which is a biologically inspired neural network Furthermore, computer-aided approaches frequently succeed in early identification of arrhythmia scope from ECG readings received, which are frequently provided by individuals or medically dispersed networks such as the internet of medical things (MIOT). Distributing computer-assisted therapy techniques that have been successfully used to treat human arrhythmia is all the rage these days, and machine learning is the latest trend in this area. Particularly well-liked in computer-aided arrhythmia prediction technologies are machine learning methodologies. The majority of recent research focuses on the use of cross-media traits in machine learning training. However, false alarms are commonly generated by machine learning models because of the huge dimensionality of the cross-media feature values employed in training. The high dimensional features of cross-media in the learning phase was addressed in this paper, and a fusion strategy was presented to minimize the total data points. It also established a method for predicting

arrhythmias from electrocardiograms using the Arrhythmia Prediction by Ensemble Classification approach (APEC). The suggested technique is a classification methodology that selects appropriate cross-media characteristics by combining diversity assessment factors. The suggested method's growth in the prediction accuracy of both labels is the subject of the experimental investigation. The cross-validation statistics of deep genetic ensemble classification (DGEC), and support vector machine (SVM-Ensemble) were compared to modern techniques of machine learning-based arrhythmia prediction algorithms to scale the performance of the APEC.

Article History
Article Received*: 25 March 2022*
Revised*: 30 April 2022*
Accepted*: 15 June 2022*
Publication*: 19 August 2022*

Key Words: Cardiovascular Disease; Coronary Arteries; Internet of Medical Thing; Medical IoT; Naïve Bayes; Particle Swarm Optimization; World Heart Federation (WHF)

1. Introduction

Cardiovascular diseases (CVD) are the leading cause of global health. The World Heart Federation (WHF) was founded in 1972 in Geneva, Switzerland. WHF is a global leader in cardiovascular health. WHF promotes heart health and reduces the global burden of heart disease and stroke, which claim 18.5 million lives annually. During this, the report of WHF (world heart-federation) in 2016, 1 out of 3 deaths were the cause of CVD, although most heart diseases, which are premature, have inevitable. The CVD financial burden was $863 billion, predicted to enhance by 22% through 2030, which costs a billion of $1.044 [1]. Because of blocked, stiffened, or narrowed blood vessels, which prevents the necessary amount of blood supply towards the brain, heart & other body parts. There are distinct CVDs types, yet blockage or narrowing of coronary arteries (CA) could be most of the heart diseases, and it ensues over time slowly.

The CAs are the blood vessels whose purpose is supplying the blood towards the heart. Heart disease is associated with valves in the heart, which might not pump properly and cause failure in the heart. The most common heart disease symptoms are breath shortness, weakness, neck pain, and throat& chest pain. Nevertheless, some of the controlling parameters assist us in lessening heart disease risk like controlled BP, evading smoking, minimum cholesterol, and daily exercise. Typically, CVD cannot be diagnosed until heart failure, stroke, angina, a heart attack occurs. Hence, it is significant to observe the cardiovascular factors and consult doctors.

The progression in information and computing schemes has enabled the health company to gather and store the daily medical data, making crucial medicaldecisions. Here, stored data of patients has been examined to make crucial decisions related to medications that might incorporate diagnosis, estimation, treatment line, & analysis of an image. The health care system has a wealth of data available, and hence it is the rich information by inappropriately of poor knowledge. For the past few years, the algorithms of machine learning act as an important part of solving intricate, maximum non-linear classification and estimation issues.

1.2. Problem Statement

In an empirical investigation, modern machine learning and AI algorithms for analysing electrocardiograms and predict arrhythmia scope show encouraging results. An important limitation in medical diagnosis is the imbalance as well as covariance of specificity and sensitivity, which results in intolerable false alarms. Their experimental investigation was carried out on a corpus of data with high sensitivity and specificity, despite the fact that the few contributions had balanced specificity and sensitivity. A vital research objective is the computer-aided diagnosis of arrhythmia utilising artificial intelligence techniques like machine learning.

1.3. Research Objective

It is possible to construct an estimating approach that may predict the presence or absence of heart disease based on a variety of heart-related symptoms. It is a critical precondition in forecasting any disease, where the prediction algorithm must classify the healthy patient as accurate. Thus, precisely anticipating disease, especially heart disease, is extremely important.

1.4. Organization of the manuscript

In the other sections of this work, the emphasis on section 2 is related. The current diversity measures to optimize cross-media features for arrhythmia prediction by ensemble classification models and several algorithms implemented by various researchers are discussed. Accordingly, section 3 proposed a solution in applying the demographic data features. Followed by section 4, the results and experimental study of this work are discussed. Conclusion for the work is discussed in section-5, followed by references.

2. Related Work

Different ML algorithms including random-forest, SVM [2], ANNs, naïve Bayes (NB), decision tree (DT), have been extensively utilized in several disease classification and estimation issues. Few of these implementations involve heart disease [3-11]. Nevertheless, the disease prediction based on ML development method and decision making related to medicine is a non-trivial challenge [5]. Here, some of the significant problems are the organization, collection and acquisition of data used for training the ML system.

The work [12] implemented several ML techniques and the performance is compared on eight medical datasets by utilizing five divergent factors: explanation, transparency, mislaid data handling and performance. Among distinct assessment factors, the NB, semi NB, KNN & back-propagation models have been assessed to be very good. Further, regarding transparency, the DT has been assessed to be very good. Here, mislaid data handling factors, NB & semi NB have been tagged to be very good. The work discusses ML techniques and implementation approaches The work [12] has notbeen offered computable prediction accuracy methods. The work [3] introduced a prototype for heart disease estimation systems possessing an interface based on the web for a query of what-if that uses three classifiers: NB, ANN & DT. Here, a survey on ML implementations in healthcare methods, specifically predicting heart disease [13], [14].

Here, the work [13] presents that NB and DT classifiers perform better than other methods while ANN, KNN & classification based on clustering could not perform better. The work [14] presents an experimental study on published 149 manuscripts from 2000-2015 for cardiology prediction, DT; SVM & ANN were identified as the most frequently used ML schemes. The work [13] [14] presents that two manuscripts contradicted in general statement-making highly accurate ML schemes for predicting heart disease. Besides, this contradiction might be due to variances in datasets and risk parameters under consideration.

The comparative analysis of classification schemes was exhibited that DT classifiers were accurate and simple [15]. Here, NB was identified as an optimal algorithm, succeeded by NNs and DT [16]. The ANN was also employed to estimate diseases. Moreover, supervised networks are utilized for diagnosing and they could be trained by using the BP algorithm. Further, simulation outcomes have depicted a reasonable accuracy [17].

Moreover, contemporary research used ensemble models to enhance classification accuracy in heart disease prediction [18]. The work [19] presents that integration of fuzzy logic-based NN and GA for the extraction of features envisioned an enhancement in accuracy up to 99.97% [20]. The

work [20] presents that GA-based trained fuzzy NN generated 97.78% accuracy to diagnose heart disease.

The work [21] presents that the accuracy of classification attained in heart-disease prediction risk by utilizing a classification system based on a rough set with a distinct dataset is 93% [21]. Moreover, NN is also used to lower the human fault in identifying BP, heartdisease & blood sugar [22-24] . The novel method CANFIS (coactive Neuro-fuzzy inference system) integrated with NNs, GA, and fuzzy logic exhibited better outcomes to estimate heart disease. In this, GA was used to tune the factors for automatic CANFIS and optimum feature set selection. The work [25], [26] presents that method is exhibited to be a resourceful tool to evaluate medical professionals in estimating heart disease [25]. The work [27] presented a model that endeavors to attain optimal accuracy, and an additional feature selection step was proposed.

The classifiers-based SVM exhibited maximum accurate outcome to classify heartbeats. The factors have been simplified by using PSO (particle swarm optimization) [28], [29]. In this, the algorithm of K-means clustering is used for extracting data from the dataset, and frequent patterns are extracted by using MAFIA (maximum frequent Item-set algorithm) to estimate heart disease based on different weightage allocated towards distinct parameters. Moreover, frequent patterns possess a maximum value of more than a particular threshold and are identified to be accurate in identifying the myocardial infarction occurrence [23].

To enhance final estimation, ensemble classifiers aggregate individual classifier decisions. Several ensemble classifier techniques are presented in the literature [30]. Models like AdaBoost [31] and Bagging [32] are used to train each classifier. The work [33] provides difficulty since it involves a large total outputs and an ensemble classifier. Other contributions focus on training each classifier with a diverse range of input characteristics. Integrating classifiers trained on different feature sets are useful, according to experiments [34], especially when single classifiers perform well. The contemporary model [35] developed an ensemble of SVM classifiers to classify a balanced dataset. Furthermore, compared to training a single SVM using all data sources, training each SVM with distinct data sourcesboosted the results.

The contemporary ensemble technique [36] selects optimal features using GA-SVM (genetic algorithm-based support vector machine) (SVM-Ensemble). The limitations of machine learning-based arrhythmia prediction are addressed by automatic detection of cardiac arrhythmia using deep genetic ensemble classification (DGEC) [37]. To detect arrhythmias in ECGs, DGEC employs linear SVM, neuro-SVM, RBFNN, and KNN. Although the DGEC performs ensemble

classification, it is ineffective to solve the problem of false alarms produced by the training variables' high dimensionality. Ensemble and hybrid approaches deal with dimensionality by training several classifiers with the same information. Dimensionality continues to generate false alarms.

Though several models are utilized to estimate heartdisease risks with better accuracy in the research, some classification algorithms detect a risk of heart disease with deprived accuracy. Many types of research, which produce maximum accuracy, employ a hybrid model that includes classification algorithms. Unlike the contemporary models stated, the suggested ensemble model APEC is using cross-media features, a fusion technique to reduce the load of the features to use in the training phase, and addressing an ensemble approach to handle the high dimensionality in features to achieve a significant reduction in false alarming.

3. Methods and Materials

In the ensemble classifiers, the approaches or methods have been divergent from each other, including variances in hypothesis, initial seed, and population and modeling mechanism. As stated in the contribution [38], three ensemble system pillars: training each ensemble system, integrating ensemble members, and diversity. Classification is the process of classifying a specified set of data into classes. Further, it may be performed in both unstructured and structured data.

3.1. The Data

The arrhythmia prognosis method uses machine learning, which uses cross-media data as input to perform training and prediction phases. The electrocardiogram's demographic features, the sequence of digital signal coordinates of the electrocardiogram signals as sequential patterns, and features of the QRS complex’s virtual images have been considered for the proposed approach. The other qualitative objective minimizing or nullifying the false alarming caused by the high dimensional features considered to train the proposal. The datasets MIT-BIH [39], PTB-XL [40] and SHHS [41] dataset has considered. Though the dataset SHHS has both the electrocardiograms and polysomnograms of the anonymized patients, to achieve uniformity in features, the electrocardiograms only have been considered from the corresponding dataset. To amplify the dimensionality in the values projected for optimal features. The electrocardiograms of diversified datasets are considered input data for both the training and testing phase of the proposal.

3.2. Demographic Features

A rhythm or heartbeat rate that is unusually low or high, or has a non-conducive pattern, is known as arrhythmia. The ECG signal series is used to recognize the heartbeat format in the data format. These signals were also utilized to identify a unique collection of tri-dimensional characteristics. These qualities and characteristics are explored further down.

3.2.1. Intervals

The RR interval is the delay between successive R waves of the ECG signal and their inverted signal to the HR. It is an inherent function of the Sinus perch and automatic effect circumstances.

For the QRS complex, the PR Interval is the estimated P-wave start. The AV perch decides on line rerouting. The average PR interval is 120-200 milliseconds. First-degree heart block is indicated by PR intervals of more than 200 milliseconds.

The average QRS complex lasts 0.08-0.10 seconds or 80-100 milliseconds. It's regarded as moderate and prolonged when the duration is 0.10-0.12s. QRS duration of more than 0.1s is considered abnormal.

The intrinsic character of the QT interval has been and recognized at various lengths. The QT interval is typically 400-440 milliseconds (0.4-0.44s). QT intervals in female patients are longer than in male individuals. QT intervals are lengthened by low heart rates.

This QTC interval is 0.40s-0.44s for normal QTC. The QTC in situations of sudden cardiac death or arrest is 431-450ms in males and 470ms in females. When QTC in males or females reaches 450ms or 470ms, it is deemed abnormal.

3.2.2. Axis

The ECG axis depicts the entire electrical activity of the heart. It can be any of the following: left, right, normal, or undetermined (northwest axis).

The sinoatrial perch depolarizes the atrium, and the P-Wave Axis Score redirects to atrial depolarization. The SA perch and the P-wave in lead II create the action potential.

The QRS axis score of the wave is crucial in determining the QRS axis, which can range from -30 to +90 degrees. Negative results for left axis deviation ranged from -30 to -90. The right axis deviates by 900-80 degrees.

During the heart's intermediate period, the T-wave is the shift in ventricular membrane potential (interval of the composite QRS to the apex of the T-wave). Non-cardiac and cardiac diseases can be caused by T-wave fluctuations and abnormally inverted T-wave amplitudes.

The QRS axis and limb scores confirm the T or P wave axis. The angle of the QRS axis should be between -30 and +90 degrees. A -30 to -90° QRS vector represents left axis anisotropy. 900o and 180o are shown on the right axis. The entire angle of the axis is unknown.

3.2.3. Signal

Biomedical signals' statistical characteristics change throughout time. Wavelet transformations are used to represent signals with both frequency and time, making it possible to analyze ECG signals. This wavelet transform is used to extract ECG characteristics [42], recognize heartbeat [43], and de-noise [44]. The suggested model extracted features using DWT, which may be divided into higher or lower frequency approximation devices.

Conflicts, eyelets, discrete, Daubechies, and Meyer [45] are some of its orthogonal features. Each heartbeat has been fragmented using wavelet transform discrete Meyers finite impulse response conditions ranging from 011.25Hz to 11.2522Hz. Wavelet features that decrease dimensionality using ICA provide 200 coefficients. To get 12 morphological input characteristics, six significant ICA components were picked from two DWT sub-bandsthe below used formula notations in table 1.

Table 1: Formula Descriptions

ECG	electrocardiograms
fP, fN	features of positive and negative
$\|r\|$	record
s	size
d_a, d_b	distribution
i	Index
cr_i	cumulative ratio
$\|v_1\|, \|v_2\|$	vector
$'sd'$	standard-deviations
n	Number
C	Centroid
cl_i	Cluster
ng	n-gram

3.3. Signal Features

A set of electrocardiograms has been presented in digital format (as x, y coordinates). Each of these electrocardiograms represents the patient who tested positive or negative for arrhythmia. The set ECG shall be partitioned into two sets pT, nT having electrocardiograms tested positive and negative in respective order. Each entry of these sets pT, nT is the sequence of y-coordinates representing the sequence of x-coordinates. The sequential patterns of size greater than zero shall be discovered from each entry of the sets pT, nT. They discovered unique sequential patterns of both sets pT, nT shall be saved further as respective sets fP, fN and denotes as features of positive and negative labels in respective order. The possible total features (sequential patterns) [46] of size greater than zero is $|r|-s+1$ the difference between the size of the record $|r|$ as well as the threshold $\{(s-1)\exists s=1,2,3,...,|r|\}$ representing the sequential pattern size s minus 1.

3.4. QRS Complex Image Features

The features of the QRS image representing the corresponding electrocardiogram have been considered the other dimension of the features used in this proposed method. To obtain these features, the electrocardiogram signals shall segment by implementing noise eradication [47] and QRS identification algorithms stated in [48]. Later, these sections have been aligned as per R-points; they have been converted into 2-dimensional signals of ECG known as QRS images having dimensions of 256 x 256. Ultimately, the resultant QRS images shall use as input to extract the entropies and morphological features [49].

For each electrocardiogram signal, the resultant virtual QRS image [47], after pre-processing, detects the values representing the entropies [50] and morphological features [49] for recognizing electrocardiogram images of class positive (having arrhythmia) and negative (not having arrhythmia). There could be a prerequisite for extracting better features, which can differentiate both kinds of electrocardiogram images. Here, in arrhythmia, the morphological features & entropies are prominent features for distinguishing [51].

This section explores the method of exploring morphological features and entropies from the virtual QRS images of each electrocardiogram considered for both training and prediction phases. The abnormal heartbeat causes changes in the electrocardiogram image micro-structure. These features of electrocardiogram images are generally particular to entropies and morphology.

3.4.1. Entropy

The term entropy is the evaluation of randomness, which is used for texture characterizing of image input. The value is high when the entire co-occurrence matrix elements are identical.

One of the most crucial components of feature extraction is entropy. Here, we distinguish between electrocardiograms of infected and healthy hearts. It is important to differentiate between positive and negative electrocardiogram images, and this distinction relies on the entropies which are accessible and discussed in this literature [51], [50]. When evaluating entropy, five distinct measures are taken into account. The standard ROI (region of interest) histogram has been recognised as a useful tool for calculating these entropies. Let the set $\varepsilon = \{r, h, c, k, y\}$ representing Renyi's[52], Havarda's [53], Charvat's[54], Kapur's[55], and Yeager's [56] entropies of the corresponding virtual QRS image.

3.4.2. The Morphologic Features

In this section, the morphological feature from the feature's extraction is explored. For detecting the electrocardiogram images, the morphometric information features are considered. Nine morphological features are extracted and used. Some of the features proposed in [57] are having invariant minutes [58]. They are deliberated in the form of Morphometric information features, which are extensive in representing the identifications of anomalous electrocardiogram images. Moreover, this can change the size & shape represented among the virtual QRS images of the negative & positive electrocardiogram images.

3.5. Distribution Diversity Measures

In this section, we'll take a look at how to fuse various distribution diversity estimation strategies to get the best results from feature selection. Many different types of statistical tests, including the WRS-Test (Wilcoxon signed-rank) [59], MWU-Test (Mann-Whitney U test) [61], KS-Test (Kolmogorov-Smirnov test) [60], as well as Dual-tailed t-Test [62], have been combined to find the most informative results for determining the best features to use in a dataset.

3.5.1. KS-Test

The KS-test calculates the gap between the total distribution and the experimental sample distribution. This could be noted as a disparity among two different distribution specimens of similar or different sizes.

Using the KS-test as a distance metric, we can see how the distribution diversity between the two datasets compares [60]. Additionally, the type of data distribution is irrelevant to this metric. The steps involved in employing the algorithmic method have also been outlined:

$ks\_test(d_a, d_b)$ Begin // receives two distribution vectors d_a, d_b.

$d_a^{agr} = \sum_{i=1}^{|d_a|} \{e_i \exists e_i \in d_a\}$ //aggregate of distribution vector d_a

$d_b^{agr} = \sum_{i=1}^{|d_b|} \{e_i \exists e_i \in d_b\}$ //aggregate of distribution vector d_b

The following sequence of expressions demonstrates the process of assessing cumulative ratios of the distribution vector d_a

$cr_o = 0$	Cumulative ratio cr_0 of the element at index 0 is zero
$i = 1$	index i begins at 1
$for\left(i \le \lvert d_a \rvert\right) begin$	index i must not be greater than the size of the distribution
$cr_i = \{e_i \exists e_i \in d_a\} * \left(d_a^{agr}\right)^{-1} + cr_{i-1}$	discovers cumulative ratio cr_i of the element of the distribution d_a, which indexed by i
$cr_a^{cr} \leftarrow cr_i$	updating set d_a^{cr} by adding resultant cumulative ratio cr_i
$i = i + 1$	Increments the index by 1
end	

Similar process stated for distribution d_a shall be applied to discover the set d_b^{cr} of cumulative ratios.

Further, discovers the absolute difference between cumulative ratios of elements indexed at i of both distributions d_a, d_b as follows

$i = 1$	index i begins at 1
$for\left(i \le \max\left(\lvert d_a \rvert, \lvert d_b \rvert\right)\right) begin$	index i must not be greater than the maximum size of both distributions
$ad \leftarrow \sqrt{\left(\left(x_i \exists x_i \in d_a^{cr}\right) - \left(y_i \exists y_i \in d_b^{cr}\right)\right)^2}$	Absolute difference of the cumulative ratio of elements indexed at i in both distributions
$i = i + 1$	Increments the index by 1

end

In addition, the highest number in the set ad is called the d-stat.

if $(d-stat > d-critic)$	A return of 0 indicates that the two distributions are not distinguishable (because the d-stat is larger than the d-critic), while a return of 1 indicates that they are.
return 0	
else *return* 1	

d-critic is the value denoted in KS-table for aggregate values d_a^{agr}, d_b^{agr} of both distribution vectors d_a, d_b, and the given degree of probability (p-value)

3.5.2. Wilcoxon-Rank Sum Test

The WRS-Test is widely used nonparametric test for comparing data from different sets. This has also been referred to as the MWU-Test, and it has been used to determine whether or not two samples come from the same population. By comparing the medians of the two groups, some of the evaluators can grasp the concept behind this test. In addition, we have recall, a parametric test for contrasting means across groups.

With the exception of, the non-parametric test two-sided and null data analysis hypothesis was reported as follows.

If both the distribution samples are distinct then returns ONE, else returns ZERO

This test is often a two-sided test, indicating that populations are not the same in the specified direction. One-sided study hypothesis is employed when interest hinges on thenegative or positive change in one population relative to others. Tracking where sample each observation is from, the test technique combines two samples asoneunited sample. Order them from 1 to "$n_1 + n_2$".The description of the test and mathematical model follows:

Find total elements as 'U1' in the vector 1 are greater than the counterpart elements of the vector2, similarly, find total elements as 'U2' in the vector 2 are greater than the counterpart elements of the vector 1

Find the greatest as 'U' of the 'U1', 'U2'

Assess the d-critic of the 'U' (of vector having highest total elements higher than the counterpart) atfixed diversity threshold

If d-critic is less than 'U', given vectors are having diversity in their distribution.

in contrast to above condition, the distribution of given both vectors is similar

$wrstest(v_1, v_2)$ // Begin

Both vectors v_1, v_2 should be sorted ascending.

$$U(v_1) = 0$$

$$U(v_2) = 0$$

$\overset{|v_1|}{\underset{i=1}{\forall}} \{e_i \exists e_i \in v_1\}$ Begin //to each element

$\overset{|v_2|}{\underset{j=1}{\forall}} \{e_j \exists e_j \in v_2\}$ Begin //the vector's for-each element

$$U(v_1) = \{(U(v_1)+1) \exists (e_j < e_i)\}$$

$$U(v_2) = \{(U(v_2)+1) \exists (e_i < e_j)\}$$

End

End

$$U = \begin{Bmatrix} U(v_1) \exists (U(v_1) < U(v_2)) \\ U(v_2) \exists (U(v_2) \le U(v_1)) \end{Bmatrix}$$

For the vectors of size $|v_1| and |v_2|$, note the d-critic of diversity threshold $d\tau(0.01, 0.05, or\ 0.1)$ using U-Table

$$return \begin{Bmatrix} 1 \exists dc < U \\ 0 \end{Bmatrix}$$

End

3.5.3. T-Test

The t-test was used to calculate the predicted distribution diversity values for two-label features. The vector v_1 and vector v_2 indicate projected values for characteristics in infected and benign records, respectively. A t-test was used to measure two distribution diversity among the vector v_1 and vector v_2, as illustrated below.

- Discover the mean of the vector 1
- Discover the mean of the vector 2
- Find the standard-deviation of vector 1
- Find the standard-deviation of vector 2
- Find the difference as 'md' between mean of the vector 1 vector 2

- Find the sum of the standard-deviations as 'sd' of vector 1 and vector 2
- Find the ratio of mean difference 'md' against square route of sum of deviations 'sd'', which results t-score
- Find the probability value (p-value) of the t-score
- If the p-value is less than given probability threshold then the given vectors has diversity in their distributions
- In contrast, the given vectors are having similar distribution

$t-test(s_1, s_2)$ // Begin //The input arguments s_1, s_2 are two vectors

$$t-score = \frac{\left(\langle s_1 \rangle - \langle s_2 \rangle\right)}{\sqrt{stdv(s_1) + stdv(s_2)}}$$

The expressions $\langle s_1 \rangle, \langle s_2 \rangle$ denote the averages of the respective vectors.

the expressions $stdv(s_1), stdv(s_2)$ denote the standard-deviations of the respective vectors

Note the probability value $(p-value)$ from t-table[63].

if $(p\tau > pv)$ *return* 1 //projects that both the vectors are distinct
else if *return* 0 //projects that both the vectors are not distinct

End

3.6. Handling Dimensionality

The FC-Means (Fuzzy C-Means) [64], [65] was used to reduce the dimensionality of the values associated with each label. The FC-Means approach separates the incoming data into tuples, each of which contains a group of records with anextensive correlation (less variability or dimensionality). There might be one or more divisions for each FC-Means record.

Based on its distance from the cluster centroid, this method awards membership to each data point. The closer the data is to the cluster's centroid, the more members are clustered there. Cluster centroids are promoted according to Eqs. 1 and 2 every membership iteration:

$$\mu_{ij} = \left[\sum_{k=1}^{c} \left(d_{ij} / d_{ik} \right)^{(2/m-1)} \right]^{-1} \quad ...(\text{Eq } 1)$$

$$\forall_{j=1}^{|codes|}\left\{v_j=\left(\sum_{i=1}^{n}\left(\mu_{ij}\right)^m x_i\right)*\left(\sum_{i=1}^{n}\left(\mu_{ij}\right)^m\right)^{-1}\right\} \text{...(Eq 2)}$$

The expression n specifies the total data points. The expression v_j shows the j^{th} centroid of the cluster, whereas the expressions "$m \in [1, \infty]$", "*codes*", "μ_{ij}", and "d_{ij}" denotes index-fuzziness, centroids, i^{th} data fitnessheaded for cluster j^{th} centroid, andEuclidean distance from j^{th} cluster-centroidto i^{th} data respectively. The main intent of this fuzzy c-means-algorithm is lessening:

$$J(U,V)=\left(\sum_{i=1}^{n}\sum_{j=1}^{c}\left(\mu_{ij}\right)^m\right)*\left(d_{ij}\right)^2 \text{...(Eq 3)}$$

Algorithmic flowof Fuzzy c-means

- The given data points denotes by theexpression $X=\{x_1,x_2,x_3,....,x_n\}$, and the centroids denotes by expression $V=\{v_1,v_2,v_3,....,v_c\}$.

1) The cluster centroid c is picked randomly.
2) Computes the fitness μ_{ij} using Eq 4:

$$\mu_{ij}=1/\sum_{l=1}^{c}\left(d_{ij}/d_{il}\right)^{(2/m-1)} \text{...(Eq 4)}$$

3) Estimates the fuzzy-centroids v_j using Eq 5:

$$v_j=\left(\sum_{i=1}^{n}\left(\mu_{ij}\right)^m x_i\right)/\left(\sum_{i=1}^{n}\left(\mu_{ij}\right)^m\right), \forall 1\le j\le c \text{ ...(Eq 5)}$$

4) The step 2 and step 3 are recurrent till attains minimal j-value or $\|U(l+1)-U(l)\|<\alpha$

The iteration index is indicated by the symbol l. The phrase α denotes the end of the criteria between [0, 1]. The expression $U=(\mu_{ij})n*c$ stands for a fuzzy membership matrix. Finally, the symbol J stands for the objective function.

3.7. The classifier

The incremental binary classifier [66] is used in the suggested ensemble classification. The classifier outperforms more sophisticated algorithms in binary classification. Subsections look into class prognosis and classifier training.

3.7.1. Formatting hierarchy

The suggested classifier functions in two phases referred as training and testing. The first phase builds a perch hierarchy with each level having additional perches that compared to the previous level, if any. Training builds positive and negative label hierarchies. The perches will be arranged in two hierarchies.

For each positive or negative cluster, sort cross-media optimal features by size. The largest n-gram features will be grouped together, with n-grams having equal size and frequency. Each group with n-grams of size n will be a perch on level $\{l \exists 1 \le l \le n\}$ of the corresponding hierarchy. Similarly, the feature n-gramshaving size $\{n-1, n-2, ..., n-(n-1)\}$ are partitioned into groups containing diverse n-grams having equal frequency. These groups are substituted as perches at level $\{l \exists 1 < l < n\}$. This process repeats until the end of the hierarchy. Each group of n-grams of size one must have the same frequency. These groups will be perches on the n^{th} level.

3.8. Classification

The second phase is testing, which predicts whether an electrocardiogram signal shows arrhythmia. Here's how classification works. Discover the unlabeled records' cross-media features. It also finds all possible subsets of cross-media features from each record. From the electrocardiogram signals, these n-grams (subsets) trace the arrhythmia scope (an unlabelled record). During the classification phase, each hierarchy framed by the clusters' optimal features is searched. The electrocardiogram's fitness for both classes is assessed as described below.

Search the hierarchy representing the positive class for each n-gram. Save the frequency of each perch that contains the corresponding n-gram in to the list fl_+^r. Similar search should be done on hierarchy representing the clusters of negative class to get a list fl_-^r of frequencies. Determining positive fitness score pfs, the absolute product of the positive frequencies fl_+^r representing the given test record's positive fitness r. Similarly, discovers negative fitness score nfs using the list fl_-^r.If the difference of fitness scores pfs, nfs is found to be greater that compared to deviation threshold. The given test record is then positive. If nfs is greater than pfs, the test record is negative. The following describes the model's algorithmic flow.

3.8.1. Feature Optimization

The proposed method relies heavily on the optimization of features. The training phase requires the input set C of records to partition into two sets, C_p and C_n, with C_p containing the positively labelled records and C_n containing the negatively labelled records. Each record r of the corresponding set represents the demographic features, sequential n-gram patterns, and entropies, morphological features of an electrocardiogram. The entropies and morphological features shall obtain from the QRS images of the corresponding electrocardiogram. FC-Means (see sec. 3.6) is used to divide the sets C_p, C_n into different clusters as $Cl_p = \{p_1, p_2, ..., p_s\} and\, Cl_n = \{n_1, n_2, ..., n_t\}$, some of which may or may not contain the common records. Finding the best features of both classes for each cluster is the most important phase of feature optimization that performs as described below.

For each cluster $\{p_i \exists p_i \in Cl_p \wedge i = 1,2,...,s\}$ of the positive label

To be more specific, ' p_i ' is a two-by-two matrix representing the cluster. The positive characteristics of an electrocardiogram are represented in each row of the matrix. The projections of feature attribute values $vf_a = \{v_1, v_2, ..., v_{|c|}\}$ across all records in a given matrix are shown in each column. The notation $|r|,|c|$ represents the row and column counts, respectively, of the associated matrix.

When the mean diversity $d_{p_i}^a$ among the values vf_a as well as the values of the respective attribute f_a in negative class clusters is found to be greater that compared to diversity threshold $d\tau$ the process of optimizing features can select an attribute fa as optimal towards to the cluster p_i of the positive class. To assess the variability of each attribute, the fusion of distance measures discussed in section 3.5 was used. The feature optimization is represented algorithmically as follows.

$diversity\_assessment(f, v, tCl)$ Begin // receives the feature attribute f, values v representing the corresponding feature in the cluster of positive class or negative, and the target clusters of label positive or negative

$ds = 0$ // diversity score

$\forall_{i=1}^{|tCL|} \{cl_i \exists cl_i \in tCL\}$ Begin // for each cluster cl_i

$ods = 0$ // overall diversity score

$fv \leftarrow cl_i(f)$ // feature f values of the cluster cl_i

$ods += ks\_test(v, fv)$ //updates with ks-test response

$ods += wrstest(v, fv)$ //updates with $mwu-test$ response

$ods += t\_test(v, fv)$ //updates with t-test response

$ds = \left\{(ds+1)\exists\left(1-ods^{-1}\right)\geq 0.5\right\}$ //increase distance-stat by one, if the overall diversity score meets the criteria

End

$return\left(\frac{|ds|}{|tCL|^{-1}}\right)$ // return theprobability of diversity

End // the $diversity\_assessment(f, v, tCl)$ ends here

*Optimal features of the clusters Cl_p *

$\underset{i=1}{\overset{|Cl_p|}{\forall}}\left\{c_i \exists c_i \in Cl_p\right\}$

$\underset{j=1}{\overset{|F|}{\forall}}\left\{f_j \exists f_j \in F\right\}\left\{\left(ofa(c_i)\leftarrow f_j\right)\exists\left(d\tau < diversity\_assessment(f_j, vf_j, Cl_n)\right)\right\}$

End

Similarly, discovers optimal featuresof each cluster $\left\{c_i \exists c_i \in Cl_n\right\}$ of the negative class

Explore all possible subsets of the optimal features F for each cluster in both classes. 'E' stands for the total distinct subsets that can be constructed from the provided set of characteristics.

Find all feasible ngf subsets of each cluster's optimum features ofa. The expression $|ngf| = 2^{|F|-1}$ denotes the total different feature subsets.

For each entry (optimum feature subset) of the set ngf, select the distinct pattern of values found in one or more records of the corresponding cluster. Also estimate empirical probability of the respective pattern of values.

3.8.2. Formation of the perch hierarchies

Establish hierarchies from both classes' clusters. The expression Cl indicates positive or negative clusters.

$\underset{i=1}{\overset{|Cl|}{\forall}}\left\{c_i \exists c_i \in Cl\right\}$

$l = 1$ // level index

$k = n$ // n-gram size initialized by maximum size of the n-grams

---

$foreach\{k \exists k \geq 1\}$ Begin

$\underset{j=1}{\overset{|ong(c_i)|}{\forall}}\{ng_j \exists ng_j \in ong(c_i)\}$ // Begin

$\left\{ong_k^{fr}(c_i) \leftarrow ng_j \exists (j=1) \vee \left(\left|ng_j\right| \equiv k\right)\right\}$

End

In the hierarchy phc_i, place each set $ong_k^{fr}(c_i)$, each of which has n-grams of size k and frequency ratio fr, as perches.

$k = k-1$ // adjust the n-gram size in decrement order by 1

$l = l+1$ // adjust the level l in incremental order

End

3.8.3. Prediction phase

Here we detail the algorithmic procedure for the label prediction strategy, which includes a positive fitness estimate.

Conduct a global search for appropriate perches that match the n-gram features $ng(tr)$ of the sample record.

#Positive fitness estimation#

$pf = 1$ // fitness is set to 1(maximum value), which meets the criteria $0 < pf \leq 1$.

$\underset{i=1}{\overset{|Cl_+|}{\forall}}\{c_i \exists c_i \in Cl_+\}$ Begin // to each cluster of the positive class

$\underset{l=1}{\overset{|phc_i|}{\forall}}\{l \exists 1 \leq l \leq |phc_i|\}$ // each level of the hierarchy built from the corresponding positive class cluster

$\underset{m=1}{\overset{||phc_i^l|}{\forall}}\{p_m \exists p_m \in |phc_i^l\}$ Begin // each perchlisted at the corresponding level

$\underset{p=1}{\overset{|ng(tr)|}{\forall}}\{(pf = pf \times fr(p_m)) \exists ng_p \in p_m\}$

End

End

End

$pF(tr) = \sqrt{(1-pf)^2}$ //maximum fitness of positive class

 # Negative class fitness estimation #

$nf = 1$// fitness is set to 1 (maximum value), which meets the criteria $0 < nf \le 1$.

$\underset{i=1}{\overset{|Cl_-|}{\forall}} \{c_i \exists c_i \in Cl_-\}$ Begin //to each cluster of the negative class

$\underset{l=1}{\overset{|phc_i|}{\forall}} \{l \exists 1 \le l \le |phc_i|\}$// each level of the hierarchy built from the corresponding negative class cluster

$\underset{m=1}{\overset{||phc_i^l|}{\forall}} \{p_m \exists p_m \in |phc_i^l\}$ Begin // each perch listed at the corresponding level

$\underset{p=1}{\overset{|ng(tr)|}{\forall}} \{(nf = nf \times fr(p_m)) \exists ng_p \in p_m\}$

End

End

End

$nF(tr) = \sqrt{(1-nf)^2}$ //maximum negative class fitness

3.8.4. Label Prediction

The fitness $pF(tr), nF(tr)$ of positive and negative classes will be used to forecast if the given sample (electrocardiogram signal) falls into the class positive (prone to arrhythmia) or in to the class negative (not prone to arrhythmia) as follows:

The criteria $(d\tau < (pF(tr) - nF(tr)))$ certifies that the test sample (electrocardiogram

signal) falls into the class positive (prone to arrhythmia).

The criterion $(d\tau < (nF(tr) - pF(tr)))$ validates that the supplied test record indicates that the given sample (electrocardiogram signal) falls into the class negative (not prone to arrhythmia).

4. Experimental Study

The performance of Arrhythmia Prognoses by Ensemble Classification (APEC) is compared to that of DGEC [37] and SVM-Ensemble [36] in this part. The experimental data are detailed in Section 3.1. The study's methodologies were implemented using Python [67]. We used data from the MIT-BIH. Standard investigation information for cardiac arrhythmia is available in the MIT-BIH arrhythmia database. It has been utilized for heart rhythm research and device development since 1980. Electrocardiograms are primarily found in PTB-XL. PTB-XL is 10 seconds long and comprises 21837 data from 18885 patients. The Sleep Heart Health Study (SHHS) was done by the National Heart, Lung, and Blood Institute to identify the cardiovascular implications of sleep-disordered breathing.

Positive records were 73337 (MIT-BIH: 59681, PTB-XL: 10506, and SHHS: 3150), whereas negative records were 57661 (MIT-BIH: 49765, PTB-XL: 4046, and SHHS: 3850). The method's performance was assessed using k-fold Leave-Pair-Out Cross-Validation [68].

4.1. Precision

	Precision									
	LPO# 1	LPO# 2	LPO# 3	LPO# 4	LPO# 5	LPO# 6	LPO# 7	LPO# 8	LPO# 9	LPO# 10
APEC	0.9072	0.9175	0.9226	0.9234	0.9247	0.9273	0.9304	0.9322	0.9368	0.948
DGEC [37]	0.8465	0.8439	0.8498	0.8563	0.8595	0.8616	0.8675	0.8737	0.8786	0.8814
SVM-ensemble [36]	0.7566	0.7587	0.7607	0.7654	0.7704	0.7784	0.7831	0.7846	0.7904	0.7923

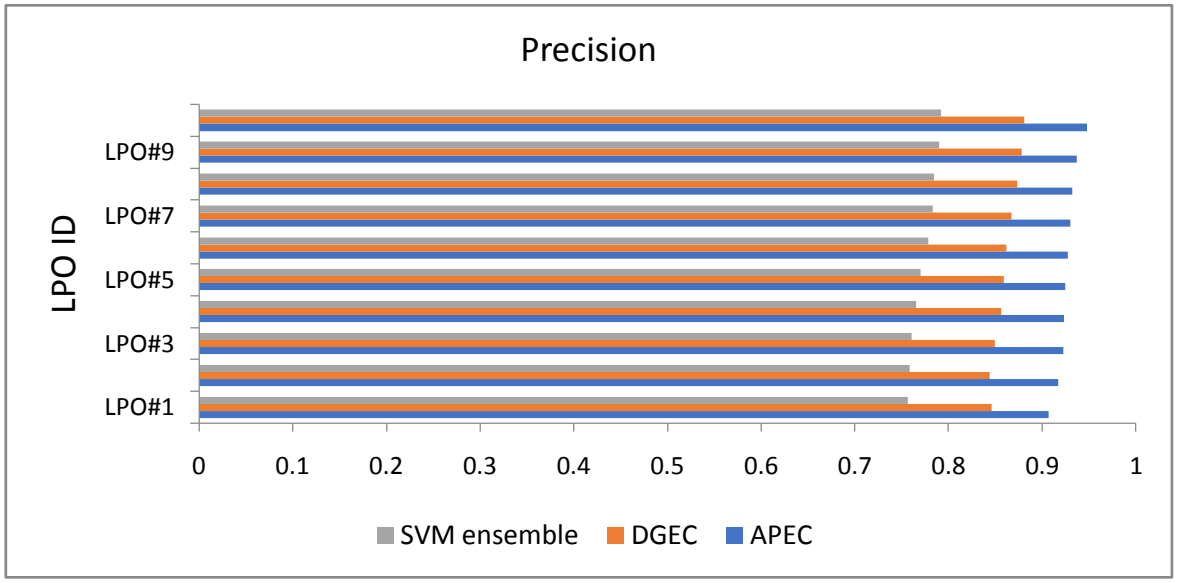

Figure 1: Precision at 10- LPOs observed for the APEC, DGEC, and SVM-Ensemble method

Figure 1 shows a graph drawn between 10-LPOs and the precision metric for the proposed APEC and the contemporary models DGEC and SVM-Ensemble methods. From the graph, it is depicted that, average and standard-deviation for the APEC is 0.92701±0.010467. Here, the proposed APEC is compared with DGEC and SVM-Ensemble method, whose average and standard-deviations are 0.86188±0.012548 and 0.77406±0.012724 respective order. It is concluded from the above-stated statistics that the APEC is more significant when compared with DGEC and SVM-Ensemble methods.

4.2. Specificity

Specificity										
	LPO# 1	LPO# 2	LPO# 3	LPO# 4	LPO# 5	LPO# 6	LPO# 7	LPO# 8	LPO# 9	LPO#1 0
APEC	0.877 2	0.893 2	0.901 8	0.901 6	0.903 5	0.908 7	0.910 2	0.914 5	0.92	0.9353
DGEC	0.804 3	0.799	0.810 2	0.815	0.817 5	0.823 2	0.83	0.842 7	0.847 5	0.8529
SVM ensemble	0.704 9	0.695 1	0.703 5	0.715 7	0.724 6	0.719 5	0.729 1	0.735 3	0.744	0.7382

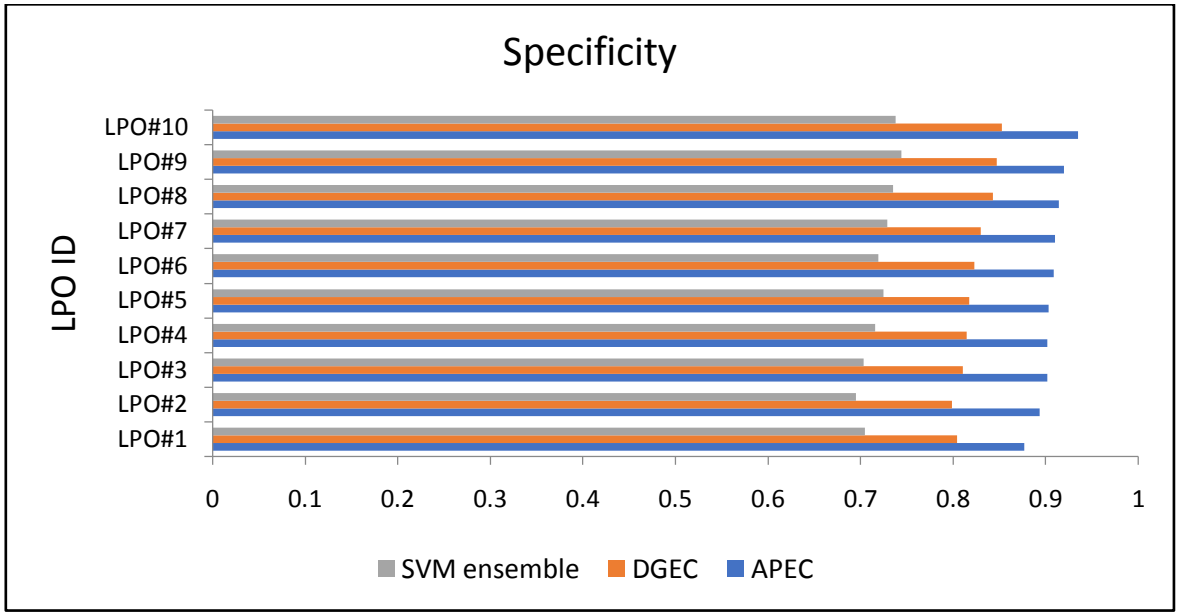

Figure 2: Specificity at 10-LPOs observed for the APEC DGEC, SVM-Ensemble method

The proposed APEC is compared with contemporary models DGEC and SVM-Ensemble method by plotting a graph between specificity and 10-LPOs over these methods, as depicted in Figure 2. The average and standard-deviation of specificity for the APEC, compared to the DGEC and SVM-Ensemble method, are 0.9066±0.01478, 0.82423±0.017607, and 0.72099±0.015409, respectively. It is exhibited from the statistics that the APEC performs better than the DGEC and SVM-Ensemble methods.

4.3. Sensitivity

	Sensitivity									
	LPO# 1	LPO# 2	LPO# 3	LPO# 4	LPO# 5	LPO# 6	LPO# 7	LPO# 8	LPO# 9	LPO#1 0
APEC	0.926 3	0.916 2	0.901	0.914 2	0.913 2	0.897 5	0.925 9	0.905 3	0.914 2	0.9114
DGEC	0.831 9	0.837 6	0.827 9	0.849 7	0.860 2	0.849 2	0.857 9	0.839 3	0.850 8	0.8423
SVM ensemble	0.707 5	0.739	0.726 8	0.715 4	0.712 5	0.759 5	0.754	0.743	0.744	0.7702

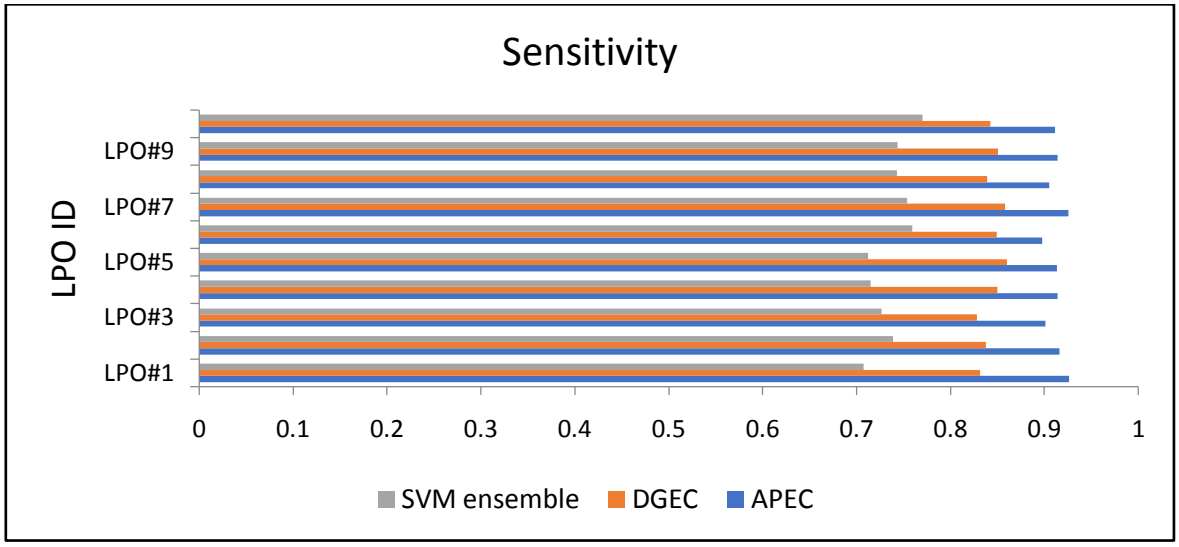

Figure 3: Sensitivity at 10-LPOs observed for the APEC, DGEC, SVM-Ensemble method

Figure 3 shows a graph drawn between the 10-LPOs and the sensitivity metric for the proposed APEC and the contemporary models DGEC and SVM-Ensemble method. From the graph, it is depicted that, average and standard-deviation of sensitivity for the APEC is 0.91252±0.008946. Here, the APEC is compared with DGEC and SVM-Ensemble method, whose average and standard-deviations are 0.84468±0.010141 and 0.73719±0.020068 in respective order. It is concluded from the above-stated statistics that the APEC is more significant when compared with DGEC and SVM-Ensemble methods.

4.4. Accuracy

	Accuracy									
	LPO# 1	LPO# 2	LPO# 3	LPO# 4	LPO# 5	LPO# 6	LPO# 7	LPO# 8	LPO# 9	LPO#1 0
APEC	0.9049	0.9062	0.9014	0.9088	0.909	0.9024	0.9191	0.9093	0.9163	0.9218
DGEC	0.8199	0.8208	0.8202	0.8346	0.8417	0.8379	0.8458	0.8408	0.8494	0.847
SVM ensemble	0.7064	0.7199	0.7167	0.7155	0.7178	0.7421	0.7432	0.7397	0.7441	0.7563

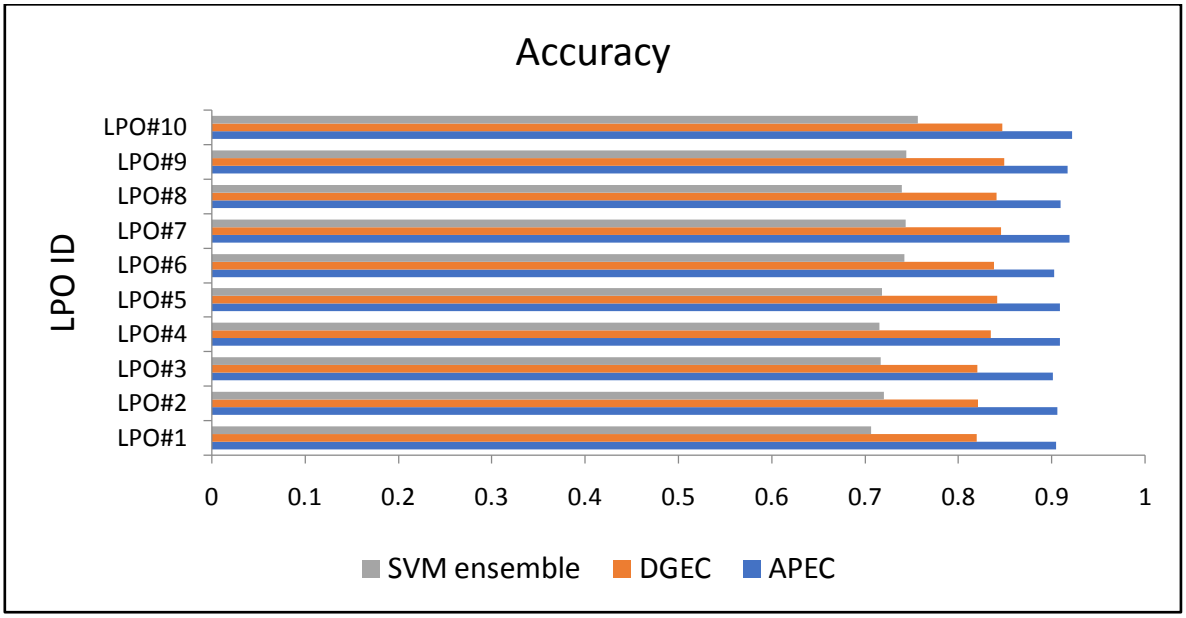

Figure 4: Accuracy at 10-LPOs observed for the APEC DGEC, SVM-Ensemble method

Figure 4 depicts the APEC's accuracy compared to the DGEC and SVM-Ensemble methods, as depicted in Figure 4. The average and standard-deviation of accuracy for the APEC, DGEC, and SVM-Ensemble methods are 0.90997±0.006659, 0.83581±0.010942, and 0.73017±0.015811, respectively. It is exhibited from the APEC statistics that it performs better compared to the DGEC and SVM-Ensemble methods.

4.5. F-measure

	F-measure									
	LPO# 1	LPO# 2	LPO# 3	LPO# 4	LPO# 5	LPO# 6	LPO# 7	LPO# 8	LPO# 9	LPO#10
APEC	0.8919	0.9052	0.9121	0.9124	0.9140	0.9179	0.9202	0.9233	0.9283	0.9416
DGEC	0.8249	0.8208	0.8295	0.8351	0.8388	0.8423	0.8483	0.8579	0.8628	0.8669
SVM ensemble	0.7298	0.7255	0.7317	0.7397	0.7468	0.7478	0.7551	0.7592	0.7662	0.7643

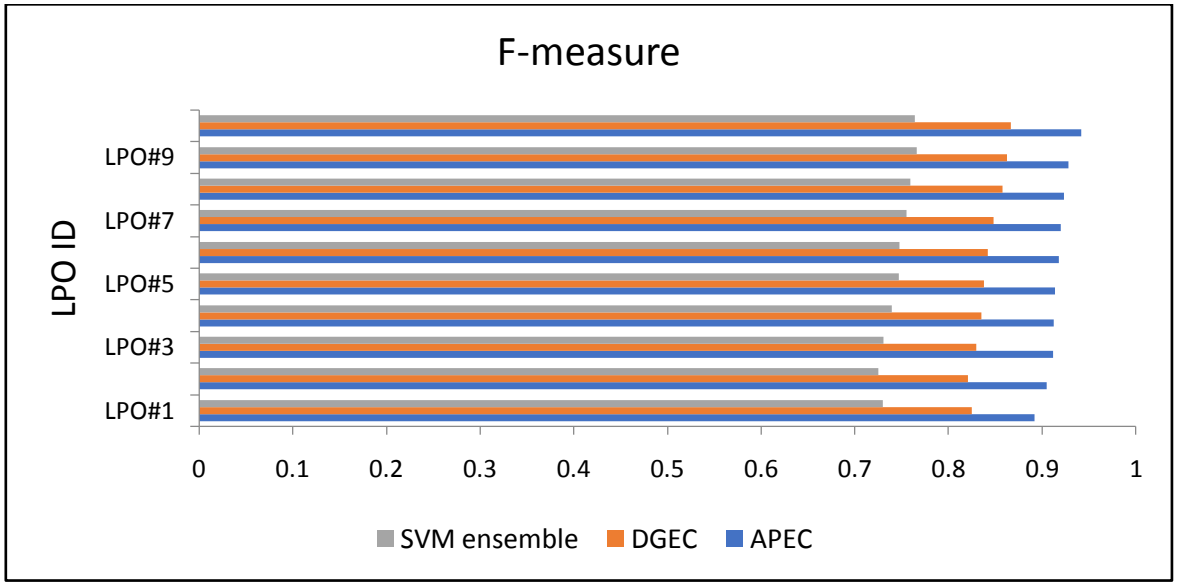

Figure 5: F-measure at 10-LPOs observed for the APEC DGEC, SVM-Ensemble method

You can see the correlation among 10-LPOs and F-measure metrics for the APEC, DGEC, as well as SVM-Ensemble approaches by looking at the graph in Figure 5. From the graph, it is depicted that, average and standard-deviation of the F-measure for the APEC is 0.91669±0.012673. Here, the APEC is compared with DGEC and SVM-Ensemble method, whose average and standard-deviations are 0.84262±0.015176 and 0.74657±0.014001 in respective order. It is concluded from the above-stated statistics that the APEC is more significant when compared with DGEC and SVM-Ensemble methods.

4.6. Matthews's correlation coefficient (MCC)

Matthews correlation coefficient										
	LPO# 1	LPO# 2	LPO# 3	LPO# 4	LPO# 5	LPO# 6	LPO# 7	LPO# 8	LPO# 9	LPO#1 0
APEC	0.806 2	0.809 2	0.800 4	0.814 8	0.815 3	0.803	0.835 6	0.816 7	0.831 7	0.8427
DGEC	0.634 7	0.635 9	0.635 9	0.664	0.677 9	0.671 9	0.686 9	0.678 7	0.695 5	0.6917
SVM ensemble	0.409 6	0.432 6	0.428	0.428	0.433 9	0.477 4	0.480 9	0.475 5	0.485	0.5066

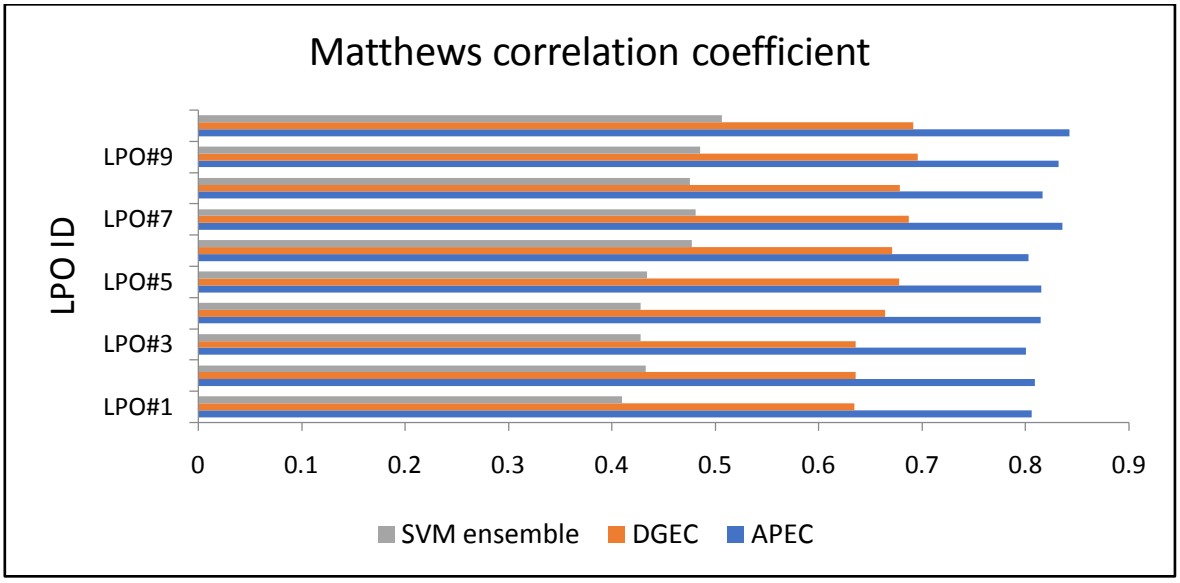

Figure 6: Matthews's correlation coefficient at 10-LPOs observed for the APEC DGEC, SVM-Ensemble method

Figure 6, the proposed APEC, is compared to the DGEC and SVM-Ensemble methods concerning MCC metric values observed from 10-LPOS. The average and standard-deviation of MCC for the APEC, DGEC, and SVM-Ensemble method are 0.81756±0.013698, 0.66722±0.022531 0.45575±0.031009, respectively. It is exhibited from the statistics that the APEC performs better than the DGEC and SVM-Ensemble methods.

Table 2: The table exhibiting the mean value of the cross-validation metrics

Mean value of the cross-validation metrics			
	APEC	DGEC	SVM ensemble
Precision	0.92701±0.010467	0.86188±0.012548	0.77406±0.012724
Specificity	0.9066±0.01478	0.82423±0.017607	0.72099±0.015409
Sensitivity	0.91252±0.008946	0.84468±0.010141	0.73719±0.020068
Accuracy	0.90997±0.006659	0.83581±0.010942	0.73017±0.015811
F-measure	0.91669±0.012673	0.84262±0.015176	0.74657±0.014001
Matthews correlation coefficient	0.81756±0.013698	0.66722±0.022531	0.45575±0.031009

Table 1 is exhibiting the mean values and respective deviations of the cross-validation metrics. The values obtained from APEC are leave behind the other two existing models. The sensitivity and

specificity of the proposed APEC method are exhibiting that both labels' detection accuracy is robust and far better than the contemporary models. The APEC model has approximately 7% more accuracy than DGEC and 12% more than SVM-ensemble.

5. Conclusion

This study uses the cross-media features of ECG data to forecast arrhythmias. The proposed method uses cross-media features to divide the training corpus into different clusters, in contrast to current models, which are solely focused on the format of features (demographic, signal patterns, or image features). One cluster's entries can show up in another's. Cross-media ECG signals are used to extract the best features from each cluster, which is then handled as a corpus. The fusion of diversity evaluation measures advises identifying the optimal cross-media characteristics for each cluster. In order to train the classifier, the best cross-media characteristics are employed. To train several clusters, different classifier objects are employed. The experimental examination of proposed and current techniques shows the relevance and effectiveness of APEC in detecting arrhythmia scope in contrast to SVM-ensemble and DGEC. In the future, the accuracy of arrhythmia prediction may be improved by the application of evolutionary methods.

Acknowledgement:

Aarathi has made significant contributions to the paper's intellectual content through her work on the paper's conception and design, data collection and analysis, and draught and critical revision.

Vasundhura has given her official confirmation of the article as well as consented to be acknowledged for the work's accuracy.

References

[1] World Heart Federation Report https://www.world-heart-federation.org/wp-content/uploads/2018/03/WHF-Report-2016.pdf.

[2] Dr. S. Vasundra, M.Dhana Lakshmi. “Predicting the treatment time in hospitals using SVM algorithm”, *JETIR*, vol-5, issue-6, ISSN: 2349-5162. (UGC approved).2018. https://www.jetir.org/papers/JETIR1806338.pdf.

[3] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." *In 2008 IEEE/ACS international conference on computer systems*

and applications, pp. 108-115. IEEE, 31st March 2008, DOI: 10.1109/AICCSA.2008.4493524.

[4] Ozcift, Akin, and Arif Gulten. "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms." *Computer methods and programs in biomedicine,*vol. 104, no.3,1st December (2011), pp. 443-451, https://doi.org/10.1016/j.cmpb.2011.03.018.

[5] Ghumbre, Shashikant U., and Ashok A. Ghatol. "Heart disease diagnosis using machine learning algorithm." *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012* (INDIA 2012) *held in Visakhapatnam,* India, January 2012. Springer, Berlin, Heidelberg, January 2012, pp. 217–225, DOI: 10.1007/978-3-642-27443-5_25.

[6] Austin, Peter C., et al. "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes." *Journal of clinical epidemiology,*vol. 66, no.4, (2013), pp. 398-407, https://doi.org/10.1016/j.jclinepi.2012.11.008.

[7] Pandey, Atul Kumar, et al. "A heart disease prediction model using the decision tree." *IOSR Journal of Computer Engineering (IOSR-JCE),*vol. 12, no.6,1st July (2013), pp. 83-86, https://www.iosrjournals.org/iosr-jce/papers/Vol12-issue6/N01268386.pdf.

[8] Ismaeel, Salam, Ali Miri, and Dharmendra Chourishi. "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis." *2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015),*IEEE, 31st May 2015, pp. 1–3, DOI: 10.1109/IHTC.2015.7238043.

[9] El-Bialy, Randa, et al. "Feature analysis of coronary artery heart disease data sets." *Procedia Computer Science,*vol. 65, 1st January (2015), pp. 459-468, https://doi.org/10.1016/j.procs.2015.09.132.

[10] Rajkumar, R., K. Anandakumar, and A. Bharathi. "Coronary artery disease (CAD) prediction and classification-a survey." *Breast Cancer,*vol. 90, (2006), pp. 94-35, http://www.arpnjournals.org/jeas/research_papers/rp_2016/jeas_0516_4179.pdf.

[11] Lo, Ying-Tsang, Hamido Fujita, and Tun-Wen Pai. "Prediction of coronary artery disease based on ensemble learning approaches and co-expressed observations." *Journal of Mechanics in Medicine and Biology,*vol. 16, no.01,23rd February (2016), pp. 1640010, https://doi.org/10.1142/S0219519416400108.

[12] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." *Artificial Intelligence in medicine*,vol. 23, no.1, 1st August (2001), pp. 89-109, https://doi.org/10.1016/S0933-3657(01)00077-X.

[13] JSoni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications*,vol. 17, no.8,8h March (2011), pp. 43-48, https://www.ijcaonline.org/volume17/number8/pxc3872860.pdf.

[14] Kadi, Ilham, Ali Idri, and J. L. Fernandez-Aleman. "Knowledge discovery in cardiology: A systematic literature review." *International journal of medical informatics*,vol. 97, 1st January (2017), pp. 12-32, https://doi.org/10.1016/j.ijmedinf.2016.09.005.

[15] Thenmozhi, K., and P. Deepika. "Heart disease prediction using classification with different decision tree techniques." *International Journal of Engineering Research and General Science*,vol. 2, no.6,2nd October (2014), pp. 6-11, http://ijergs.org.managewebsiteportal.com/files/documents/HEART-1.pdf.

[16] Soni, Jyoti, et al. "Intelligent and effective heart disease prediction system using weighted associative classifiers." *International Journal on Computer Science and Engineering*,vol. 3, no. 6, (2011), pp. 2385-2392, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.6636&rep=rep1&type=pdf.

[17] Guru, Niti, Anil Dahiya, and Navin Rajpal. "Decision support system for heart disease diagnosis using neural network." *Delhi Business Review*,vol. 8, no.1,8th January (2007), pp. 99-101, DOI: 10.1109/ICCCT.2010.5640377.

[18] Fida, Benish, et al. "Heart disease classification ensemble optimization using genetic algorithm." *2011 IEEE 14th International Multitopic Conference,IEEE,22nd December 2011*, pp. 19–24, DOI: 10.1109/INMIC.2011.6151471.

[19] Singh, Jagwant, and Rajinder Kaur. "Cardio vascular disease classification ensemble optimization using genetic algorithm and neural network." *Indian J Sci Technol*,vol. 9 S1, 9th December(2016), pp. 1-5, DOI: 10.17485/ijst/2016/v9i(S1)/98900.

[20] Uyar, Kaan, and Ahmet İlhan. "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks." *Procedia computer science*,vol. 120,1st January (2017), pp. 588-593, https://doi.org/10.1016/j.procs.2017.11.283.

[21] Liu, Xiao, et al. "A hybrid classification system for heart disease diagnosis based on the RFRS method." *Computational and mathematical methods in medicine*,vol. 2017, 3rd January (2017), pp. 1–11, https://doi.org/10.1155/2017/8272091.

[22] Nahar, Jesmin, et al. "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach." *Expert Systems with Applications*,vol. 40, no.1,1st January (2013), pp. 96-104, https://doi.org/10.1016/j.eswa.2012.07.032.

[23] Patil, Shantakumar B., and Y. S. Kumaraswamy. "Extraction of significant patterns from heart disease warehouses for heart attack prediction." *IJCSNS*,vol. 9, no.2,28th February (2009), pp. 228-235, doi=10.1.1.554.7404.

[24] Kuruba, Chandrakala, N. Pushpalatha, Gandikota Ramu, I. Suneetha, M. Rudra Kumar, et al., "Data mining and deep learning-based hybrid health care application." *Applied Nanoscience*, 6th Febuary (2022): 1-7, DOI: https://doi.org/10.1007/s13204-021-02333-1.

[25] Parthiban, Latha, and R. Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm." *International Journal of Biological, Biomedical and - Medical Sciences*,vol. 1, no.5, (2007), pp. 278-281, doi.org/10.5281/zenodo.1082439.

[26] Lakshmi, T. Naga, S. Jyothi, and M. Rudra Kumar. "Image Encryption Algorithms Using Machine Learning and Deep Learning Techniques—A Survey." *In Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough*, pp. 507-515. Springer, Cham, 2021, DOI: 10.1007/978-3-030-68291-0_40.

[27] Chauhan, Shraddha, and Bani T. Aeri. "The rising incidence of cardiovascular diseases in India: Assessing its economic impact." *Journal Preventive Cardiology*,vol. 4, no.4, (2015), pp. 735-740, https://www.researchgate.net/profile/Bani-Aeri/publication/313758654_The_rising_incidence_of_cardiovascular_diseases_in_India_Assessing_its_economic_impact/links/58a4fbb292851cf0e39306e2/The-rising-incidence-of-cardiovascular-diseases-in-India-Assessing-its-economic-impact.pdf.

[28] Samhitha, B. Keerthi, et al. "Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms." *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, 28th July 2020, pp 1326-1330, DOI: 10.1109/ICCSP48568.2020.9182303.

[29] Verma, Luxmi, Sangeet Srivastava, and P. C. Negi. "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data." *Journal of medical systems*,40.7 (2016): 1-7, DOI: 10.1007/s10916-016-0536-z.

[30] Dietterich, Thomas G. "Ensemble methods in machine learning." *In International workshop on multiple classifier systems*, pp. 1-15. Springer, Berlin, Heidelberg, 21st June 2000, DOI: 10.1007/3-540-45014-9_1.

[31] Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of computer and system sciences*,vol. 55, no.1,1st August (1997), pp. 119-139, https://doi.org/10.1006/jcss.1997.1504.

[32] Breiman, Leo. "Bagging predictors." *Machine learning*,vol. 24, no.2,August (1996), pp. 123-140, https://doi.org/10.1007/BF00058655.

[33] Dietterich, Thomas G., and Ghulum Bakiri. "Solving multiclass learning problems via error-correcting output codes." *Journal of artificial intelligence research*,vol. 2, (1994), pp. 263-286, DOI: https://doi.org/10.1613/jair.105.

[34] Duin, Robert PW, and David MJ Tax. "Experiments with classifier combining rules." *International Workshop on Multiple Classifier Systems, Springer, Berlin, Heidelberg, 21st June 2000*, pp. 16–29, DOI: 10.1007/3-540-45014-9_2.

[35] Waske, Björn, and Jon Atli Benediktsson. "Fusion of support vector machines for classification of multisensor data." *IEEE Transactions on geoscience and remote sensing*,vol. 45, no.12,19th November (2007), pp. 3858-3866, DOI: 10.1109/TGRS.2007.898446.

[36] Kadam, Vinod, Shivajirao Jadhav, and Samir Yadav. "Bagging based ensemble of support vector machines with improved elitist GA-SVM features selection for cardiac arrhythmia classification." *International Journal of Hybrid Intelligent Systems*, vol.16, no. 1, 1st January (2020),pp. 25-33, DOI: 10.3233/HIS-190276.

[37] Pławiak, Paweł, and U. Rajendra Acharya. "Novel deep genetic ensemble of classifiers for arrhythmia detection using ECG signals." *Neural Computing and Applications*, vol. 32, no. 15,August (2020), pp. 11137-11161, https://doi.org/10.1007/s00521-018-03980-2.

[38] Zhang, Cha, and Yunqian Ma, eds. "Ensemble machine learning: methods and applications."*Springer Science & Business Media*, 17th March 2012, pp. 1–34, DOIhttps://doi.org/10.1007/978-1-4419-9326-7.

[39] Moody, George B., and Roger G. Mark. "The impact of the MIT-BIH arrhythmia database." *IEEE Engineering in Medicine and Biology Magazine*,vol. 20, no.3, May (2001), pp. 45-50, DOI: 10.1109/51.932724.

[40] Strodthoff, Nils, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL." *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, 9th September (2020): 1519-1528, DOI: 10.1109/JBHI.2020.3022989.

[41] https://sleepdata.org/datasets/shhs

[42] Goldstein, Mark R. "Sudden death due to cardiac arrhythmias." *The New England journal of medicine*,vol. 346, no.12,15th November (2002), pp. 946-7, DOI: 10.1056/NEJMra000650.

[43] Kadambe, Shubha, Robin Murray, and G. Faye Boudreaux-Bartels. "Wavelet transform-based QRS complex detector." *IEEE Transactions on biomedical Engineering*,vol. 46, no.7,July (1999), pp. 838-848, DOI: 10.1109/10.771194.

[44] Ye, Can, BVK Vijaya Kumar, and Miguel Tavares Coimbra. "Heartbeat classification using morphological and dynamic features of ECG signals." *IEEE Transactions on Biomedical Engineering*,vol. 59, no.10,15th August (2012), pp. 2930-2941, DOI: 10.1109/TBME.2012.2213253.

[45] Addison, Paul S. "Wavelet transforms and the ECG: a review." *Physiological measurement*,vol. 26, no.5, 8th August (2005), pp. R155-R199, https://iopscience.iop.org/article/10.1088/0967-3334/26/5/R01/meta.

[46] Jech, Thomas. Set theory. *Springer Science & Business Media, 2013*, DOIhttps://doi.org/10.1007/3-540-44761-X.

[47] Poungponsri, Suranai, and Xiao-Hua Yu. "An adaptive filtering approach for electrocardiogram (ECG) signal noise reduction using neural networks." *Neurocomputing*,vol. 117,6th October (2013), pp. 206-213, https://doi.org/10.1016/j.neucom.2013.02.010.

[48] Pan, Jiapu, and Willis J. Tompkins. "A real-time QRS detection algorithm." *IEEE transactions on biomedical engineering*,vol. BME-32, no. 3, (1985): 230-236, DOI: 10.1109/TBME.1985.325532.

[49] De Chazal, Philip, Maria O'Dwyer, and Richard B. Reilly. "Automatic classification of heartbeats using ECG morphology and heartbeat interval features." *IEEE transactions on biomedical engineering*,vol. 51, no.7, (2004), pp. 1196-1206, DOI: 10.1109/TBME.2004.827359.

[50] Ghosh, Madhumala, Devkumar Das, and Chandan Chakraborty. "Entropy based divergence for leukocyte image segmentation." 2010 International Conference on Systems in Medicine and Biology,*IEEE, 16th December 2010*, pp. 409-413, DOI: 10.1109/ICSMB.2010.5735414.

[51] Pharwaha, Amar Partap Singh, and Baljit Singh. "Shannon and non-shannon measures of entropy for statistical texture feature extraction in digitized mammograms." *Proceedings of the world congress on engineering and computer science*,vol. 2. 20th October 2009, pp. 20-22, http://www.iaeng.org/publication/WCECS2009/WCECS2009_pp1286-1291.pdf.

[52] Klebanov, Igor R., et al. "Rényi entropies for free field theories." *Journal of High Energy Physics*,vol. 2012, no. 74,April (2012), pp.1-28, https://doi.org/10.1007/JHEP04(2012)074.

[53] Pavesic, Nikola, and Slobodan Ribaric. "Gray level thresholding using the Havrda and Charvat entropy." *In 2000 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries. Proceedings*,MeleCon 2000 (Cat. No. 00CH37099), vol. 2, pp. 631-634. IEEE, 29th May 2000, pp. 631-634, DOI: 10.1109/MELCON.2000.880013.

[54] Sahoo, Prasanna K., and Gurdial Arora. "Image thresholding using two-dimensional Tsallis–Havrda–Charvát entropy." *Pattern recognition letters*, 27.6, 15th August (2006), 520-528, https://doi.org/10.1016/j.patrec.2005.09.017.

[55] Raja, N., et al. "Segmentation of breast thermal images using Kapur's entropy and hidden Markov random field." *Journal of Medical Imaging and Health Informatics*,col. vol. 7, no.8,1st December (2017), pp. 1825-1829, DOI: https://doi.org/10.1166/jmihi.2017.2267.

[56] Li, Yin, Jian Tao, and Yazhi Song. "An Incremental-Hybrid-Yager's Entropy Model for Dynamic Portfolio Selection with Fuzzy Variable." *Discrete Dynamics in Nature and Society*, 2018, 1st January (2018), https://doi.org/10.1155/2018/7387210.

[57] Gonzalez, R. C. Processing, (2002).

[58] Das, Devkumar, et al. &quot. "Invariant moment based feature analysis for abnormal erythrocyte recognition".*International Conference on Systems in Medicine and Biology*,IEEE, 16th December 2010, pp: 242-247, DOI: 10.1109/ICSMB.2010.5735380.

[59] Rey, Denise, and Markus Neuhäuser. "Wilcoxon-signed-rank test." *International encyclopedia of statistical science.Springer Berlin Heidelberg, 2011*. 1658-1659, https://doi.org/10.1002/9780471462422.eoct979.

[60] Ghasemi, Asghar, and Saleh Zahediasl. "Normality tests for statistical analysis: a guide for non-statisticians." *International journal of endocrinology and metabolism*,vol. 10, no.2, (2012), pp. 486, doi: 10.5812/ijem.3505.

[61] McKnight, Patrick E., and Julius Najab. "Mann Whitney U Test." *The Corsini encyclopedia of psychology*,30th January (2010), pp. 1-1, https://doi.org/10.1002/9780470479216.corpsy0524.

[62] Budak, Hüseyin, and Semra Erpolat Taşabat. "A modified t-score for feature selection." *Anadolu University Journal of Science and Technology A-Applied Sciences and Engineering*, vol. 17, no. 5, (2016), pp. 845-852, https://doi.org/10.18038/aubtda.279853.

[63] t-table. http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf. 2017.

[64] Shasidhar, M., V. Sudheer Raja, and B. Vijay Kumar. "MRI brain image segmentation using modified fuzzy c-means clustering algorithm." *2011 International Conference on Communication Systems and Network Technologies*,Katra, India, IEEE, 3rd June 2011, pp 473-478m DOI: 10.1109/CSNT.2011.102.

[65] S. Vasundra, C. Swathi. "A Fuzzy C-Means Based Feature Selection to Process Medical Data." *International Journal on Recent and Innovation Trends in Computing and Communication*,5, no. 7, (2017), pp. 609-612, DOI: https://doi.org/10.17762/ijritcc.v5i7.1096.

[66] Abdualrhman, Mohammed Ahmed Ali, and M. C. Padma. "CS-IBC: Cuckoo search based incremental binary classifier for data streams*." Journal of King Saud University-Computer and Information Sciences*,vol. 31, no.3, 1st July (2019), pp. 367-377, https://doi.org/10.1016/j.jksuci.2017.05.008.

[67] https://www.python.org/

[68] Airola, Antti, et al. "An experimental comparison of cross-validation techniques for estimating the area under the ROC curve." *Computational Statistics & Data Analysis*,vol. 55, no.4, 1st April (2011), pp. 1828-1844, https://doi.org/10.1016/j.csda.2010.11.018.