How to Cite:
Jyothi, S., & Geethanjali, N. (2022). Arrythmia prediction from high dimensional electrocardiogram's data corpus using ensemble classification. *International Journal of Health Sciences*, *6*(S1), 4790–4810. https://doi.org/10.53730/ijhs.v6nS1.5898

Arrythmia prediction from high dimensional electrocardiogram's Data Corpus using ensemble classification

Sreedhar Jyothi
Research Scholar, Department of Computer Science & Technology, S.K. University, Ananthapuramu - 515003
Email: jsreedharcs@yahoo.com

N. Geethanjali
Professor, Department of Computer Science & Technology, S.K. University, Ananthapuramu - 515003
Email: geethanjali.sku@gmail.com

***Abstract*---**In clinical practice, software aided arrhythmia diagnosis from electrocardiographic signals is critical, and it has the capability to minimize mortality induced by untrained clinicians. Furthermore, computer-assisted methods are generally successful in detecting arrhythmia extent from ECG readings early. The buzzword in computer-assisted clinical settings is branch of artificial intelligence. Computer-assisted arrhythmia forecasting approaches, particularly, are widely used machine learning methodologies. Most recent research is focused on the utilization of high-dimensional learning data sets to build machine learning models. The large dimensions of data points used for the machine learning techniques, on the other hand, frequently leads to false alarms. Though the few contemporary models endeavored to handle this by using multiple classifiers as ensemble model, they evince improved decision accuracy when trained on high volume of data. They do, however, frequently exhibit significant false alerting, with the training data representing the high dimensional data points of the enormous amount of training data provided. This paper discussed an ensemble learning approach that selects optimal subset of data-points by fusing diversity evaluation method.

***Keywords*---**Arrythmia Prediction, Electrocardiogram, KS-test, classification, ECG Heartbeat Categorization Dataset, MIT-BIH, Mann-Whitney U Test.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Corresponding author: *Jyothi, S.; Email: jsreedharcs@yahoo.com*
Manuscript submitted: 27 Feb 2022, Manuscript revised: 18 March 2022, Accepted for publication: 09 April 2022

Introduction

Leading cause of worldwide deaths recognized by WHO is cardiovascular diseases (CVDs) [1]. In rent past variety of policies and programs introduced across the broader range of areas to help reduce the number of new and recurring cardiac events. Since then, the electrocardiogram (ECG) has risen in popularity as a tool for early diagnosis of cardiovascular disease (CVD). Heart illness and irregularities can be detected with the help of an electrocardiogram (ECG), a visual picture of electrical impulses in the heart. For more than 70 years, doctors have used electrocardiogram (ECG) signals to identify heart conditions like arrhythmias and myocardial infarction. P, QRS complex, and T waves [3-5] make up an electrocardiogram signal. A U wave could also be present. Many heart illnesses can be detected by studying the fluctuations in these impulses. ECG equipment are low-risk and low-cost options for cardiac monitoring. Spikes in Electrocardiogram readings can, nevertheless, be caused by noise as well as other variables known as artefacts. Patients' physical movement, electrode motions on the body, as well as power line disruptions are examples of artefacts. To achieve precise ECG studies, noise and artefacts from the Electrocardiogram must be eliminated. ECG data are preprocessed with several transforms to reduce noise and artefacts, with the wavelet analysis being the most often used transform [6]. To achieve noise- as well as artifact-free ECG readings, several techniques for detecting the P, QRS complex, and T waves have been previously described. These algorithms have also been verified across the MIT-BIH electrocardiogram dataset [8–13]. Using the ratios among neighboring signals, an algorithm has been proposed in [8] for the recognition of R-peaks that exploits the existence of R-peaks like a cue. This algorithm only uses two datasets o electrocardiogram signals and is more sophisticated. It was presented in [9] to use a technique based on observable mode segmentation and the Hilbert transform to find Electrocardiogram signals' R-peaks. Nevertheless, this approach is difficult to understand and uses a huge number of nodes to detect R-peaks. Furthermore, only R-peaks can be detected using either one of these techniques. Other contemporary contributions [10–13] presented some techniques basis of two event-related moving-averages to detect P and T-peaks together with R-peaks. Filtration, augmenting, BOI creation with each peak, as well as thresholding are all components among these algorithms. To enhance high values and reduce tiny values, such algorithms use a Butterworth technique to filter the electrocardiogram signals. The measurement results are then squared. Windows are lengthd according to the QRS wave's amplitude and recurrence intervals once it has been enhanced. Moving-averages are then used to calculate BOI at each peak. Depending on the time window, the width per each chunk is computed and compared to a threshold. Afterwards, each chunk's peaks being located. OPEN specimens are those that are taken in between any identified R-peaks, such as any R-peak specimens, in order to locate P as well as T-peaks. When it comes to peak recognition, this method performs admirably.

Review of Literature

A contemporary model has used neural networks to classify Electrocardiogram arrhythmias [14]. The obtained electrocardiogram signal is processed, and data-points are extracted using the wavelet transformation (DWT). A recurrent neural network neural network is used to classify the data. For the machine-learning classification, electrocardiogram signals from the MIT-BIH database have been employed. Ten files, including two arrhythmias, were used in the testing and training. The classifier is sensitive enough to detect two different types of arrhythmias with an accuracy ratio 0.965.

Using a convolutional neural network, the other contribution [15] classified the electrocardiogram signals. They investigated the amplitude of Electrocardiogram intervals, such as the PR interval, ST segments, QRS complex, and ST interval, by extracting various aspects from the electrocardiogram signals. The pulse rate has been used as a criterion for identifying cardiac disease. The information was gathered out from MIT-BIH dataset.

The contemporary classification model [16] examined the electrocardiogram signal prediction performance of various machine learning techniques. They used a database that contained 50 electrocardiogram signals from 50 subjects. To reduce noise from the signal, the Wavelet transform was applied. The complete dataset is separated into three pieces for training, testing, and validation after normalizing. RBF surpassed MLP with an accuracy ratio 0.94, according to the study.

RS and subatomic artificial neural network has been used by the other contemporary classification model [17] classified the Electrocardiogram signals (QNN). The signals were first homogenized, and then the wavelet transform was used to extract the data-points. The RS reduction technique is used to remove redundant properties. After attribute reduction, QNN-based classification modelling is performed. According to the findings, classification employing RS–QNN is superior to traditional approaches.

The neural network contribution [18] employed classifiers of different forms of neural networks to classify electrocardiogram signals: the multilayered convolutional classifier and the neural networks of vector quantization learning. The two classifiers have been kept to the test with Electrocardiogram, and the MLP classifier outperformed the neural network of Vector Quantization.

The other classification technique [19] used an artificial neural network to classify electrocardiogram signals (ANN). The research is based on data from the MIT-BIH arrhythmia database. The electrocardiogram signal is preprocessed using the Wavelet transform. By identifying the QRS complex of an electrocardiogram signal, several problems in the signal can be predicted. To detect abnormalities in the electrocardiogram signal, a GUI was created.

An artificial neural network has portrayed [20] to evaluate electrocardiogram signals. The study is based on electrocardiogram signal parameters such as the Poincare plot, Lyapunov exponent, and spectral entropy. The anomalies in the Electrocardiogram are detected using neural network.

The contemporary classification technique [21] used neural networks to classify cardiac arrhythmias. The classification is done via incremental back propagation and data-point selection based on correlation. The information has been obtained from the UCI dataset. With 100 simulations, an accuracy ratio 0.8771 is attained.

The contemporary classification model [22], [23] used a relevance vector machine to detect and classify Electrocardiogram arrhythmia. Cardiac beats are automatically classified using a relevance vector machine. The data for the experiment came from the MIT-BIH database. Signal processing methods are used to extract data-points.

A multilayered deep neural network to classify heartbeats [24] in an electrocardiogram signal, CNN recognizes distinct types of heartbeats automatically. An accuracy ratio of 0.9403 has been achieved after removal of the noise from a open access dataset.

A new deep learning strategy [25] for classifying cardiac arrhythmia efficiently has been portrayed in contemporary literature. The study made use of electrocardiograms of forty-five subjects from the MIT-BIH dataset. The Deep "1D-convolution neural network has been utilized, and the accuracy ratio is 0.9133.

Using an evolving neural network strategy, a classification strategy [26] has been established a technique that enabling efficient classification of diversified kinds of cardiovascular diseases by using the corpus of electrocardiogram signals. The study made use of electrocardiogram signals of forty-five subjects from the MIT-BIH dataset. The study is carried out by creating an evolving neural network model by using the SVM classifier that has portrayed accuracy ratio of 0.9885.

Using an evolving neural system, the other classification methodology [27] has been proposed that used an ensemble classification strategy to efficiently classify diversified cardiovascular diseases. The study was conducted utilizing electrocardiogram signals of the twenty-nine subjects from the MIT-BIH dataset.

The work [28] presents that ventricular extra ectopic or systole beats were classified with the morphology matching assistance, clustering algorithms & R-R intervals. The work [29] presents that 17 kinds of ECG beats were categorized by utilizing local hexadecimal patterns computed from sub-bands wavelet. The work [30] presents that 5 primary kinds of heart-beats were categorized by utilizing EEMD (empirical ensemble mode decomposition) based data-points exposed to SMO-SVM (sequential minimum optimization)-SVM. Also, NN acts as a prominent role in the biological analysis of signals [31]. Contemporarily, class-oriented strategies based on deep-learning came into existence. The schemes of deep learning are a part of ML schemes applied based on more hidden NN.

The work [32] presents that 17 kinds of heart-beats were categorized by utilizing ID-CNN & new three-layer ensemble deep genetic classifiers. In the subject-oriented method, the overall database of MIT-BIH is segmented into 5 clusters of heart-beats as per ANSI/AAMI. The list of these clusters is ventricular ectopic, supraventricular ectopic, unknown, non-ectopic & fusion. Again 2 schemes were perceived to categorize these divergent clusters: inter-patient & intra-patient strategies. The basic disagreement among these 2 schemes is the departure of testing & training datasets. The work [33] presents that models based on intra-patient strategy are explored extensively in this review. Nevertheless, these models have a minimum effect in real-world cases. Due to real-time implementations, the unrecognized subject, who undergoes generally testing, would be foreign towards the constructed method. Hence, the method is sufficient for capturing inter-individual changes among ECG. When the intra-patient method is designing, there is a scope of possessing common information subject in both testing & training. For mitigating such a problem, the work [34] classified heart-beat based on inter-patient. The entire database of MIT-BIH is segmented into 2 clusters. One cluster is allocated for training, and the other is to test by assuring no identical data is subjected in both clusters.

The goal of "Automatic detection of cardiac arrhythmia (ADCA)" [35] was to solve the limitations of current computer aided arrhythmia prognosis approaches. Despite the fact that the ADCA is indeed an ensemble model, it does not solve the false alerting created by the large dimensional data projecting the data points of the learning phase. “Cardiac Arrhythmia Detection Using Ensemble of Machine Learning Algorithms (EMLA)" [36], a recent ensemble strategy, acknowledged an innovative method of selection and optimization of data points to classify the given electrocardiographic signals as prone to or not. However, EMLA, overlooked the issue of false alarming caused by high dimensional data points. This manuscript described a unique ensemble classification technique that employs signal flow characteristics to reduce the high dimensionality in data points.

Methods and Materials

Model Description

The suggested method for predicting arrhythmias from electrocardiograms is a supervised learning strategy that learns from the ECG's ideal interval and axis peaks, which are determined using a fusion of diversity assessment criteria. The heuristic search technique called Incremental binary classifier [37] has adapted that performs the Arrhythmia Prediction in a hierarchical order. A method of weighing the interval and axis peaks of the ECG towards the diversified labels positive (prone to arrhythmia), negative (not prone to arrythmia) have been built under the Likert Scale's influence [38]. The gist of the approach is as follows.

The given electrocardiograms, which fall into one of the labels positive and negative, each record contains a set of interval and axis peaks of the ECG. The last column of each record engages the respective positive or negative label. Further, partitions the given labeled records of labels positive and negative into multiple groups such that each group contains the records of a specific positive or negative label.

Further, for each record of the group that representing one of the positive or negative label, discovers interval and axis peaks of the ECG, which includes emoticons concerning to identify optimal interval and axis peaks of the ECG towards each positive and negative label, the fusion of diversity assessment metrics shall apply on each length of the interval and axis peaks of the ECG records fall in one label positive with the other label negative corresponding column. According to the diversity observed, the corresponding column shall consider as optimal or not. The later phase derives interval and axis peaks of the ECG from the optimal interval and axis peaks of the ECG of each label, which shall use in learning phase of the classification process [37]. Finally, the proposal performs the Arrhythmia Prediction from the given electrocardiograms. Concerning to fusion of diversity measures, the proposal adopted two distribution diversity assessment measures called KS-Test (Kolmogorov–Smirnov test) [39], and MWU-Test (Mann-Whitney U Test) [40]

Data and characteristics

Cardiac arrhythmia is a disorder with the pace or frequency of a person's pulse, in which the pulse rate are abnormally fast, slow, or get a non-conducive sequence. The electrocardiogram signal sequence of the electrocardiogram. All of these signal sequences are further used to detect the pulse structure, is the structure of such data being considered. Such signals were then utilized to identify a variety of data-points, which were then categorized and referred to as tridimensional characteristics. In the next sections, we'll look at such dimensions including their characteristics.

Intervals

RR Intervals: This interval between sub consecutive Electrocardiogram R-waves associated with QRS signal, as well as its counterpart in terms of HR, which is dependent on specific properties with such a sinus node and situations with systemic influence.

PR Intervals: This is the time to begin the QRS-complex in P-Wave, which replicates on the AV node-based conduction. In regards to time, the usual PR interval approximately 120-200ms (0.12-0.20s). The existence of first-degree heart block is required if indeed the PR-interval has been greater than 200ms.

QRS Intervals: The QRS complex usually lasts 0.08 to 0.10 seconds, which is equivalent to 80 to 100 milliseconds. It is classified as transitional or somewhat protracted if the time extent ranges from 100ms to 120 milli seconds. Abnormal circumstances are defined as QRS durations above of 100 milli seconds.

Because numerous durations were recorded, QT intervals shall consider as the intrinsic parameter. The general QT interval is usually between 0.4 and 0.44 seconds. A1 longer QT interval can appear in female gender that compared to male gender patients, according to research. Longer QT intervals are also associated with lower pulse rate.

QTc intervals have been used to define the appropriate QTc, which can be either equal or less than 400 ms, 410 ms, 420 ms, or 440 ms. In sudden cardiac death situation, the threshold QTc is 0.431-0.45 sec in male gender patients and 0.47 sec in female gender patients. "Abnormal" QTc signal in men are QTc ranges between 0.45 s and 0.47 s in gender females.

Axis

In regard to overall electrical impulses of the heart, the Electrocardiogram axis is one of the most important directions. This can be conventional, rightward as well as leftward, or even undetermined criteria, such as the northwest-axis.

An atrial depolarization with sinus node, which designates sinoatrial node, provides an action that depolarizes the atria, is reflected in the degree of axis of P-Waves. Often P-wave shall be vertical within lead II if the SA node generates potential action.

Degree of QRS-Wave Axis: Determining the QRS axis is critical, with the usual QRS axis ranging from -30 to +90 degrees. Major QRS vector spanning between -30 and -90 degrees is referred to as left axis deviation. In the QRS axis, right axis deviation ranges between +90 and +180 degrees.

T-wave axis degree: Electrocardiogram signal's ventricular repolarization represents by T-wave. An absolute refractory period is defined as the time interval between the QRS-complex as well as the apex of the T-wave. In the Electrocardiogram, the T-wave is perhaps the most malleable-Wave. T-wave modifications occur in tandem due to T-Wave amplitude, which often peculiarly caplengthd T-Waves. This peculiar, caplengthd T-Waves can be caused by a variety of cardiovascular as well as non-cardiac cardiovascular diseases. Besides the type right-precordial leads, the normal T-Wave is normally in the same direction also as QRS.

However, either T-Wave or P-Wave axis, QRS axis and the limb leads reveal through these measurements, which are significant to analyze.. The normal QRS axis must be between -30 as well as +90 degrees. Left-axis divergence is defined as a main QRS vector ranging from -30 to -90, and right-axis divergence is ranging between +90 and +180. As a result, the undetermined axis range has been defined as $\{(+/-)190,-90\}$

Preprocessing

This step discards any input electrocardiograms in the training corpus that do not match one of the labels positive (arrhythmia) or negative (benign). The properties of formats mentioned as intervals CS and Axis of each input ECG in the supplied training corpus are also extracted. The different intervals $\{r\exists r\in CS\}$ stacked in the relevant ECG are represented by the record of intervals. The diverging axis formats are represented by the electrocardiogram's axis values $\{ar\exists ar\in AS\}$. Similarly, derives each electrocardiogram's signal $\{fo\exists fo\in FS\}$ record for the supplied corpus.

Mann-Whitney U Test

Mann-Whitney U Test (MWU-Test) [40] is among the multiple diversity assessment methods, which does not include centric to the distribution format that deserves in most of the datasets having recorded with diversified labels. The description of the MWU-Test implementation process is as follows:

The notations v_1, v_2 denote the data-point vector distributions used as input to the method MWU-Test to conclude the scope of diversity between corresponding data-point vectors, which is as follows:

Initially, all the entries of data-point vectors v_1, v_2 are moved to a new data-point vector v. Further, sort the data-point vector v in ascending order of the values and let the indices of the ordered values of the data-point vector v as corresponding ranks R. The average of the identical values' indices will be the rank of all the respective identical values. Further description denotes the ranks assigned to the data-point vector's values v_1 as a set R_1 and the ranks assigned to the data-point vector's values v_2 as a set R_2. Later the process finds the aggregate of the entries in the set R_1 as RS_1, which is further used to determine the rank-sum threshold RST_1 of the data-point vector v_1 as follows in (Eq 1):

$$RST_1 = RS_1 - \frac{|v_1| \times (|v_1| + 1)}{2} \quad \text{...(Eq 1)}$$

// the notation $|v_1|$ denotes the length of the data-point vector v_1.

Similarly, the rank-sum threshold RST_2 of the data-point vector v_2 will be determined as follows

$$RST_2 = RS_2 - \frac{|v_2| \times (|v_2| + 1)}{2} \quad \text{...(Eq 2)}$$

// In (Eq 2), notation $|v_2|$ denotes the length of the data-point vector v_2, and the notation RS_2 denotes the sum of the ranks of the entries in data-point vector v_2 those listed in a set R_2.

Then the rank-sum threshold RST of the data-point vectors' entries v_1, v_2 is the sum of rank-sum thresholds RST_1, RST_2 of the data-point vectors v_1, v_2 that are followed in (Eq 3).

$$RST = RST_1 + RST_2 \quad \text{...(Eq 3)}$$

Then find the z-score [41] as follows:

Initially, find the mean m_{RST} and standard deviation d_{RST} as follows in (Eq 4), (Eq 5):

$$m_{RST} = \frac{RST}{2} \quad \text{...(Eq 4)}$$

$$d_{RST} = \sqrt{\frac{|v_1| * |v_2| * (|v| + 1)}{|v|}} - \sqrt{\frac{|v_1| * |v_2|}{|v|}\left((|v| + 1) - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{|v| * (|v| - 1)}\right)} \quad \text{...(Eq 5)}$$

Here in (Eq 4), (Eq 5), the notation k denotes the number of distinct ranks, t_i denotes the number of entries sharing the same rank i
Further, the z-core assesses as follows in (Eq 6):

$$z = \frac{RST - m_{RST}}{d_{RST}} \quad ...(Eq\ 6)$$

Then, in the z-table [42], determine the p-value of the illustrated z-score. The distribution of the vectors v_1, v_2 is deemed to be varied whereas if p-value is bigger than the provided probable threshold (typically 0.1, 0.05, or 0.01). Aside from that, the distribution is quite consistent.

KS-test

Kolmogorov-Smirnov test (KS-test) [39] is a distribution diversity assessment measure, which has been used to assess the diversity between the values projected for each data-point attribute of given two datasets. The significance of the ks-test is that it can apply to assess the diversity between two data-point vectors of variable length $(|fv_a| \neq |fv_b|)$. The diversity assessment of two data-point vectors by KS-Test is as follows:

The given two data-point vectors fv_a, fv_b, representing the data-point values of data-point attribute in distinct datasets. The KS-Test will implement in concern to evaluate the distributions of 2 data-point vectors are similar or divergent as follows:

Assessing the aggregates $Ag(fv_a), Ag(fv_b)$ of the given two data-point vectors fv_a, fv_b and assessing cumulative ratio of each element exists in data-point vectors fv_a, fv_b is the initial process of the KS-Test (see Eq 7).

$$\begin{aligned}
&cr = 0 \\
&\overset{|fv_j|}{\underset{i=1}{\forall}} \left\{ el_i \exists el_i \in fv_j \right\} begin \\
&cr = \frac{el_i}{Ag(fv_j)} + cr \qquad ...(Eq\ 7) \\
&CR_{fv_j} \leftarrow cr \\
&end
\end{aligned}$$

The notations used in Eq 7 are,
The notation el_i denotes the data-point value representing the corresponding data-point attribute.
The expression $Ag(fv_j)$ indicates the total of the data-point values listed in data-point vector fv_j

The expression CR_{fv_j} indicates the set of aggregate ratios of the data-points $\left\{el \exists el \in fv_j\right\}$ of the vector fv_j.
Concerning to the aforesaid ks-test process, the cumulative ratios CR_{fv_a}, CR_{fv_b} of the values representing the data-point vectors fv_a, fv_b in respective order.

Further discovers the absolute Difference as a set $AD_{\left\{CR_{fv_a} \leftrightarrow CR_{fv_b}\right\}}$ of the cumulative ratios of the data-points listed in sets CR_{fv_a}, CR_{fv_b} in respective order, which is as follows.

```
$\overset{\max(|CR_{fv_a}|,|CR_{fv_b}|)}{\underset{i=1}{\forall}} \left\{cr_i(fv_a), cr_i(fv_b) \exists cr_i(fv_a) \in CR_{fv_a} \wedge cr_i(fv_b) \in CR_{fv_b}\right\}$ Begin // for all cumulative ratios exists in sets $CR_{fv_a}, CR_{fv_b}$
    $AD_{CR_{fv_a} \leftrightarrow CR_{fv_b}} \leftarrow abs\left(cr_i(fv_a) - cr_i(fv_b)\right)$ // discovering the absolute Difference $abs\left(cr_i(fv_a) - cr_i(fv_b)\right)$ as a set $AD_{\left\{CR_{fv_a} \leftrightarrow CR_{fv_b}\right\}}$ of the cumulative ratios of the data-points listed in sets $CR_{fv_a}, CR_{fv_b}$ in respective order
End
```

The maximum value of the set $AD_{CR_{fv_a} \leftrightarrow CRf_{v_b}}$ denotes further as d-stat helps to find diversity scope. If the d-stat is greater than the d-centric, then confirms the diversity of the given data-point vectors is poor, else the diversity of both vectors is significant. The aforesaid d-centric is the degree of probability threshold of the sets $Ag(fv_a), Ag(fv_b)$ that tracked from the KS-table [43].

The classifier

The classification has been carried by using incremental binary classifier [37], which is noted to be optimal. The subsequent sections have portrayed the process of supervised learning and class prediction phases of the classifier.

Supervised learning phase

The classification process is divided into two stages. This phase, known as training, builds a nest hierarchy in which each order has more perches that compare to the previous order. During the training process, two hierarchies are established: one for positive labels, and another for negative ones. In both hierarchies, the perches will be arranged as follows:

The n-grams of the data-points extracted from the respective label's training data will be sorted in the descending order of their length for both labels. N-gram data-points of maximum length should be divided into groups, so that all n-grams having same length and frequency can be found in one group. Groups with n-grams of length n will be placed as perches in the first order of their respective hierarchies. To do this with the n-gram data-points of length $\{(n-i)\exists 1 \leq i < n\}$, it's recommended that they be divided into groups with a variety of n-grams sharing the same frequency. Perched on order l. This process is repeated until the final order of the hierarchy. In order to categorize the n-grams of length one into groups, each group must contain a set of n-grams with the same frequency. The final order will house all of these groups as perches.

Classification

The unlabeled records' arrhythmia scope is predicted by the classification phase. Following is a breakdown of the classification procedure. The provided electrocardiograms will be preprocessed in order to extract the data-points (see sec. 3.2). All possible n-gram patterns are discovered from the data-point values. The arrhythmia scope of the provided electrocardiograms will be traced using these n-grams. Classification performs a hierarchical search on each perch hierarchy built from the optimal data-points of the labels associated with the respective perch. Following these steps, we can determine the input record's suitability for both positive and negative labels:

A search for all perches in the arrhythmia scope-positive record hierarchy that have n-grams of the same length should be conducted for each n-gram. Gather up the frequency of each individual perch that has the input string n-gram and add it to fl_+^r. Perch hierarchies are built from negative label n-grams, which yields a list of frequencies for each perch, denoted as fl_+^r in this example. The expression fl_+^r is a positive fitness index for the given electrocardiograms, and r is a positive fitness index for the given electrocardiograms. From the negative frequencies (nfs_+^r), it calculates the negative fitness score nfs. For example, let's say that a positive fitness score is greater than a negative fitness score by the given deviation threshold, and that this difference is greater than the given deviation threshold. In this case, an electrocardiogram's test results will be categorized as positive. This test record of electrocardiogram will be labelled as negative if the negative fitness score (nfs) exceeds the positive fitness score (pfs). The following description depicts the proposed model's algorithmic flow:

Hierarchy of the Perches

For all the resultant clusters, respective perch hierarchies shall be framed for both class labels. The expression Cl denotes resultant clusters of both class labels

```
$\forall_{i=1}^{|Cl|} \{c_i \exists c_i \in Cl\}$
$l = 1$ // denotes the present order of hierarchy $pHc_i$
```

```
$k = n$ // denotes the n-gram's length  that begins with maximum length $n$
$while(k \geq 1)$ Begin
    $\underset{j=1}{\overset{|ong(c_i)|}{\forall}} \{ng_j \exists ng_j \in ong(c_i)\}$ Begin
        $if\ (j=1)\ ong_k^{fr}(c_i) \leftarrow ng_j$
        $else\ if\ (|ng_j| \equiv k)$ begin
            $ong_k^{fr}(c_i) \leftarrow ng_j$ // n-grams of length k having frequency $fr$ are being stored as a set $ong_k^{fr}(c_i)$
        End
    End
    Place the n-grams of the set $ong_k^{fr}(c_i)$ as a perch in the hierarchy $phc_i$ at order indexed as $l$
    $k = k - 1$ // reduce the length index $k$ by1
    $l = l + 1\ \underset{j=1}{\overset{|ong(c_i)|}{\forall}} \{ng_j \exists ng_j \in ong(c_i)\}$ //the order index of the hierarchy is increased to index the next order
End
```

Class Label Prediction

Label estimation strategy is depicted in this section, which involves the assessment of positive fitness.

To recognize arrhythmia-proneness of the test record tr. Preprocessing (see section 3.3) was applied to test record tr, and the resultant word vector is wv.
All possible n-grams found in the data-point's vector wv has been represented by the notation $ng(tr)$. Find the most relevant perches for the n-gram data-points $ng(tr)$ of the test record tr by performing a perch exploration on all hierarchies:

Estimating positive fitness#

```
$pf = 1$// expression denotes fitness indicator having a value between 0 and 1, which is set to one
$\underset{i=1}{\overset{|Cl_+|}{\forall}} \{c_i \exists c_i \in Cl_+\}$ Begin // for each respective cluster of the positive class label
    $\underset{l=1}{\overset{|phc_i|}{\forall}} \{l \exists l = 1,2,3,....|\ phc_i\ |\}$ Begin//  in the given hierarchy, in the sequence $l$ of the orders
        $\underset{m=1}{\overset{||phc_i^l|}{\forall}} \{p_m \exists p_m \in|\ phc_i^l\}$ Begin // in the sequence of perches at order $l$
            $\underset{p=1}{\overset{|ng(tr)|}{\forall}} \{(pf = pf \times fr(p_m)) \exists ng_p \in p_m\}$ //estimating the fitness that denoted by expression $pf$
            End
        End
End
$pF(tr) = 1 - pf$ //finding the max fitness

#Estimating negative fitness#

$nf = 1$// expression denotes fitness indicator having a value between 0 and 1, which is set to one
$\underset{i=1}{\overset{|Cl_-|}{\forall}} \{c_i \exists c_i \in Cl_-\}$ Begin // sequence of resultant clusters of the negative class label
    $\underset{l=1}{\overset{|phc_i|}{\forall}} \{l \exists l = 1,2,3,....| phc_i |\}$ // in the given hierarchy, in the sequence $l$ of the orders
        $\underset{m=1}{\overset{||phc_i^l|}{\forall}} \{p_m \exists p_m \in | phc_i^l\}$ Begin // in the sequence of perches at order $l$
            $\underset{p=1}{\overset{|ng(tr)|}{\forall}} \{(nf = nf \times fr(p_m)) \exists ng_p \in p_m\}$ estimating the fitness that denoted by expression $nf$
            End
        End
End
$nF(tr) = 1 - nf$ // finding the max fitness
```

Class Label Prediction

$pF(tr), nF(tr)$, the portrayed fitness scores, will be used to determine whether a person is arrhythmia-prone in the following way:

Using the abbreviation $if\left((pF(tr) - nF(tr)) > d\tau\right)$, it is confirmed that the provided electrocardiogram record shows an arrhythmia proneness.

It is clear from the $elseif\left((nF(tr) - pF(tr)) > d\tau\right)$ abbreviation that the given electrocardiogram test results are negative for arrhythmia proneness.

Experimental Study

EHCD (ECG Heartbeat Categorization Dataset) [44] and MIT-BIH [5] have been used to create the dataset, which includes both positive and negative records. "Each record considered" means either "positive" or "negative," as stated in [45].

For the experimental study, the maximum count of labelled records perceived is 81614, which includes 46103 positive records as well as 35511 negative records. Attempting to compare the suggested model's performance to other recent models, such as ADCA Using Ensemble Learning [35] and EMLA Using Ensemble of Machine Learning Algorithms (36), has allowed us to scale it up. The proposed methodology and emerging concepts ADCA as well as EMLA have been cross-validated ten - fold in experimental studies. Cross-validation metrics have been used to evaluate the proposed APHDE method's performance (Arrhythmia

Prediction from High Dimensional Electrocardiogram). Using 10-fold cross-validation on the proposed APHDE, evaluation metrics have indeed been compared to those obtained using contemporary methods in terms of scale, accuracy rate, false alarm scope, and label predictability robustness.

The statistics of the electrocardiogram data corpus has projected in Table 1. Performance has been assessed under diversified metrics like overall prediction accuracy, sensititvity that denotes True-positive-rate (TPR), specificity that denotes True-negative-rate (TNR), F-Measure, Mathew's correlation coefficient and precision scaled under diversified optimal data-points selected using fusion of distance thresholds. The proposed supervised learning approach method "APHDE" has been critically assessed by comparing it with the performance of the ADCA [35] and EMLA [36] in a similar context of the data and the data-points. The python [46] language has considered to implement the proposed contribution.

Table 1
The statistics of chest x-ray corpus

Total positives	46103
Total negatives	35511
Positives for training for each fold	41493
Negatives for training for each fold	31960
Positives for testing for each fold	4610
Negatives for testing for each fold	3551

Accuracy

The figure it is proved that accuracy is stable for the data-points selected under distance scale of 0.5 and below, which obtained optimal data-points are 17, whereas for distance scale > 0.5 accuracy is not stable.

The term accuracy is defined as the ratio of error to the feasible output values. The graph is drawn between the Accuracy and diversity threshold for the proposed "APHDE" and contemporary methods ADCA and EMLA from the statistics, as shown in Figure 1, it is noticed that the "APHDE" performs better when compared with ADCA and EMLA.

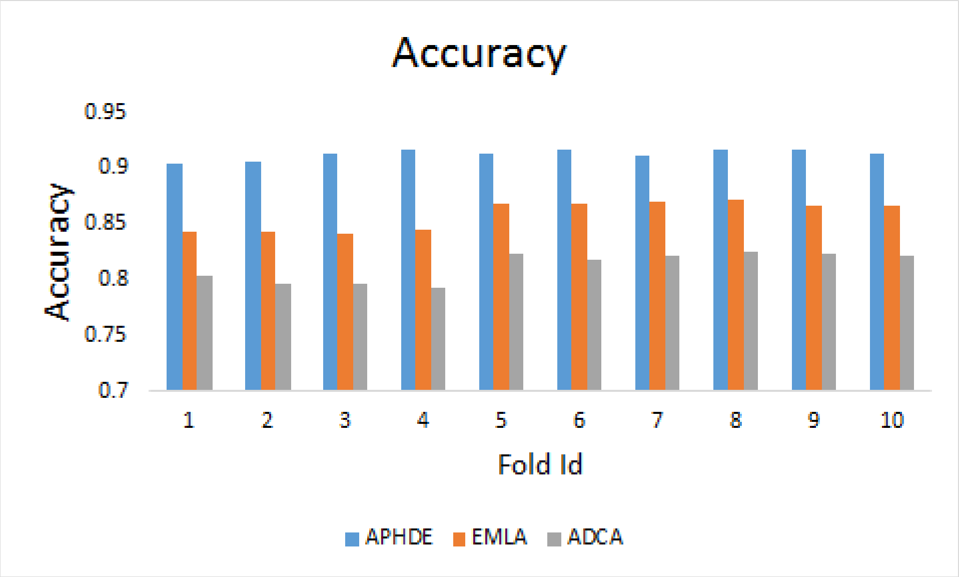

Figure 1. The Accuracy noticed for a diverse number of data-points beneath varying diversity thresholds

Precision

The metric precision is defined as the volume of information that is conveyed through value. The graph is plotted between precision and at various thresholds of diversity. From the statistics, as shown in Figure 2, it is observed that the proposed method "APHDE" is having superior performance when compared with ADCA and EMLA.

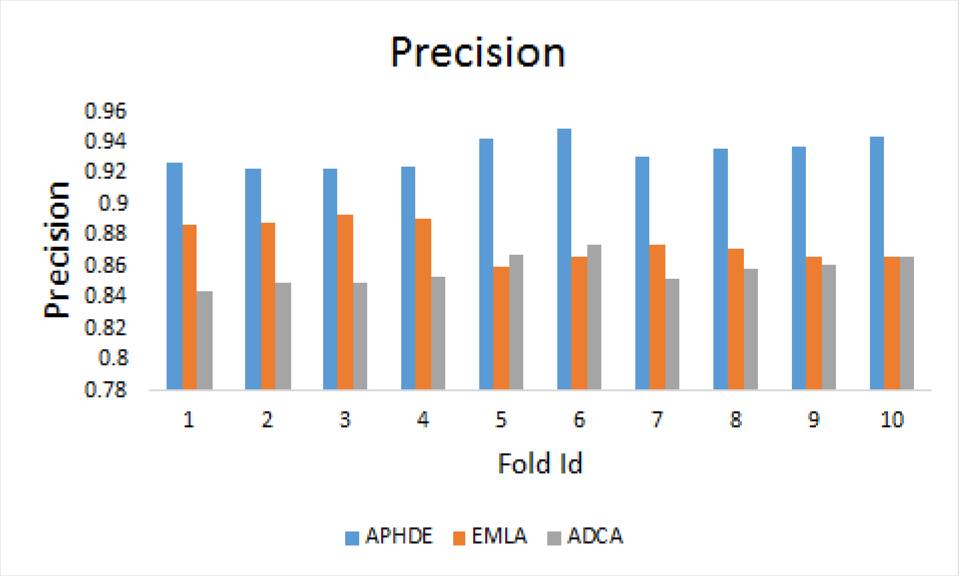

Figure 2. Precision noticed for a diverse number of data-points beneath varying diversity thresholds

Specificity

The metric specificity is the other metric used to measure the proposed model APHDE and contemporary models ADCA and EMLA over the four-folds. The true negative-rate (TNR), also known as the ratio of TNs to the aggregate of FPs and TNs, is a statistic of specificity. Figure 3 depicts a plot of metric specificity and four folds to compare the proposed model APHDE to extant models EMLA and ADCA. When it comes to specificity, the suggested model APHDE outperforms the current models EMLA and ADCA.

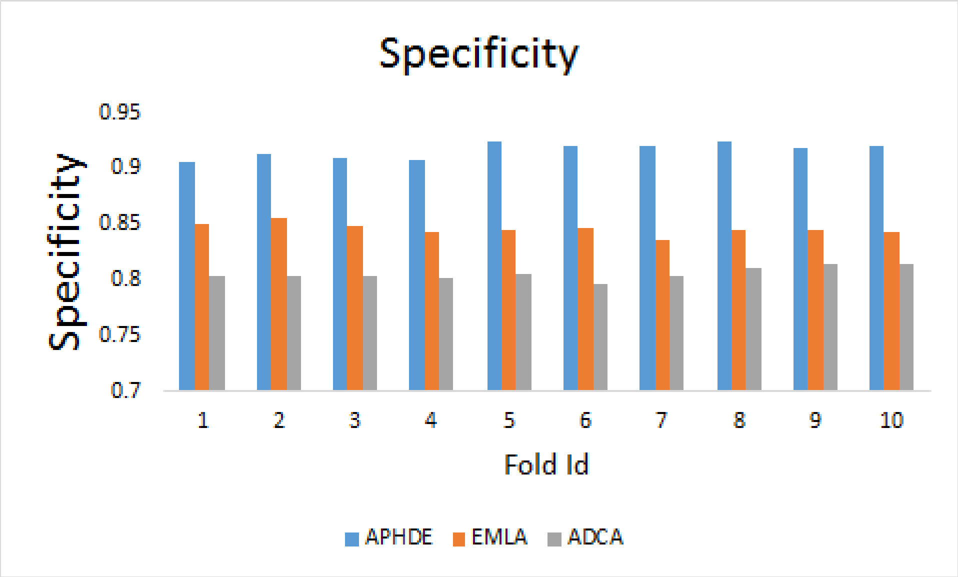

Figure 3. Value of Specificity perceived for a distinct number of data-points beneath varying diversity thresholds

Sensitivity

The metric recall or sensitivity is defined as the ratio of true positives to the sum of true positives and false negatives. The graph is plotted between sensitivity and divergent labels like positive, negative, neutral, and mismatch over the proposed model APHDE and contemporary methods EMLA and ADCA, as exhibited in Figure 4. It is envisioned from the statistics that the sensitivity of APHDE is higher for all the labels compared to the values attained for the contemporary EMLA and ADCA method

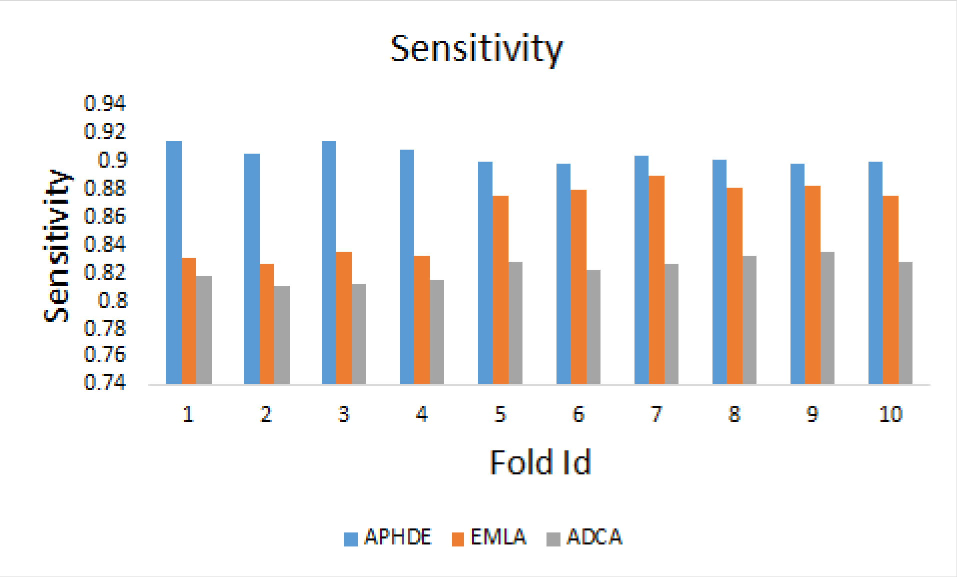

Figure 4. Recall/sensitivity (label order accuracy) statistics of both APHDE, EMLA, and ADCA observed for diversified labels

F-Measure

Figure 5 refers to the conditions of how the proposed model APHDE and the other comparative models EMLA and ADCA used indicate the performance in terms of f-measure. It is evident from the computations that the model discussed in this study has more potential in comparison to the other models.

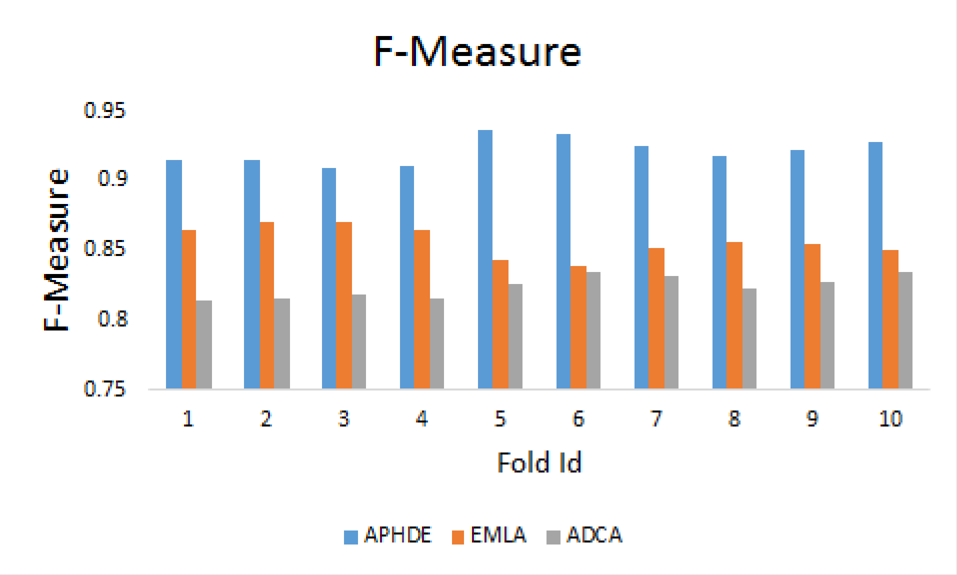

Figure 5. F-Measure noted for the compared models chosen for comparison

Matthews Correlation Coefficient

In Figure 6 terms of assessing the binary classification conditions, the metric MCC holds critical importance, and thus for the proposed model and the comparative models, the efficacy in terms of binary classification is observed. Based on the information furnished in the visual representation of the computation, performance of the proposed model is APHDE more significant than the other models EMLA and ADCA used for comparison.

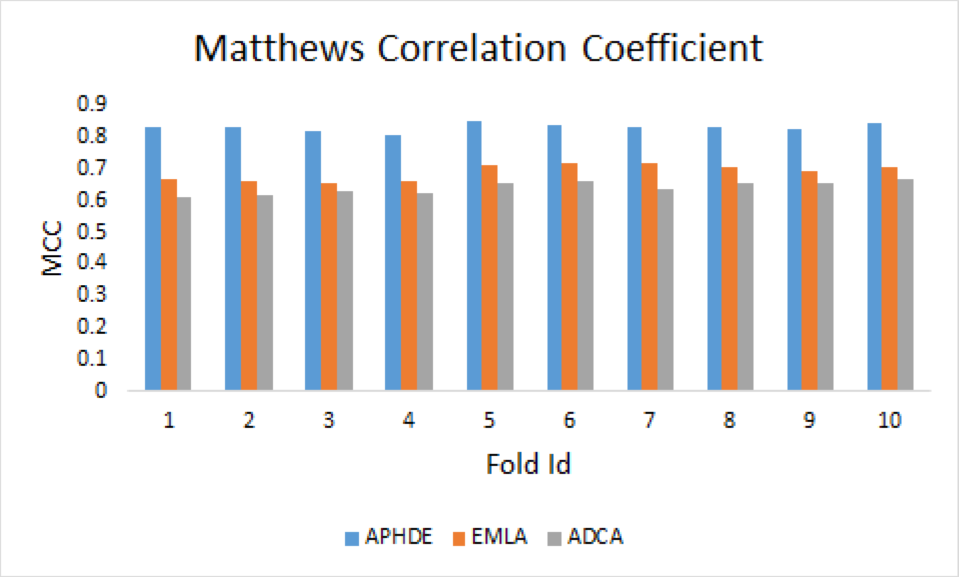

Figure 6. MCC (Mathews Correlation Coefficient) observed for the proposed APHDE and models chosen for comparison EMLA and ADCA

Conclusion

A unique heuristic search-based machine learning and classification technique was created over Cuckoo Search in order to predict if a particular ECG signal is normal, prone to arrhythmia, or prone to atrial fibrillation in the future. There were four distinct parts to the methodology: data-point selection, data-point optimization, classifier training using perch hierarchies, and disease prediction based on ECG signals. There are two datasets used to categorise the records: EHCD (ECG Heartbeat Categorization Dataset) [44] and MIT-BIH [5]. There are just two possible outcomes in the experimental study: good or negative. The best data points were found using the KS-Test diversity assessment measure on the intervals and axis data points of the provided ECG signals. The suggested "Arrythmia Prediction from High Dimensional Electrocardiogram's (APHDE) Data Corpus using Ensemble Classification" model was tested using a multi-label and multifold cross validation on the adopted benchmark dataset. The Likert Scale's influence has led to the development of a mechanism for weighing the ECG's interval and axis peaks. It has been shown that both labels have risen in prediction accuracy as a result of the proposed strategy. cross validation statistics were compared to current machine learning-based arrhythmia prediction methodologies in order to scale up the APHDE's performance. An analysis based on cross-validation metrics reveals that sensitivity, specificity, and accuracy are all significantly higher for the proposed APHDE than for the competition's present models. It is expected that future research will focus on overcoming the data's many shortcomings. Electrocardiogram signals can be used to anticipate arrhythmia and atrial fibrillation scope.

References

[1] Dagenais, G. R., Leong, D. P., Rangarajan, S., Lanas, F., Lopez-Jaramillo, P., Gupta, R., ... & Yusuf, S. (2020). Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study. The Lancet, 395(10226), 785-794.

[2] Rajni, R., & Kaur, I. (2013). Electrocardiogram signal analysis-an overview. International Journal of Computer Applications, 84(7), 22-25.

[3] Clifford, G. D., Azuaje, F., & McSharry, P. (2006). Advanced methods and tools for ECG data analysis (p. 12). Boston: Artech house.

[4] Malmivuo, J., & Plonsey, R. (1995). Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields. Oxford University Press, USA.

[5] Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. IEEE Engineering in Medicine and Biology Magazine, 20(3), 45-50.

[6] Misiti, M. I. C. H. E. L., Oppenheim, G., & Poggi, J. M. (1996). Wavelet toolbox for use with Matlab, the Math Works: Natick.

[7] Thiamchoo, N., & Phukpattaranont, P. (2016, June). Application of wavelet transform and Shannon energy on R peak detection algorithm. In 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (pp. 1-5). IEEE.

[8] Adeluyi, O., & Lee, J. A. (2011). R-READER: A lightweight algorithm for rapid detection of ECG signal R-peaks. In 2011 2nd International Conference on Engineering and Industries (ICEI) (pp. 1-5). IEEE.

[9] Mabrouki, R., Khaddoumi, B., & Sayadi, M. (2014, March). R peak detection in electrocardiogram signal based on a combination between empirical mode decomposition and Hilbert transform. In 2014 1st International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 183-187). IEEE.

[10] Elgendi, M., Jonkman, M., & De Boer, F. (2009, April). R wave detection using Coiflets wavelets. In 2009 IEEE 35th Annual Northeast Bioengineering Conference (pp. 1-2). IEEE.

[11] Elgendi, M. (2013). Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases. PloS one, 8(9), e73557.

[12] Elgendi, M. (2016). TERMA framework for biomedical signal analysis: An economic-inspired approach. Biosensors, 6(4), 55, pp- 1-19.

[13] Elgendi, M., Meo, M., & Abbott, D. (2016). A proof-of-concept study: Simple and effective detection of P and T waves in arrhythmic ECG signals. Bioengineering, 3(4), 26-40.

[14] Sarkaleh, M. K., & Shahbahrami, A. (2012). Classification of ECG arrhythmias using discrete wavelet transform and neural networks. International Journal of Computer Science, Engineering and Applications, 2(1), 1.

[15] Jaiswal, G. K., & Paul, R. (2014). Artificial neural network for ecg classification. Recent Research in Science and Technology, 6(1).

[16] Sadr, A., Mohsenifar, N., & Okhovat, R. S. (2011). Comparison of MLP and RBF neural networks for Prediction of ECG Signals. International Journal of Computer Science and Network Security (IJCSNS), 11(11), 124-128.

[17] Tang, X., & Shu, L. (2014). Classification of electrocardiogram signals with RS and quantum neural networks. International Journal of Multimedia and Ubiquitous Engineering, 9(2), 363-372.

[18] Belgacem, N., Chikh, M. A., & Reguig, F. B. (2003). Supervised classification of ECG using neural networks. In Biomedical Engineering Laboratory, Department of Electronics, Science Engineering Faculty. Belkaid University.

[19] Kshirsagar, P. R., Akojwar, S. G., & Dhanoriya, R. A. M. K. U. M. A. R. (2017). Classification of ECG-signals using artificial neural networks. Researchgate. Net, pp. 1-4.

[20] Sao, P., Hegadi, R., & Karmakar, S. (2015, April). ECG signal analysis using artificial neural network. In International Journal of Science and Research, National Conference on Knowledge, Innovation in Technology and Engineering (pp. 82-86).

[21] Mitra, M., & Samanta, R. K. (2013). Cardiac arrhythmia classification using neural networks with selected data-points. Procedia Technology, 10, 76-84.

[22] Gayathri, B. M., & Sumathi, C. P. (2016, December). Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-5). IEEE.

[23] M. Rudra Kumar, V. K. (2020). Review of Machine Learning models for Credit Scoring Analysis. Revista Ingeniería Solidaria, 16(1).

[24] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., & San Tan, R. (2017). A deep convolutional neural network model to classify heartbeats. Computers in biology and medicine, 89, 389-396.

[25] Yıldırım, Ö., Pławiak, P., Tan, R. S., & Acharya, U. R. (2018). Arrhythmia detection using deep convolutional neural network with long duration ECG signals. Computers in biology and medicine, 102, 411-420.

[26] Pławiak, P. (2018). Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. Expert Systems with Applications, 92, 334-349.

[27] Pławiak, P. (2018). Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals. Swarm and evolutionary computation, 39, 192-208.

[28] Cuesta-Frau, D. e. (2007). Unsupervised classification of ventricular extrasystoles using bounded clustering algorithms and morphology matching. Medical & biological engineering & computing, 45(3), 229-239.

[29] Tuncer, T. e. (2019). Automated arrhythmia detection using novel hexadecimal local pattern and multiorder wavelet transform with ECG signals. Knowledge-Based Systems, 186, 104923.

[30] Rajesh, K. N. (2017). Classification of ECG heart-beats using nonlinear decomposition methods and support vector machine. Computers in biology and medicine, 87, 271-284.

[31] Tadeusiewicz, R. (2015). Neural networks as a tool for modeling of biological systems. Bio-Algorithms and Med-Systems, 11(3), 135-144.

[32] Pławiak, P. a. (2019). Novel deep genetic ensemble of classifiers for arrhythmia detection using ECG signals. Neural Computing and Applications, 1-25.

[33] Elhaj, F. A. (2016). Arrhythmia recognition and classification using combined linear and nonlinear data-points of ECG signals. Computer methods and programs in biomedicine, 127, 52-63.

[34] De Chazal, P. M. (2004). Automatic classification of heart-beats using ECG morphology and heart-beat interval data-points. IEEE transactions on biomedical engineering, 51(7), 1196-1206.

[35] Peimankar, A., Jajroodi, M. J., & Puthusserypady, S. (2019, October). Automatic detection of cardiac arrhythmias using ensemble learning. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 383-388). IEEE.

[36] Abirami, R. N., & Vincent, P. D. R. (2020). Cardiac Arrhythmia Detection Using Ensemble of Machine Learning Algorithms. In Soft Computing for Problem Solving (pp. 475-487). Springer, Singapore.

[37] Abdualrhman, Mohammed Ahmed Ali, and M. C. Padma (2019). "CS-IBC: Cuckoo search based incremental binary classifier for data streams." Journal of King Saud University-Computer and Information Sciences 31.3 (2019): 367-377.

[38] Mu, E., & Pereyra-Rojas, M. (2016). Practical decision making: an introduction to the Analytic Hierarchy Process (AHP) using super decisions V2. Springer.

[39] Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10(2), 486.

[40] McKnight PE, N. J. (2010, Jan 30). Mann-Whitney U Test. The Corsiniencyclopedia of psychology., 1-1.

[41] Sahoo, P., &amp; Riedel, T. (1998). Mean value theorems and functional equations. World Scientific.

[42] Z-table. (n.d.). Retrieved from http://www.sjsu.edu/faculty/gerstman/StatPrimer/z-table.PDF

[43] https://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/].

[44] https://www.kaggle.com/shayanfazeli/heartbeat.

[45] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/.

[46] Python. (n.d.). Retrieved from https://www.python.org/.