DOI: 10.17725/j.rensit.2024.16.053

A Robust and Efficient Scheme for ECG Signal Classification Based on Digital Signal Processing, Feature Engineering Selection, and Random Forest Classifier

Anas Fouad Ahmed, Baraa M. Albaker

Al-Iraqia University, College of Engineering, Electrical Engineering Department, http://www.aliraqia.edu.iq/
Al Adhmia-Haiba Khaton, 6029, Baghdad, Iraq
E-mail: anas.ahmed@aliraqia.edu.iq, baraamalbaker@aliraqia.edu.iq

Hadeel N. Abdullah

University of Technology, Electrical Engineering Department, http://www.uotechnology.edu.iq/
Al wehada-Neighborhood, 19006, Baghdad, Iraq
E-mail: 30002@uotechnology.edu.iq

Received September 11, 2023, peer-reviewed September 18, 2023, accepted September 25, 2023, published March 15, 2024.

Abstract: **Determining the optimal integration between features and classifiers has a significant effect on the performance of automatic heartbeat diagnostic systems. This importance stands out when dealing with critical applications that contain limited resources devices and require accurate and fast heartbeat classifiers to help the doctor make an accurate and quick diagnosis of heart diseases. Aiming at this task, this paper introduces a novel approach for choosing the optimal features of the ECG signal to be used with the Random Forest (RF) classifier following the inter-patient method for ECG signals division and obeying the instructions of the "Association for the Advancement of Medical Instrumentation (AAMI)." The features were chosen based on the concept of "Mutual Information Ranking (MIR)." The presented framework is comprehensive in terms of performing all the necessary processes efficiently, starting from ECG digital signal processing, segmentation, feature extraction, feature selection, and ending with ECG classification. The results of the experiments demonstrate that features corresponding to the normalized QRS width and the normalized RR intervals are the most influential features in the heartbeat classification. All tests were conducted using real ECG signals taken from the "MIT-BIH" Arrhythmia Database (MIT-BIH-ARR-DB). The suggested scheme attained the following F1-scores: 91.02%, 73.17%, and 98.04% in the classification of the Ventricular Ectopic Beats (V or VEB), Supraventricular Ectopic Beats (S or SVEB), and Normal Beats (N or NB), respectively. The overall accuracy was 96.26%. Despite its relative simplicity and reliance on few features, the proposed approach outperforms most of the reported state-of-the-art.**

Keywords: **Mutual Information; Feature Selection; Cardiac Arrhythmia; Discrete Wavelet Transform (DWT); ECG classification; Random Forest (RF)**

UDC 53.047:57(075.8)

For citation: Anas Fouad Ahmed, Hadeel N. Abdullah, Baraa M. Albaker. A Robust and Efficient Scheme for ECG Signal Classification Based on Digital Signal Processing, Feature Engineering Selection, and Random Forest Classifier. *RENSIT: Radioelectronics. Nanosystems. Information Technologies*, 2024, 16(1):53-66e. DOI: 10.17725/j.rensit.2024.16.053.

Contents

1. INTRODUCTION

Cardiovascular ailments are the main reason for killing millions of people around the world, and the electrocardiogram (ECG) signal is a magnificent and effective means in their diagnosis [1,2]. The ECG is a low-cost and non-invasive means frequently used by cardiologists to record electrical activity and to detect abnormalities in the heart [3-5]. The ECG instrument senses this electrical activity using electrodes placed on the body and shows it on the screen [6-9]. An ECG signal is composed of a collection of periodical heartbeats, which are examined by a doctor to make a diagnosis and treatment [10]. This procedure relatively requires a lot of time and intensive efforts, in addition to human errors that may occur due to workload [11]. Nowadays, it has become easier to acquire the ECG signal due to the availability of portable and wearable devices and the possibility of using them in ambulatory healthcare applications [12]. Employing these devices requires evolution not only in accuracy and reliability but also simple and computationally efficient methods for intelligent automatic heartbeat detection and classification to limit the amount of information or data required to be sent to the doctor to evaluate the case and give counsel [13].

By referring to the AAMI, the approaches that classify the heartbeats must at most discriminate the following categories: Fusion Beats (F or FB), Unclassifiable or paced Beats (Q or QB), (VEB), (SVEB), and (NB). It is necessary to detect and count the number of period ectopic beats to avoid dangerous heart disease. For instance, the repetition of SVEB is related to dangerous illness, "Atrial Fibrillation," also the recurrence of VEB is exceptionally indicator because it can be utilized as a predictor of the most dangerous disease, "Heart Failure," which often leads to death [14,15].

Many efforts have been made by researchers in recent years to solve the problem of automatic diagnosing or classifying heartbeats by proposing various techniques to fulfill this task. Generally, these techniques can be classified into two approaches: the first one is based on a deep-learning strategy, while the other is based on a feature-engineering. Some researchers have applied Deep Neural Networks (DNNs) for classifying ECG signals [16-18], but the DNN experiences redundancy of parameters and complexity of calculations. It is significant to mention here that the analytical classification based on features engineering can assist cardiologists in accurately diagnosing and precisely drawing up suitable therapy schedules by taking advantage of the understandable extracted features.

Deep learning algorithms, on the other hand, are unable to provide cardiologists with interpretable features and cannot analyze the effects of feature selection on ECG classification efficiency because the features are tacitly captured by deep layers in the network.One of the most important surveys of heartbeat classification schemes was introduced by Luz et al. [19]. Machine learning (ML) techniques have enticed great interest nowadays, and perhaps the most famous of these is the RF technique, in which a set of classifiers is produced as an "ensemble" of Decision Trees (DTs) [20]. The input is classified by the ensemble (forest) based on the majority principle [21]. Therefore, the RF classifier is considered extremely effective and has been recommended for resource-restricted devices that are intended to work in real-time [22-25]. Emanit [26] applied the DWT to the ECG signals and used the resultant coefficients as features to train the RF classifier. Llamedo et al. [27] presented a simple ECG classifier that classifies the SVEB and VEB based on coefficients of the DWT and RR intervals. Alickovic and Subasi [28] suggested utilizing the distribution of DWT coefficients to extract various features that feed into RF for heartbeats classification purposes. Gutierrez et al. [29] classified ECG signals using Quadratic Wavelet Transform (QWT) and neural networks. Ganesh and Kumaraswamy [30] proposed a method for ECG signals classification based on RR intervals and features extracted by taking the Discrete Cosine Transform (DCT) for ECG signals to train the RF classifier. Park et al. [31] used collections of different time-domain features with an RF classifier for heartbeat classification. Aravind et al. [32] offered an algorithm for heartbeat classification using the Convolutional Neural Network (CNN) and Continuous Wavelet

Transform (CWT). Yuanlu et al. [33] classified the ECG signal based on deep residual CNN.

The aforementioned studies achieved excellent accuracy (ACCo); however, some limitations exist, such as few of them: obeyed the AAMI guidelines regarding the type of heartbeat to be detected and classified, offered a robust model that can be applied to imbalanced data or signals directly, and employed the other metrics to evaluate the classification quality, such as F1-Score, besides the Positive Predictivity (PPr), and Sensitivity (Sen), that may be degraded because of the majority class domination. Most importantly of all, there are no previous studies (as far as the authors know) that deal with the use of the RF classifier obeying the inter-patient mode. In this mode, the signals are separated into two groups: the first is used in the training phase and the second is used in the test phase, and there is no intersection between the two groups (there are no signals taken from the same patient in both groups) this leads to a more reliable assessment of the performance of the classifier. In this research, we used an efficient RF classifier for classifying the ECG signals according to the AAMI recommendations and the inter-patient mode. The following are the contributions of this work:

- We introduce a comprehensive, robust, and efficient framework for ECG signals classification based on an improved RF classifier, taking into consideration all the required digital signal processing and analysis for the collected signals and adhering to the instructions of the AAMI and the inter-patient mode.
- A simple and efficient mechanism for detecting R-peaks based on trigonometry is proposed.
- New considerations for evaluating the normalized features are presented.
- Optimizing the performance of the RF classifier without requiring complex computations by using the MIR to obtain an optimal and reduced group of features from a large feature set that contains the most important features suggested in previous studies.
- Successful dealing with the problem of imbalanced data of ECG signals where the abnormal signals are much less than normal signals; this attained by using the RF classifier, which is actually composed of an ensemble of several DTs that are sensitive to imbalanced classes and it takes a Uniform Random Sample (URS) from the data of ECG signal with replacement strategy.
- An understandable analysis for ECG signal is attained by providing feature sets from various ECG signals, which are indispensable for medical purposes.

2. PROPOSED FRAMEWORK

The complete block diagram of the suggested framework is illustrated in **Fig. 1**. It consists of the following stages: collecting ECG signals, digital signal preprocessing, preparing heartbeats, features extraction, features selection, heartbeats classification, and classifier evaluation. Each one of these stages is further detailed in the subsequent sections.

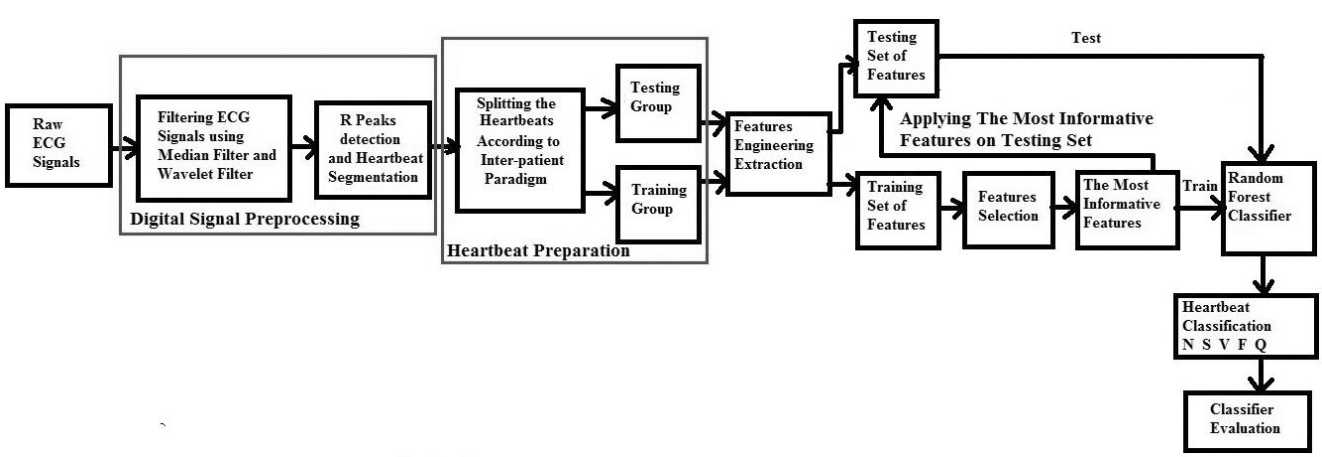

Fig. 1. *Block diagram of suggested framework.*

2.1. Collecting ECG signals

The ECG signals were collected from MIT-BIH-ARR-DB, which is available in [34]. It is well-known and used by many researchers, and this facilitates the process of comparing the suggested scheme with previous works. **Table 1** describes these signals briefly.

Table 1

Description of the collected ECG signals

Parameter	Value
No. Records	48
Duration (Minutes)	30
Sampling Frequency (Hz)	360
Leads	Single (MLII)
No. Heartbeats	109,494
No. Heartbeats types	15
Gender of the Participants (%)	53% males, 47% females
Age Range (Years)	23-89

The relabeling process for the collected signals was conducted to obtain five heartbeats classes according to the AAMI instructions, as shown in **Table 2.**

Table 2

Relabeling the MIT-BIH-ARR-DB according to the AAMI instructions

AAMI Classes with Explaination	MIT-BIH Beats with Labels	
Normal beats (**N**)	N	Normal beat
	L	Left bundle branch block beat
	R	Right bundle branch block beat
	e	Atrial escape beat
	j	Nodal (junctional) escape beat
Supraventricular ectopic beats (**S**VEB)	A	Atrial premature beat
	a	Aberrated atrial premature beat
	J	Nodal (junctional) premature beat
	S	Supraventricular premature beat
Ventricular ectopic beats (**V**EB)	V	Premature Ventricular contraction
	E	Ventricular escape beat
Fusion beats (**F**)	F	Fusion of ventricular and normal beat
Unknown beats (**Q**)	/	Paced beat
	f	Fusion of paced and normal beat
	Q	Unclassified beat

2.2. Digital Signal Preprocessing

In practice, the ECG signals during their acquiring or transmitting are usually corrupted with two main types of noises: Baseline Wanders (BW) and White Gaussian Noise (WGN) [35]. These noises must be reduced as possible to avoid their negative impact on the quality of the signals and the classification performance. **Fig. 2** demonstrates an example of an ECG signal contaminated with BW and WGN.

To obtain baseline-corrected ECG signals, each signal passed through two Median Filters (the width of the first is 200 ms, while the width of the second is 600 ms). Next, it is subtracted from the raw ECG signal. After that, to eliminate the WGN, the DWT is

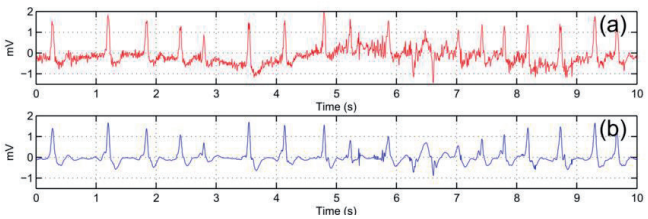

Fig. 2. *An example of ECG signal corrupted with BW and WGN.*

taken for the baseline-corrected signal to decompose it into various frequency bands. The Daubechies-6 (DB6) wavelet filter, four levels of decomposition, soft thresholding scheme (defined in Eq. (1)), and universal threshold determination strategy (given in Eq. (2)) are optimal for removing the WGN from the ECG signal.

$$S_{thresholding}(De) = \text{sign}(De)\left(|De| - t_{hr}\right) \quad \text{for } |De| > t_{hr}, \text{ and } 0 \text{ for } |De| \le t_{hr}, \tag{1}$$

where De are the detail coefficients; t_{hr} is the threshold value.

$$t_{hr} = \text{SDN}.\sqrt{2\log(\text{No.Samples})}, \quad \text{SDN} = \text{Median}(D)/0.6745 \tag{2}$$

where SDN – is the Standard Deviation of Noise.

The DB6 is characterized by the following features [35]:

- It has a very close shape to the ECG signal, which is essential for the ideal reconstruction.
- It is suitable for non-stationary signals due to its short vanishing moments.
- It has relatively low-cost computations due to its orthogonality.

Fig. 3 illustrates the shape of the DB6 wavelet filter.

Fig. 3. *The shape of the DB6 wavelet filter.*

2.3. R Peak Detection

This research suggests a simple and efficient strategy for R peak detection by evaluating the angles in degrees between sequential samples of the amplitude of the signal followed by applying a dynamic threshold to these angles and a time window as described in equations (3), (4), (5), (6), and (7), respectively:

$$\theta(k) = \tan^{-1}\left(\frac{V}{W}\right) = \tan^{-1}\left[\frac{\beta \times \left|\delta(k) - \delta(k-1)\right|}{\dfrac{360}{\text{Sampling frequency}}}\right] \quad (3)$$

$$\beta = \begin{cases} 500, & \textit{initial value} \\ 1000, & \left(\frac{V}{W}\right) < 57.5 \ \textit{or} \ \theta(k) < 89^{\circ} \ \textit{during} \ 2s \ \text{interval} \\ 500, & \left(\frac{V}{W}\right) < 117 \ \textit{or} \ \theta(k) > 89.5^{\circ}, \end{cases} \quad (4)$$

where β is the magnification factor, δ(*k*) is the k^{th} sample of the amplitude.

Note that β in Eq. (4); is adjusted in a way that ensures detection of low-amplitude beats. **Fig. 4** illustrates this concept; when sampling frequency = 360 Hz, it is clear that the range of angles is maintained between 80° and 90°. The dynamic threshold value in degrees *D*(*k*) is varied according to Eq. (5) to tackle the variations of the heartbeats due to different patient cases and recording circumstances. The values of *θ*(*k*) and *D*(*k*) are compared, then the time window *G*(*k*) is used for *R* peak detection.

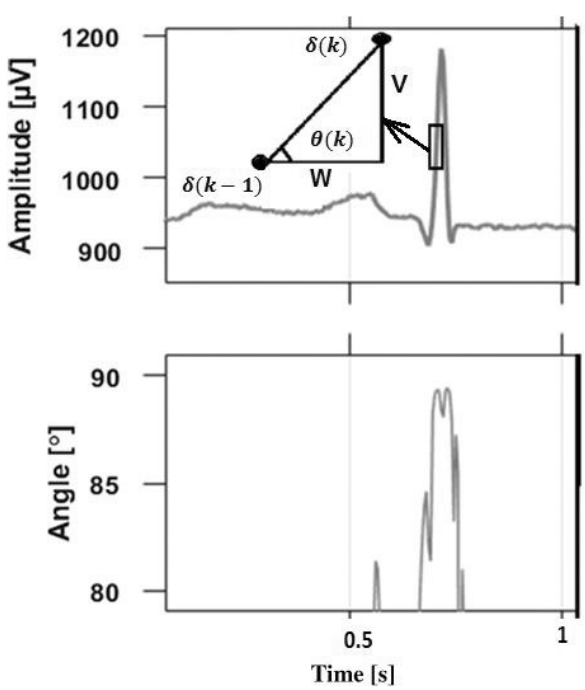

Fig. 4. *Evaluating the angles of the heartbeat based on trigonometry.*

$$D(k+1) = \begin{cases} 0, & \textit{initial value} \\ \theta(k) - Y_1, \| I \| \theta(k) > D(k) + Y_1 \\ D(k), \| I \| D(k) > \theta(k) \le D(k) + Y_1, \beta t = 0, \\ D(k) - (Y_2 \times \beta t), \| \theta(k) \le D(k), \beta t = \beta t + 1 \\ 80, & | D(k) \le 80, \end{cases} \quad (5)$$

where Y_1 and Y_2 are empirically set to 0.4° and 0.00012°, sequentially to get the smallest detection error as illustrated in **Appendix A.**

$$G(k) = \begin{cases} \textit{Signal amplitude}, & 0 \le \beta t \le Y_2 \\ 0, & \textit{otherwise} \end{cases} \quad (6)$$

$$Y_2 = \begin{cases} 0.25s, & \textit{initial value} \\ 0.450s \times F_s, & \textit{Mean of last 8 RR intervals} \ge 0.68s \\ 0.25s \times F_s, & \textit{otherwise.} \end{cases} \quad (7)$$

After the time interval of *G*(*k*) is evaluated, a comparison between the |local maximum| and |local minimum| amplitude of the signal within *G*(*k*) is conducted; the highest one is specified as the *R* peak point, as illustrated in **Fig. 5.** The strengths of the proposed strategy can be summarized as follows: Eq. (1) guarantees that the relationship between *θ*(*k*) and |δ(*k*) – δ(*k* – 1)| is nonlinear, and *θ*(*k*) is restricted for a fixed range (0° and 90°). In other words, *θ*(*k*) is immune to the different changes of the QRS complex. Moreover, the $\tan^{-1}$ term makes the computation of *θ*(*k*) easy and fast.

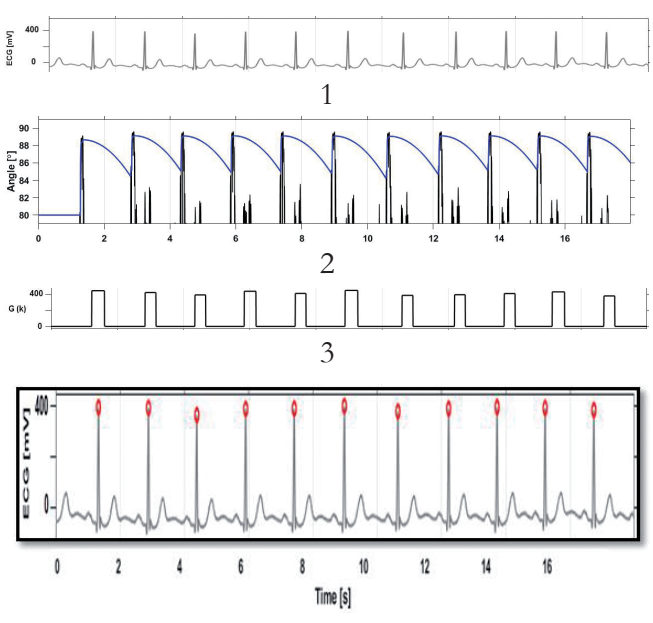

Fig. 5. *A demonstrative example of the suggested approach for detecting the R peaks. (1). Denoised signal, (2). Angles (black) and dynamic thresholding (blue), (3). Time window G (k), (4). Locating and tagging the R peaks.*

After performing the filtering process, and *R* peak detection, the segmentation process is carried out. Each ECG signal is segmented into several heartbeats based on *R* peaks locations. The heartbeat is defined by specifying 92 sampling points before the *R* peak and 146 sampling points after the *R* peak (i.e., the length of each heartbeat is 238 samples).

2.4. Preparing Heartbeats

The segmented heartbeats are divided (according to the inter-patient scheme [36]) into two groups: the training group, which results from filtering and segmenting the signals of set 1, and the testing group, which results from filtering; and segmenting the signals of set 2. There is no intersection between the two sets (they come from different patients). The recordings for set 1 and set 2 are shown in **Table 3.** According to AAMI standards [37], the four recordings that contain pace beats are eliminated.

Table 3

Signals of set 1 and set 2 with their associated recordings.

Signals	Recordings from MIT-BIH-ARR-DB
signals of set1	101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230
signals of set2	100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234

2.5. Features Extraction

A large number of features were examined in this paper, and they represent the most notable eighty-five features, most of which were addressed in previous works [31,37-41], and they are as follows:

- **The Coefficients of DWT:** The DWT has the ability to extract detailed information from both the frequency and temporal domains, making it ideal for ECG representation. In this study, the first-order Daubechies mother wavelet (DB1) with three decomposition levels was used to introduce twenty-three features.
- **The Coefficients of Hermite Basis Functions (HBF):** The coefficients of HBF were utilized as the features describing the pattern of the ECG signal. The sample points situated 250 ms before and following each R peak represent the beat segment. The parameters for evaluating the coefficients of HBF were described in [38]. The order of HBF was fixed to twenty, and the width variable was estimated to reduce the reconstruction error (as possible) for each beat [37]. Three, four, and five HBFs were used to get fifteen features.
- **Higher-Order Statistics (HOS)**: They are functional in capturing the slight variations in ECG signals [39]. Each heartbeat was segmented into 5 intervals over each one; the third and fourth-order statistics (skewness and kurtosis) were determined to generate ten features. The range of delay variables, which are centered on the R peak, is (–250 to 250) ms as described in [39].
- **QRS Complex Durations and Their Normalizations:** They include the following five durations and their normalizations (ten features): overall QRS duration (QRSo), the QRS duration at the middle value of the R peak (QRSm), the QRS duration at the quarter value of the R peak (QRSq), the duration between Q and S waves (dQS), the duration between P wave and the starting point of the QRS complex (dPQRS). These features are demonstrated in **Fig. 6**. The normalization for the above features can be obtained by dividing the feature value by the mean of its value in the last thirty-two heartbeats. A comprehensive explanation for extraction mechanism is demonstrated in **Appendix B.**
- **Euclidean Distances:** These four features depend on Euclidean distances between the R peak and four points of the beat that represent the amplitude values through several samples (as proposed in [37]): maximum amplitude (Beat [0, 40]), minimum amplitude (Beat [75, 85]), minimum amplitude (Beat [95, 105]), and maximum amplitude (Beat [150, 180]).

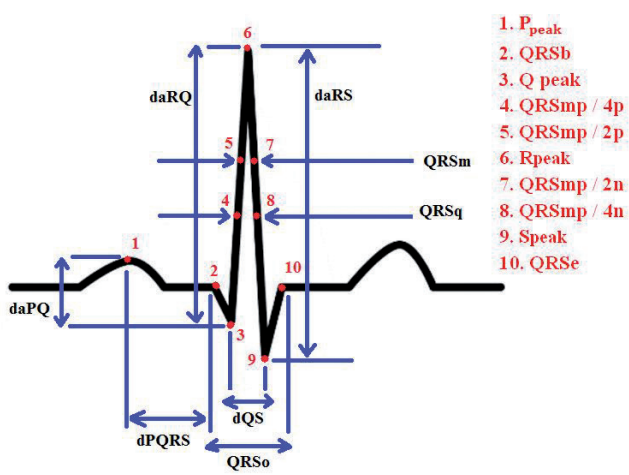

Fig. 6. *Demonstration of features of the QRS complex durations and amplitude differences.*

- **Amplitude Differences and Their Normalizations**: They consist of the following seven amplitudes and their normalizations (fourteen features): the difference in amplitude between R and Q waves (daRQ), the difference in amplitude between P and Q waves (daPQ), the differences in amplitude between R and S waves (daRS), and the peak values of (S, R, P, and Q) waves. All these features are illustrated in Fig. 6. The normalization for the above features can be determined by dividing the feature value by the mean of its value in the latter thirty-two heartbeats. A comprehensive explanation for the extraction mechanism is demonstrated in **Appendix B.**
- **RR Intervals and Their Normalizations**: They are the most utilized features for classification purposes. These features can be calculated from the time distance between consequent beats. If we denote RRp for the present RR interval, then RRp – 1 and RRp + 1 represent the pre RR and post RR intervals, respectively. The number of RR intervals considered in this paper is three. The following six features are the normalized RR intervals:

  Normalized RRp = (RRp)/mean of the latter thirty-two RR intervals

  Normalized RRp – 1 = (RRp – 1)/mean of the latter thirty-two RR intervals

  Normalized RRp + 1 = (RRp + 1)/mean of the latter thirty-two RR intervals

  Normalized RRr1 = (RRp – 1)/RRp

  Normalized RRr2 = (RRp + 1)/RRp

  Normalized RRpt = [RRp – mean of the latter thirty-two RR intervals]/ standard deviation of the latter thirty-two RR intervals.

2.6. Feature Selection

To minimize the computational cost of the introduced framework without sacrificing the performance of the classifier, the number of features for the training stage was adjusted to ten; this is consistent with the conclusions of the specialists and previous work [27,41]. The MIR principle is used to select just the most significant features according to the class of the heartbeats. From the perspective of selection features, MIR has been recognized to be an impressive strategy as it can identify non-linear relevance between features of a given features-vector or features-matrix. The scores of MIR (or MI values) evaluated between the labels of the classes and features indicate the discrimination ability of these features.

$$MI(\text{feature}, \text{class label}) = E(\text{feature}) - E(\text{class label} \mid \text{feature}) \quad (8)$$

where E (feature): is the entropy of a feature; E (class label | feature): is the conditional entropy of the class label given feature evaluating the uncertainty about the class label whenever the feature is known. In this research, the MI values are estimated based on work in [42] and [43] and using a python function called "mutual_info_classif.".

2.7. The Classification of Heartbeats

In this study, the RF classifier is utilized for the heartbeats classification task. It is an ensemble of Z trees R1(I), R2(I), …, RZ(I), where I = i_1, i_2, …, i_j is a *j*-dimension vector of inputs and the obtained group generates Z outputs O1= R1(I), O2= R2(I), …, Oz = Rz(I). Oz is the value of the prediction taken by tree number z. The final prediction O is determined by aggregating the output of all random trees (majority voting). The RF produces Z number of DTs from L-trained samples. The new training group is produced by conducting bootstrap sampling for all trees in the forest individually. This group is utilized to develop DT without trimming. In each division of a DT node, just little numbers of *j* features are chosen randomly rather than all of them. In order to construct a randomly produced forest, this procedure iterated to generate *J* of the DTs.

The following is a brief description of the training process for a randomly created forest:

Stage One: From the training group, choose an arbitrary sample.

Stage Two: For every arbitrary sample, create a tree with the following alteration: at every node, choose the optimal division among an arbitrary chosen subgroup of input parameters, which is the RF's tuning parameter. The tree is completely created until no more divisions are potential and not trimmed.

Stage Three: Reiterate stages one and two until Z such trees are created.

The RF classifier was constructed by utilizing the Scikit-learn library of Python [44]. All the variables remained at their default configuration, except the number of trees. To discover the best number of trees, the training was performed with an increased

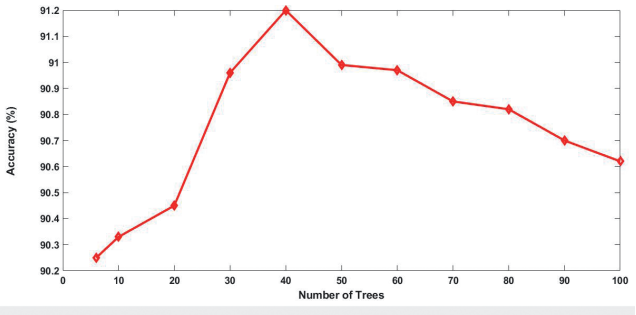

Fig. 7. *Accuracy vs. the number of trees attained from LOOCV on the training set for the RF classifiers.*

number of trees by utilizing the optimal ten features, and the ACC achievement was examined. **Fig. 7** depicts the results of applying the LOOCV on the training set. It is clear that when the number of trees exceeds forty, the accuracy of the classifiers does not improve and may even deteriorate.

2.8. Classification Assessment Metrics

The measurement for the classification performance and comparison with state of the art was conducted by determining the following metrics: Overall Accuracy (ACCo), F1-score, Positive Predictivity (PPr), and Sensitivity (Sen).

$$\mathrm{F1-Score} = \frac{2 \cdot \mathrm{PPr} \cdot \mathrm{Sen}}{\mathrm{PPr} + \mathrm{Sen}} \tag{8}$$

$$\mathrm{PPr} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\mathrm{ACCo} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{11}$$

where TP – True Positive, TN – True Negative, FP – False Positive, and FN – False Negative.

3. RESULTS AND DISCUSSION

This paper performs an investigation into the exploitation of the RF classifier for ECG signal classification by employing the optimal choice for the extracted Time-Domain (TD) features and obeying the AAMI standards and the inter-patient approach. Various eighty-five features were examined in this research. All the features suggested in [37], which are considered important research work in this direction, were implemented. Furthermore, most of the features offered or based on [41] and [38] were used. In this study, normalization was accomplished by dividing the features by the mean of the most recent heartbeats. This process can be conducted more practically in real-time medical wearable and portable monitoring devices. Further normalizations for the RRp – 1 and RRp + 1 were performed by dividing them by the RRp interval; this was presented as an effort to comprise the timing relevance between the sequential R peaks, which is also addressed in [45]. The normalized RRpt was calculated to implicate a quantitative measure of the standard deviations for a specific interval from the mean. **Table 4** illustrates the most significant ten features evaluated by the MIR principle. Because the high proportion of the above normalizations is presented in the table of the most significant features (as indicated in Table 4), our findings reveal that the normalized RR intervals provide further information and are more effective for classifying ECG signals than other classical RR intervals. Just the best ten with the highest significant features were utilized in the tests. The results exhibit that only six features are adequate to attain the optimum outcomes (ACCo = 96.26%), as illustrated in **Fig. 8**, this is consistent with the obtained results

Table 4
The most significant ten features resulting from MIR with their MI values

Position According to MIR	Feature Name	MI Value
1	Normalized QRSm	0.19037819
2	Normalized QRSq	0.17508354
3	Normalized RRp	0.16311836
4	Normalized RRr2	0.14615831
5	QRSm	0.14439339
6	First fitting coefficient of the fourth degree of HBF (HBFC1D4)	0.14157413
7	HBFC2D4	0.14001679
8	QRSq	0.13641600
9	HBFC0D4	0.13299729
10	Normalized RRr1	0.13147040

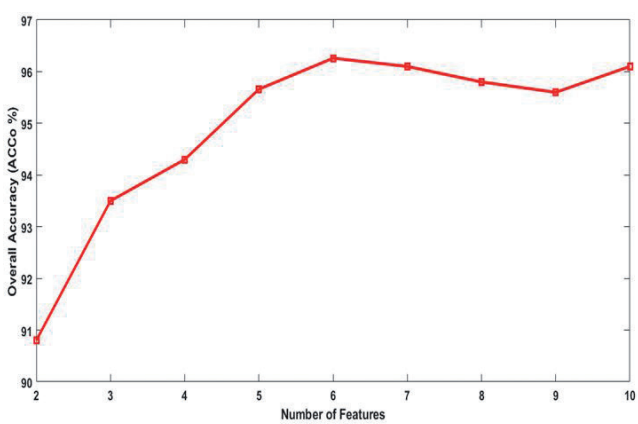

Fig. 8. *Overall accuracy (ACCo%) vs. the number of features for the RF classifier.*

Table 5

Performance of the proposed method in terms of the confusion matrix using testing set

			Prediction	
		NB	SVEB	VEB
	NB	43494	653	69
	SVEB	367	1447	23
True	VEB	377	15	2818
	FB	269	2	67
	QB	3	1	5

in [38], which show the number of most distinctive features in a subset is typically few and eliminating the less significant ones can effectively boost the classifier's performance.

Besides the normalized RR intervals features, the largest number of the ten most significant features is concerned with the duration of the QRS complex computed at specific magnitudes. This indicates that those particular features are more robust identifiers of the QRS shape than fiducial points that are characterized by small-amplitude variations, which are often disturbed by baseline noise. **Tables 5** and **6** demonstrate the performance of the proposed framework in terms of the confusion matrix and various assessment metrics, respectively. Since the

Table 6

Performance of the suggested approach in terms of Sen, PPr, F1-Score, and ACCo using the testing set

	Sen	PPr	F1-Score	ACCo
NB	98.04%	97.71%	98.04%	
SVEB	78.77%	68.32%	73.17%	**98.26%**
VEB	87.78%	94.50%	91.02%	

wrong classification of QB and FB is not taken into consideration, according to the AAMI guidelines, therefore, they are not regarded in this study (as the most prior works), but they are still involved in the assessment. It is noteworthy that the F1-Score gives a more reliable indication for the performance of the classifiers, especially when the data are unbalanced. In the context of a comparison with previous works, the proposed framework is distinguished from approaches in previous researches by being comprehensive and fully automated. In contrast to the previous methods, which did not automatically detect the locations of the R peaks (they work only if these locations are pre-defined in the database), the proposed framework presented a new, simple, and efficient method to detect these locations based on trigonometry. **Table 7** compares the results achieved with the testing set of this paper with

Table 7

A comparison of the performance of the suggested framework with state-of-the-art methods

Referer	Extracted Features	Size of used Features	Classifier	F1 Class of NB	F1 Class of SVEB	F1 Class of VEB	ACCo
[48]	RR intervals, time morphology, and WT	101	Ensemble of SVM	97.59%	57.68%	92.37%	93.8%
[37]	RR intervals, time morphology, HOS, and WT	45	Ensemble of SVM	97.04%	60.74%	94.29%	94.5%
[46]	GM, HOS, and RR intervals	33	Ensemble of BDT	97.89%	**88.64%**	85.82%	96.15%
[49]	DCT Random Projections, RR intervals	33	SVM	96.88%	33.37%	77.29%	93.1%
[27]	RR intervals, time domain morphology, and WT	8	Linear Dis-criminant	96.48%	17.16%	83.89%	93%
[50]	HOS	4	Naive Bayes	65.6%	0.4%	84.7%	94.0%
	RR intervals, HOS, WT and time domain morphology	45	SVM and DNN	91.50%	54.45%	**94.47%**	88.69%
[33]	–	–	RCNN	93.93%	45.9%	83.89%	88.99%
Presented Paper	**RR intervals, time domain morphology, and HBF**	**6**	RF	**98.04%**	73.17%	91.02%	**96.26%**

WT: Wavelet Transform, SVM: Support Vector Machine,
BDT: Bootstrap aggregated Decision Trees.

several methodologies presented in the literature that attained a very good ECG signal classification performance and used the AAMI advice and the inter-patient concept. It has been shown that the overall accuracy obtained in this paper is one of the best achieved using the same assessment indicators (ACC = 96.26%), outperforms the performance accomplished in [46] (ACCo = 96.15%), which is, according to our research, the highest in the previous works. The researchers in [46] utilized the Gaussian Mixture (GM) approach (which is computationally expensive) to obtain various features, and they used the Bagging Tree technique for classification purposes. The RF classification scheme can give better results than Bagging Tree because RF provides a minimum correlation between trees; this was also highlighted in [47]. The suggested method in [46] used one hundred trees to obtain their results, while the results introduced in this paper use forty trees only. Moreover, the proposed method is more general and practical in the real world because it deals with the problem of WGN that can negatively affect the classification results. This problem is neglected in [46] and in most previous methods. The best detection results for SVEB were achieved in [46]; however, the suggested framework outperformed in terms of detection of VEB. The presented framework ranked second in the detection of VEB (F1-score = 91.02%) after the first rank which, was achieved in [51] (F1-score = 94.47%). Nevertheless, our model is more efficient in terms of complexity, simplicity, and fit for wearable and portable devices than the one offered in [51], which is based on DNN. The detection of VEB is especially important because its recurrence can be utilized as a predictor of the most dangerous disease, "Heart Failure," which often leads to death [14,15].

4. CONCLUSION

In this research, a comprehensive, robust, efficient, and fully automated framework for ECG signal classification based on digital signal processing, feature engineering selection, and RF classifier was suggested. Optimal incorporation between features engineering and the RF classifier was introduced to meet the heartbeat classification task for devices with limited resources. The ECG signal classification performance was reasonably calculated by obeying the AAMI advice and the inter-patient approach. The most significant features for the ECG classification problem were the normalized features associated with the RR intervals and durations of the QRS complex computed at specific magnitudes. The optimal outcomes were attained with the best six most significant features and forty decision trees of the RF classifier. The assessment was based on MIT-BIH-ARR-DB. By comparing the proposed scheme in this paper with the reported literature, it can be noted that our findings are one of the best performances achieved to present. The outcomes not only asserted that the RF classifier is a superior approach for classifying the ECG signals, but also, comparatively, few numbers of features and trees are adequate to achieve or outperform literature performance.

Appendix A

Tuning Y_1 and Y_2 to obtain the smallest detection error

			Y_2		
Y_1	0.00006	0.00008	0.00018	0.00012	0.00014
0.2	0.350	0.281	0.262	0.251	0.624
0.3	0.235	0.218	0.216	0.207	0.218
0.4	0.198	0.191	0.190	0.185	0.193
0.5	0.198	0.196	0.195	0.193	0.200
0.6	0.200	0.195	0.208	0.203	0.220

Appendix B

Extracting the key points of ECG signal

The highest absolute amplitude of the signal in the period 100 ms prior to and following the R peak is assigned as a reference base point. The following procedures have been conducted for obtaining the ECG key points: (note the R peaks are detected and extracted as explained in section 2.3)

1 Initially, suppose that the peaks of P, Q, and S = 0 (no associated waves exist).

2. Go back before QRSmp and determine the transition points as follows:

i. Set [QRSmp/2p] = the first position at which the signal value is under QRSmp/2.

ii. Set [QRSmp/4p] = the first position at which the signal value is under QRSmp/4.

iii. Set Qpeak = amplitude of the first negative transition point.

iv. If the first transition point $t_{rp} \geq 0$, then specify this point as QRSb, set Speak = QRSmp, and Qpeak = 0.

v. If QRSmp > 0, and the second $t_{rp} < 0$, and Qpeak = 0, then specify the amplitude of this point as Qpeak.

vi. If Qpeak ≠ 0, and the signal passes through 0, then set QRSb = the first point that is ≥ 0.

vii. If the second $t_{rp} \geq 0$, and QRSb has still not been determined, then specify this point as QRSb.

3. Go ahead after QRSmp and determine the transition points as follows:

i. Set [QRSmp/2n] = the first position at which the signal value is under QRSmp/2.

ii. Set [QRSmp/4n] = the first position at which the signal value is under QRSmp/4.

iii. Set Speak = amplitude of the first t_{rp}, that is < 0.

iv. If Speak ≠ 0, and the signal passes through 0, then set QRSe = the first point that is ≥ 0.

v. If the second $t_{rp} \geq 0$ and QRSe has still not been determined, then specify this point as QRSe.

4. Evaluate the maximum amplitude of the signal in the period between 236 ms and 70 ms prior to QRSb (MApriQRSb[236,70]). If this amplitude is > 3·STD of the signal through the period of 70 ms before the period under consideration and its location matches the trp of the signal, then set Ppeak = MApriQRSb[236,70].

Note: QRSb: beginning of the QRS, QRSe: ending of the QRS, STD: Standard Deviation.

REFERENCES

1. Xu J, Xiao W, Liang X, Shi L, Zhang P, Wang Y, Wang Y, Yang H. A meta-analysis on the risk factors adjusted association between cardiovascular disease and COVID-19 severity. *BMC Public Health*, 2021, 21(1):1533; doi: 10.1186/s12889-021-11051-w.
2. Tyagi A, Mehra R. 2021. Intellectual heartbeats classification model for diagnosis of heart disease from ECG signal using hybrid convolutional neural network with GOA. *SN Applied Sciences*, 2021, 3(2):1-14.
3. Pal HS, Kumar A, Vishwakarma A, Singh GK, Lee HN. An effective ECG signal compression algorithm with self-controlled reconstruction quality. *Computer Methods in Biomechanics and Biomedical Engineering*, 2023, pp. 1-11. DOI: 10.1080/10255842.2023.2206933.
4. Ali AM, Ahmed AF, Najim AH. 2020, November. Efficient and Effective Scheme for ECG Compression. Proc. *2nd Annual International Conference on Information and Sciences (AiCIS)*, 2020, pp. 91-94. DOI: 10.1109/AiCIS51645.2020.00024.
5. Martínez JP, Almeida R, Olmos S, Rocha AP, Laguna P. A wavelet-based ECG delineator: evaluation on standard databases. *IEEE Transactions on biomedical engineering,* 2004, 51(4):570-581.
6. Sharma M, Dhiman HS, Acharya UR. 2021. Automatic identification of insomnia using optimal antisymmetric biorthogonal wavelet filter bank with ECG signals. *Computers in Biology and Medicine*, 2021, 131:104246.
7. Hussain I, Park SJ. Big-Ecg: Cardiographic Predictive Cyber-Physical System for Stroke Management. *IEEE Access*, 2021, 9:123146-123164.
8. Chandra S, Sharma A, Singh GK. A comparative analysis of performance of several wavelet based ECG data compression methodologies. *IRBM*, 2021, 42(4):227-244. DOI: 10.1016/j.irbm.2020.05.004.
9. Yan Z, Zhou J, Wong WF. 2021. Energy efficient ECG classification with spiking neural network. *Biomedical Signal Processing and Control*, 2021, 63:102170.
10. Saini I, Singh D, Khosla A. 2013. QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of advanced research*, 2013, 4(4):331-344.
11. Martis RJ, Acharya UR, Mandana KM, Ray AK, Chakraborty C. Cardiac decision making using higher order spectra. *Biomedical Signal Processing and Control*, 2013, 8(2):193-203.
12. Kim H, Kim S, Van Helleputte N, Artes A, Konijnenburg M, Huisken J, Van Hoof C, Yazicioglu RF. A configurable and low-power mixed signal SoC for portable ECG

monitoring applications. IEEE transactions on biomedical circuits and systems, 2013, 8(2):257-267.

13. Scirè A, Tropeano F, Anagnostopoulos A, Chatzigiannakis I. Fog-computing-based heartbeat detection and arrhythmia classification using machine learning. *Algorithms*, 2019, 12(2):32.
14. Acharya T, Tringali S, Bhullar M, Nalbandyan M, Ilineni VK, Carbajal E, Deedwania P. Frequent atrial premature complexes and their association with risk of atrial fibrillation. *The American journal of cardiology*, 2015, 116(12):1852-1857.
15. Liaqat S, Dashtipour K, Zahid A, Assaleh K, Arshad K, Ramzan N. Detection of atrial fibrillation using a machine learning approach. *Information,* 2020, 11(12):549. DOI:10.3390/info11120549.
16. Pyakillya B, Kazachenko N, Mikhailovsky N. Deep learning for ECG classification. *Journal of Physics Conference Series,* 2013, 913(1):012004. DOI: 10.1088/1742-6596/913/1/012004.
17. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 2019, 25(1):65-69.
18. Murat F, Yildirim O, Talo M, Baloglu UB, Demir Y, Acharya UR. Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. *Computers in biology and medicine*, 2020, 120:103726.
19. Luz EJDS, Schwartz WR, Cámara-Chávez G, Menotti D. 2016. ECG-based heartbeat classification for arrhythmia detection: A survey. *Computer methods and programs in biomedicine*, 2016, 127:144-164.
20. Ani R, Krishna S, Anju N, Aslam MS, Deepa OS. Iot based patient monitoring and diagnostic prediction tool using ensemble classifier. *Proc. International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 1588-1593.
21. Saffari A, Leistner C, Santner J, Godec M. Bischof H. On-line random forests. *Proc. IEEE 12th international conference on computer vision workshops*, 2009, pp. 1393-1400.
22. Lempitsky V, Verhoek M, Noble JA, Blake A. Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography. *Proc. International Conference on Functional Imaging and Modeling of the Heart,* 2009, pp. 447-456. Springer, Berlin, Heidelberg.
23. Vamsi IV, Abhinav N, Verma AK, Radhika S. Random forest based real time fault monitoring system for industries. *Proc. 4th International Conference on Computing Communication and Automation (ICCCA),* 2018, pp. 1-6.
24. Zhao X, Song Z, Guo J, Zhao Y, Zheng F. Real-time hand gesture detection and recognition by random forest. *Proc. Int. Conf. "Communications in computer and information processing" (ICCIP)*, 2012, 289:747-755. Springer, Berlin, Heidelberg.
25. Li J, Zhong PA, Yang M, Zhu F, Chen J, Liu W. Xu S. Intelligent identification of effective reservoirs based on the random forest classification model. *Journal of Hydrology,* 2020, 591:125324.
26. Emanet N. ECG beat classification by using discrete wavelet transform and Random Forest algorithm. *Proc. Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control,* 2009, pp. 1-4.
27. Llamedo M, Martínez JP. 2010. Heartbeat classification using feature selection driven by database generalization criteria. *IEEE Transactions on Biomedical Engineering,* 2010, 58(3):616-625.
28. Alickovic E, Subasi A. Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier. *Journal of medical systems*, 2016, 40(4):108.
29. Gutierrez-Gnecchi JA, Morfin-Magaña R, del Carmen A, Tellez-Anguiano DLE, Reyes-Archundia E, Díaz OH. Cardiac Arrhythmia Classification Using a Combination of

Quadratic Spline-Based Wavelet Transform and Artificial Neural Classification Network. *Proceedings 2nd International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)*, 2014, Vols 1, pp.1743-1754.

30. Kumar RG, Kumaraswamy YS. Investigating cardiac arrhythmia in ECG using random forest classification. *Int. J. Comput. Appl,* 2012, 37(4):31-34.
31. Park J, Lee S, Kang K. Arrhythmia detection using amplitude difference features based on random forest. *Proc. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 5191-5194.
32. Aravind S, Sanjay M. ECG Classification and Arrhythmia Detection Using Wavelet Transform and Convolutional Neural Network. *Proc. International Conference on Communication, Control and Information Sciences (ICCISc)*, 2021, Vol. 1, pp. 1-5.
33. Li Y, Qian R, Li K. Inter-patient arrhythmia classification with improved deep residual convolutional neural network. *Computer Methods and Programs in Biomedicine*, 2022, 214:106582.
34. https://physionet.org/content/mitdb/1.0.0/.
35. Wang T, Lu C, Sun Y, Yang M, Liu C, Ou C. 2021. Automatic ECG Classification Using Continuous Wavelet Transform and Convolutional Neural Network. *Entropy,* 2021, 23(1):119.
36. De Chazal P, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE transactions on biomedical engineering,* 2004, 51(7):1196-1206.
37. Mondéjar-Guerra V, Novo J, Rouco J. Penedo MG, Ortega M. Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers. *Biomedical Signal Processing and Control*, 2019, 47:41-48.
38. De Lannoy G, François D, Delbeke J, Verleysen M. Weighted SVMs and feature relevance assessment in supervised heart beat classification. *Proc. International Joint Conference on Biomedical Engineering Systems and Technologies*, 2010, pp. 212-223. Springer, Berlin, Heidelberg.
39. Osowski S, Linh TH. ECG beat recognition using fuzzy hybrid neural network. *IEEE Transactions on Biomedical Engineering,* 2001, 48(11):1265-1271.
40. Ince T, Kiranyaz S, Gabbouj M. A generic and robust system for automated patient-specific classification of ECG signals. *IEEE Transactions on Biomedical Engineering*, 2009, 56(5):1415-1426.
41. Mar T, Zaunseder S, Martínez JP, Llamedo M, Poll R. Optimization of ECG classification by means of feature selection. *IEEE transactions on Biomedical Engineering,* 2011, 58(8):2168-2177.
42. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. *Physical review E,* 2004, 69(6):066138.
43. Ross BC. Mutual information between discrete and continuous data sets. *PloS one*, 2014, 9(2):e87357.
44. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. *The Journal of machine Learning research,* 2011, 12:2825-2830.
45. Tsipouras MG, Fotiadis DI, Sideris D. An arrhythmia classification system based on the RR-interval signal. *Artificial intelligence in medicine,* 2005, 33(3):237-250.
46. Afkhami RG, Azarnia G, Tinati MA. Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals. *Pattern Recognition Letters*, 2016, 70:45-51.
47. Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. A comparison of decision tree ensemble creation techniques. *IEEE transactions on pattern analysis and machine intelligence*, 2006, 29(1):173-180.
48. Huang H, Liu J, Zhu Q, Wang R, Hu G. A new hierarchical method for inter-patient heartbeat classification using random projections and RR intervals. *Biomedical engineering online*, 2014, 13(1):1-26.
49. Chen S, Hua W, Li Z, Li J, Gao X. Heartbeat classification using projected and dynamic

features of ECG signal. *Biomedical Signal Processing and Control*, 2017, 31:165-173.

50. Marinho LB, de MM Nascimento N, Souza JWM, Gurgel MV, Reboucas Filho PP, de Albuquerque VHC. A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification. *Future Generation Computer Systems*, 2019, 97:564-577.
51. Li Y, He Z, Wang H, Li B, Li F, Gao Y, Ye X. CraftNet: a deep learning ensemble to diagnose cardiovascular diseases. *Biomedical Signal Processing and Control,* 2021, 62:102091.