EJONS International Journal on Mathematic, Engineering and Natural Sciences | ISSN 2602 - 4136

Article Arrival Date: 25.11.2020 | Article Type: Research Article | Article Published Date: 15.12.2020

Doi Number: http://dx.doi.org/10.38063/ejons.361

ARRHYTHMIA DIAGNOSIS FROM ECG SIGNAL USING TREE-BASED MACHINE LEARNING METHODS

Önder YAKUT

Kocaeli University, Department of Distance Education Research and Application Center, Kocaeli, Turkey, onder.yakut@kocaeli.edu.tr, ORCID: 0000-0003-0265-7252 (corresponding author)

Emine DOĞRU BOLAT

Kocaeli University, Faculty of Technology, Department of Biomedical Engineering, Kocaeli, Turkey, ebolat@kocaeli.edu.tr, ORCID: 0000-0002-8290-6812

ABSTRACT

Today, heart diseases that cause the death of people are becoming more common. Pre-diagnosis of heart diseases is important for both patients and clinicians. Electrocardiography (ECG) is a bioelectrical signal used in the diagnosis of heart disease. Noise reduction, feature extraction, feature selection and classification methods that diagnose cardiac arrhythmia are being developed to be used in computer-aided diagnostic systems. In this study, a computer-aided diagnosis system that detects arrhythmia using the MIT-BIH Arrhythmia Database (MIT-BIH AD) is proposed. ECG recordings in MIT-BIH AD are denoised of baseline wander using Chebyshev Type II Filter. Then, the positions of the R peaks belonging to the heartbeats in the ECG recordings were obtained by using the annotation files in MIT-BIH AD. The R peaks of the heartbeats in the ECG signal were divided into sub-bands using the DWT method using a 256 sample-wide window. The features have been created by using the sub-band coefficients. The obtained features have been normalized in the interval of [0,1]. Significance levels of features have been found using SelectKBest method, chi2 score function. The most effective 27 features have been obtained by using these levels. In this study, two data sets consisting of 170 features and 27 selected features have been obtained. These data sets have been divided into two as the training set (2/3) and testing set (1/3). In this study, Random Forest, Extra Trees and Decision Tree Classifiers have been used as machine learning methods. Among these methods, Random Forest classifier has obtained the best performance result. Finally, a computer-aided diagnosis system has been proposed to assist healthcare professionals in the diagnosis of arrhythmia using the data set containing 10-class arrhythmia heartbeats and with DWT-based features. The software and hardware requirements of the machine learning methods in this study have been met using Google Cloud Computing based Google Colaboratory.

Keywords:

Arrhythmia Classification, Cloud Computing, Google Colaboratory, Machine Learning, Signal Processing

1. INTRODUCTION

Cardiac arrhythmias are common today due to factors such as diseases, sedentary life, obesity, stress and an unhealthy diet. According to World Health Organization data, cardiovascular diseases are the leading cause of death worldwide (Uri1, 2020). Early diagnosis of heart diseases, appropriate counselling services for sick or high-risk people and fulfilling the need for treatment

will increase the quality and lifespan of human life. Today, early diagnosis of heart diseases has become easier with the help of developing technology. Computer-aided diagnostic systems are helpful tools for clinicians in diagnosing disease and making disease-related decision processes. Fast and accurate analysis are made by help of new methods and techniques to be developed in computer-aided diagnosis systems. In order to analyze ECG recordings, research has been carried out in various fields such as noise removal, finding fiducial points of the ECG signal, extracting features, feature selection and machine learning methods. Some studies in the literature are as follows.

Ullah and Anwar have proposed a deep one-dimensional convolutional neural network (1D-CNN) that can accurately classify five types of ECG heartbeats (Ullah and Anwar 2020). Liu et al. proposed an approach using wavelet scattering transform to automatically classify the four categories of arrhythmia (Liu et al., 2020). Hsu et al. presented a waveform-based signal processing method in their study. Using this signal processing method, they performed a machine learning and deep learning-based arrhythmia classification (Hsu et al., 2020). Ramírez et al. have presented a method combining neural networks and fuzzy logic to construct a hybrid model as a classification system using 2-leads for cardiac arrhythmias (Ramírez et al., 2020). Sahoo et al. They proposed an automated system of classifying ECG beats based on multi-field characteristics derived from ECG signals (Sahoo et al., 2020).

In this study, the ECG signal was denoised from the baseline wander. Then, the location information of the R peaks in the ECG signal was obtained with the help of annotation files. DWT based features are extracted from the ECG signal. The significance levels of these features were determined by the SelectKBest method and the most effective ones were selected. Finally, the heartbeats were classified with Tree-based Random Forest, Extra Trees and Decision Tree machine learning methods using the feature data sets.

2. MATERIALS AND METHODS

2.1. Data Set

In this study, The MIT-BIH AD which has ECG recordings containing arrhythmic heartbeats (Moody et al., 2001; Uri2, 2020) was used. The distribution of beat types in the data set including arrhythmic heartbeats is shown in Table 1.

Table 1. Distribution of Arrhythmic heartbeats

No	Beat Types	Total	Training	Testing
1	NORMAL (N)	75017	50011	25006
2	LBBB (L)	8072	5381	2691
3	RBBB (R)	7255	4837	2418
4	APC (A)	2546	1697	849
5	NAPC (x)	193	128	65
6	PVC (V)	7129	4752	2377
7	FLWAV (!)	472	314	158
8	FUSION (F)	802	534	268
9	PACE (P)	7024	4682	2342
10	PFUS (f)	982	654	328

When the data in Table 1 are examined there are a total of 109492 heartbeats. From these heartbeats, 72990 beats (2/3) were reserved for the training set and 36502 beats (1/3) for the test set. Tree-based machine learning methods were fed for classification by using training and test data sets that have a distribution as in Table 1.

2.2. ECG Signal Preprocessing

Baseline wandering is low-frequency noise below 1 Hz in the ECG signal (Yakut et al., 2018). In this study, a low pass Chebyshev Type II filter was used to eliminate or reduce the change in baseline wander in ECG recordings.

Figure 1 shows the effect of the developed filter on the signal. The raw ECG signal of the 105th ECG recording is shown in Figure 1 (a). Baseline wander noise in the ECG signal is shown in Figure 1 (b). The filtered ECG signal is shown in Figure 1 (c).

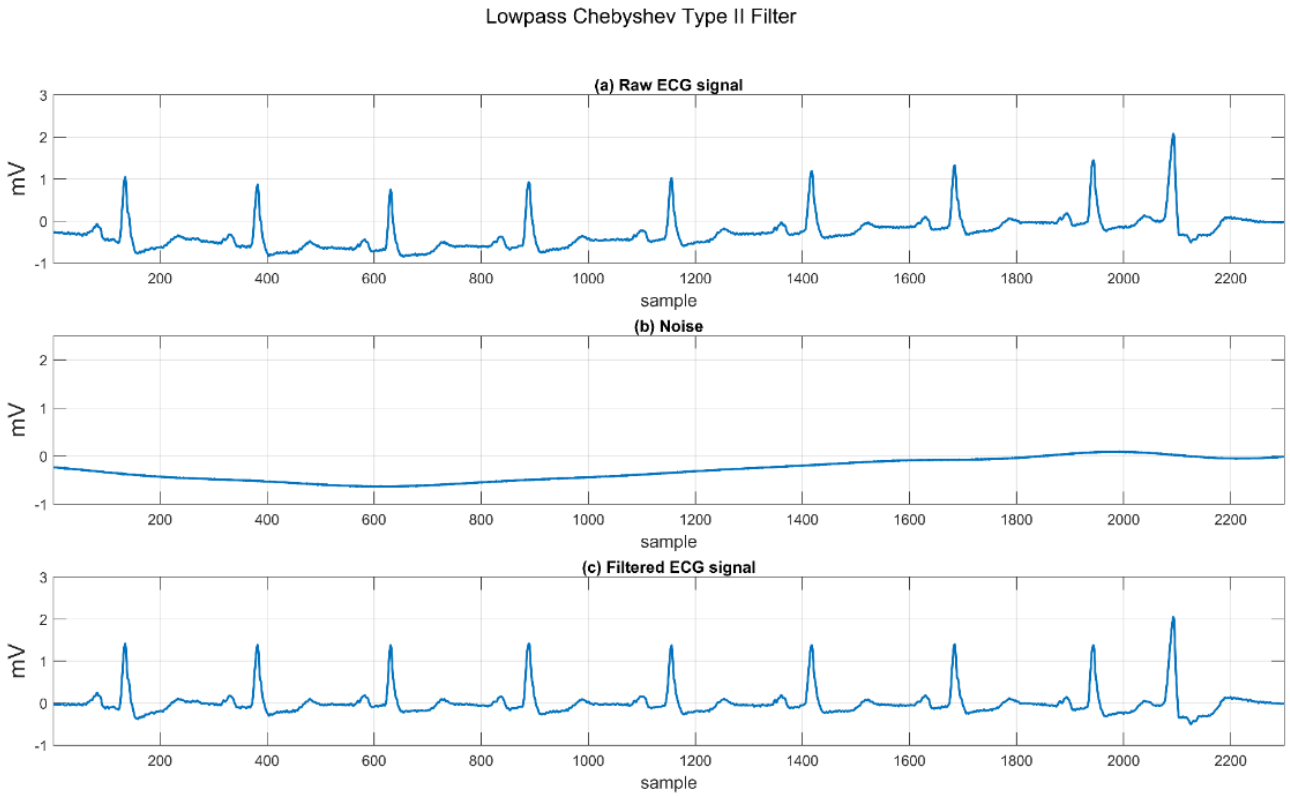

Figure 1. Elimination of ECG signal noise with Chebyshev Type II filter (105th ECG signal)

2.3. Discrete Wavelet Transform (DWT) Based Feature Extraction

ECG heartbeats divided as 256-sample segments of which 127 samples before R peaks and 128 samples after the R peaks were examined with DWT for each heartbeat. With the DWT method, the ECG signal was divided into 5 sub-bands as seen in Figure 2. The db8 (Daubechies) was used as a wavelet family. In this study, since the arrhythmic beats in the ECG signal would be examined, wavelet transform was performed up to the 5th decomposition level with a window of 256 samples including a cardiac cycle (Yakut, 2018).

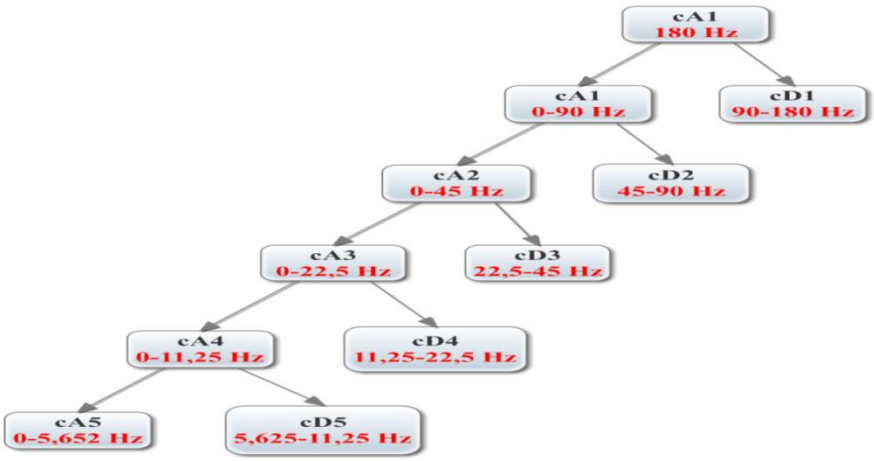

Figure 2. DWT level 5 decomposition tree, frequency ranges of detail (cA) and approximation (cD) coefficients at each level (Yakut, 2018)

When we examined the components of the ECG signal as P wave, QRS complex, and T wave morphology, it was seen that it was similar to the db8 wavelet family. For this reason, the ECG signal was divided into 5 sub-bands using the DWT method using the db8 wavelet family, and feature extraction was performed (Yakut, 2018).

Table 2. DWT-based extracted features

Number	**Features**
DWT 1-22	DWT 5th level approximation (cD5) coefficients
DWT 23-32	DWT absolute averages of the obtained coefficients
DWT 33-40	Ratios of absolute mean values of adjacent coefficients
DWT 41-50	Autocorrelation values of the coefficients
DWT 51-60	Median values of the coefficients
DWT 61-70	Lowest and highest rates of coefficients
DWT 71-80	Variance values of the coefficients
DWT 81-90	Standard deviation (std) values of the coefficients
DWT 91-100	Quadrant values (iQR) values of the coefficients
DWT 101–110	Mean absolute deviation (mad) values of the coefficients
DWT 111-120	Square root values of the square averages of the coefficients
DWT 121-130	Average Shannon Entropy values of the coefficients
DWT 131-140	Energy density (rms) values of the coefficients
DWT 141-150	Kurtosis values of the coefficients
DWT 151-160	Skewness values of the coefficients
DWT 161-170	Moment values of the coefficients

Features were obtained using various statistical methods and techniques, detail and approximation coefficients of DWT. By this way, since wavelet transform technique is mostly used in feature extraction process, useful information about ECG signal in both time and frequency domain was obtained in this study. For each heartbeat separated into 256-sample segments, DWT was applied and used for feature extraction by obtaining the detail and approximation coefficients at each level of wavelet decomposition. In the wavelet transform, which is detailed in Table 2, the features of the ECG signal were calculated by using approximation and detail coefficients of each sub-band in different methods and techniques (Yakut, 2018).

2.4. Feature Selection

In this study, the SelectKBest feature selection method which sorts the DWT-based 170 features according to their significance, was used. SelectKBest removes all features except the highest-scoring ones (Uri3, 2020). The SelectKBest method was used with the chi2 (chi-square test) (Uri4, 2020) score function. Chi2 score function was used to select the highest value features for chi-square statistics. Chi-square test extracts features that do not have any effect on classification due to the dependency between stochastic variables (Uri4, 2020).

In this study, the significance levels of the DWT-based 170 feature were determined by using the SelectKBest method and chi2 score function. Among these features, the most effective selected 27 features are shown in Figure 3. Thus, a second data set with 27 features was obtained. This data set with the reduced size was used in machine learning methods for classification.

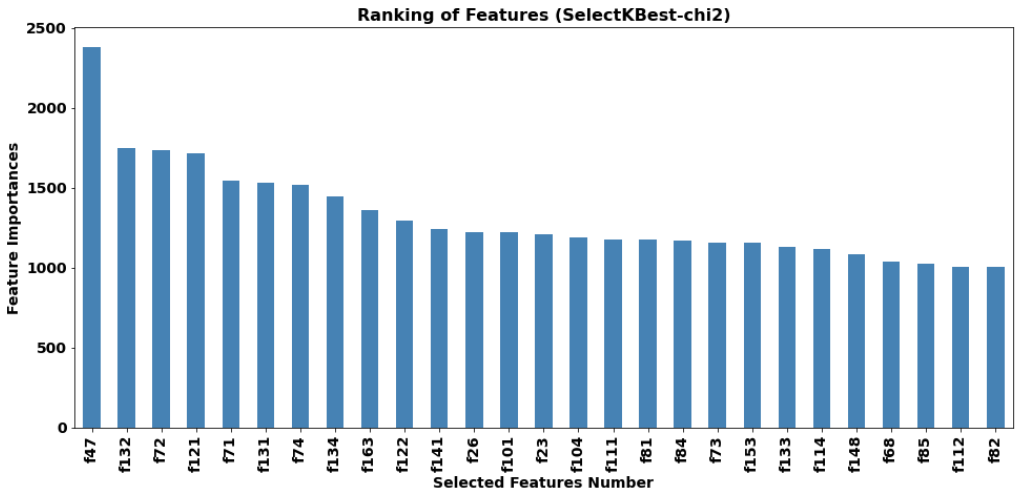

Figure 3. Selected 27 features by the SelectKBest method

In Figure 3, features are represented on the x-axis. The significance levels of the features are represented on the y-axis. The features are ranked from high to low significance according to their importance.

2.5. Extra Treess Classifier

Extra Trees Classifier is a collective learning technique that combines the results of multiple unrelated decision trees gathered in a forest to output the classification result. In concept, it is very similar to the Random Forest Classifier and differs from it in terms of the construction of decision trees in the forest (Uri5, 2020). The first difference is that identifiers are completely randomly split in the node. The second difference is that each tree is enlarged with the entire data set rather than a bootstrap sampling (Geurts et al., 2006; Mishra et al., 2017).

2.6. Decision Tree

Decision trees are commonly used structures for classification purposes. The root node of the tree contains the basic feature that is sought. Each internal node contains one of the test parameters while each branch holds the result of the corresponding test parameter. The value found in leaf nodes is the decision expression. While reaching the decision, the decision tree runs a series of tests. When you want to find the value of a feature it is progressed on the tree until the result is reached (Sathyadevan and Nair, 2015).

2.7. Random Forest

Random Forest (RF) is a supervised learning algorithm proposed by Breiman (Breiman, 2001) in 2001 and can be used in classification and regression problems. RF which is formed by the combination of estimates of trees trained separately from each other is a community learning method that makes new predictions by averaging these estimates (Denil et al., 2014). Each decision tree in the community votes for each situation to be classified. The prediction of this algorithm is obtained by the majority of the trees (Nisbet et al., 2018; Shalev-Shwartz et al., 2014).

3. EXPERIMENTAL STUDY

In this study, Matlab R2015b (The MathWorks, Inc., Natick, MA, USA) software was used to extract DWT-based feature and denoise the ECG signal. The methods that use feature selection and classification have been developed utilizing the Sci-kitlearn machine learning library (Uri6, 2020). Python was used as the programming language. The software and hardware requirements of the machine learning methods in this study were met using Google Cloud Computing-based Google Colaboratory (Uri7, 2020).

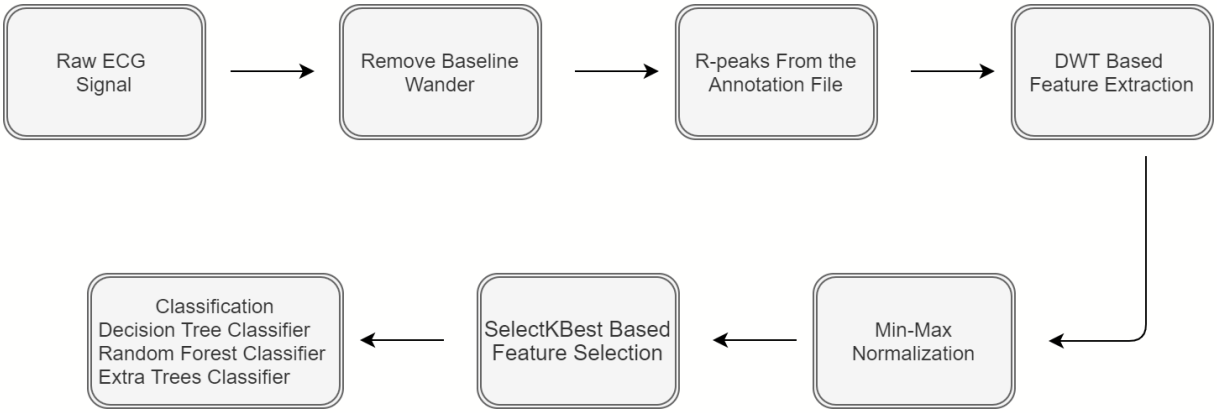

Figure 4. Block diagram of the proposed method

The block diagram of the proposed method in this study is shown in Figure 4. Signal preprocessing is performed by removing the raw ECG signal baseline noise. In the study, the noise in the ECG signal was removed using the Chebyshev Type II filter. Then, the locations of the R peaks of each heartbeat were obtained using annotation files. Each heartbeat was sub-banded utilizing the DWT method using a window of 256 samples. The features were extracted by using the coefficients of sub-bands and statistical methods. Thus, 170 features detailed in Table 1 were obtained. These features were normalized in the interval [0,1] using the Min-Max normalization. In this study, 10 different types of heartbeats including normal and arrhythmic beat were used. The extracted features were ranked according to their significance with the SelectKBest (Uri3, 2020) method and chi2 (Uri4, 2020) score function in the Sci-kitlearn (Uri6, 2020) machine learning library. So, the most successful 27 features were selected as a result of the feature selection process. The data set with 170 features extracted by the DWT method has been reduced to 27 features. In this study, both the data set with 170 features and the data set with 27 features were used. These data sets were

divided into the training set (72990 heartbeats, 2/3) and the test set (36502 heartbeats, 1/3). Decision Tree, Random Forest and Extra Trees Classifiers which are tree-based machine learning methods developed within the scope of the study were used. The obtained classification results were compared. Thus, a method for arrhythmic heartbeat detection was proposed.

4. EXPERIMENTAL RESULTS

In this study, the results obtained by using the data set with 170 features are shown in Table 3. Machine learning methods in Table 3 are listed from high to low performance. Also, the Receiver Operating Characteristic Area Under the ROC Curve (ROC AUC) graph showing the performance of Random Forest, Extra Treess and Decision Tree Classifiers is shown in Figure 5.

Table 3: Performance results of tree-based methods using data set with 170 features

Methods		Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Random Forest Classifier	Training	99,876	99,618	99,869	99,743
	Testing	98,971	97,201	98,965	98,075
Extra Treess Classifier	Training	98,504	95,097	98,151	96,599
	Testing	98,499	94,599	98,115	96,325
Decision Tree Classifier	Training	99,190	97,700	99,098	98,390
	Testing	96,371	95,984	96,458	96,220

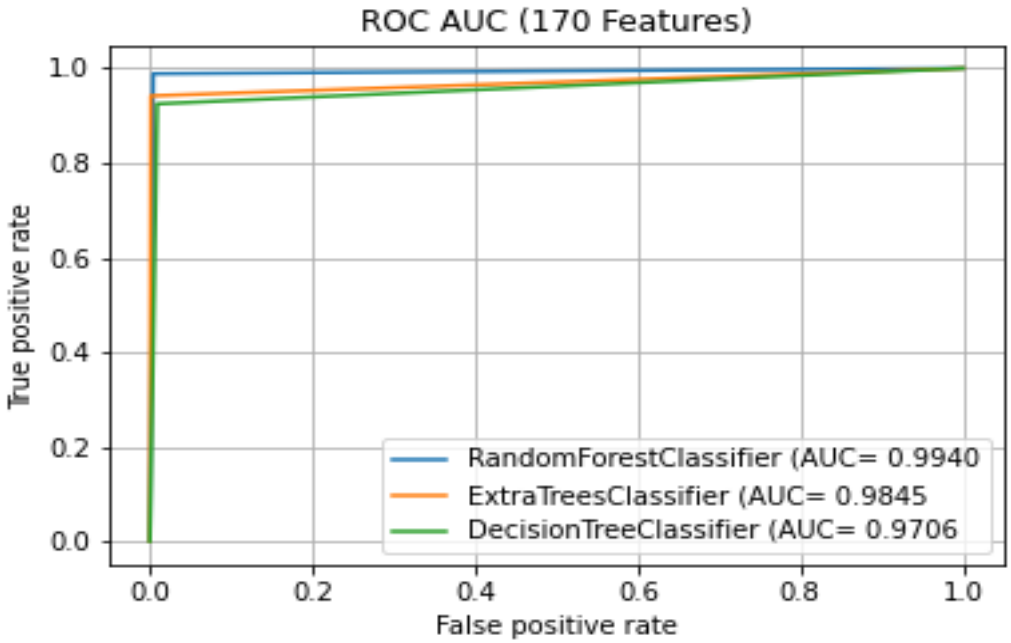

Figure 5: Comparison of the performance results of machine learning methods (170 features)

Table 4: Performance results of tree-based methods using selected data set with 27 features

Methods		Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Random Forest Classifier	Training	99,467	98,174	99,323	98,740
	Testing	97,814	98,078	98,822	98,448
Extra Trees Classifier	Training	99,094	98,912	99,041	98,970
	Testing	98,110	97,559	98,118	97,830
Decision Tree Classifier	Training	98,628	96,514	98,429	97,292
	Testing	97,845	96,102	97,591	96,608

The results obtained by using the 27-feature data set obtained as a result of the feature selection process are shown in Table 4. Machine learning methods in Table 4 are listed from high to low performance. At the same time, the ROC AUC graphs showing the performance of Random Forest, Extra Trees and Decision Tree Classifiers are shown in Figure 6.

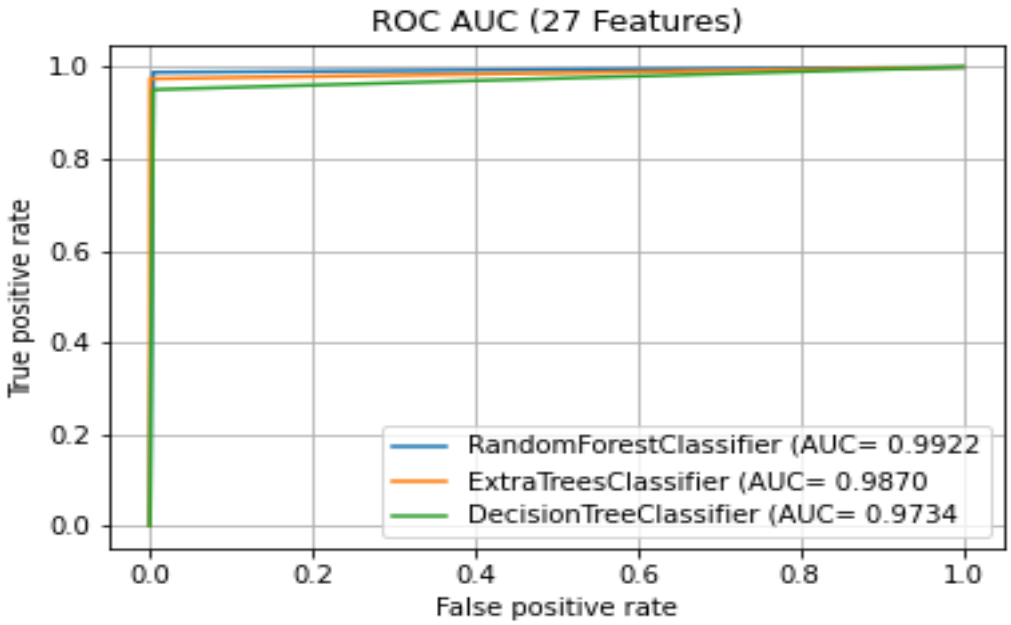

Figure 6: Comparison of the performance results of machine learning methods (27 features)

5. DISCUSSIONS

In this study, a low-pass filter was used that eliminates the baseline wander in the ECG signal. This filter eliminates the noise in the ECG signal, allowing the ECG signal to fit into the baseline.

DWT method is used to extract features from the ECG signal. The ECG signal was divided into sub-bands. The coefficients of these sub-bands and the features obtained by statistical methods from the coefficients of the sub-bands formed the data set (170 features). This data set was reduced in size using the SelectKBest method and chi2 score function and a data set with 27 features was created. Thus, the computational complexity and cost of the proposed method have been reduced. As a result of Dimensionality Reduction, the performance of machine learning methods increased

because the features that were not important and the features that had a negative effect on classification were eliminated from the data set.

When the performance results obtained by using the data set with 170 features belonging to the developed machine learning models in Table 3 are examined all methods yielded results close to each other in the same range. Random Forest method results were obtained as Accuracy 98.971%, Recall 97.201%, Precision 98.965 and F1-Score 98.075%. The results of the Extra Trees method were obtained as Accuracy 98.499%, Recall 94.599%, Precision 98.115 and F1-Score 96.325%. The results of the Decision Tree method were found as Accuracy 96.371%, Recall 95.984%, Precision 96.458 and F1-Score 96.220%. It was observed that the method with the highest success was the Random Forest method. In addition, when the ROC AUC graph in Figure 5 for arrhythmic heartbeat detection of the models was examined, it was found that the performance of the models had a perfectly satisfactory performance result.

When the performance results obtained by using the 27-feature data set belonging to the developed machine learning models in Table 4 are analyzed all methods gave results close to each other in the same range. Random Forest method results were obtained as Accuracy 97.814%, Recall 98.078%, Precision 98.822 and F1-Score 98.448%. The results of the Extra Trees method were obtained as Accuracy 98.110%, Recall 97.599%, Precision 98.118 and F1-Score 97.830%. The results of the Decision Tree method were found as Accuracy 97.845%, Recall 96.102%, Precision 97.591 and F1-Score 96.608%. It was observed that the method with the highest performance was the Random Forest method. In addition, when the ROC AUC graph in Figure 6 for arrhythmic heartbeat detection of the models was examined, it was understood that the performance of the models had an extremely satisfactory performance result.

When the F1-Score values of the test results of each machine learning method in Table 3 and Table 4 were examined it was seen that the results of arrhythmic heartbeat detection using the reduced data set (27 features) were higher than the other data set (170 features). Thus, it has been shown that the performance of machine learning methods will increase by removing unnecessary features from the data set.

In this study, when the results obtained using both data sets were analyzed the Random Forest method was proposed for the detection of the arrhythmic heartbeat.

6. CONCLUSIONS

In this study, a computer-aided diagnosis system that detecting the arrhythmic heartbeats using MIT-BIH AD is proposed. ECG recordings were filtered for eliminating the baseline wander. Then, the locations of the R peaks belonging to the heartbeats in the ECG recordings were obtained by using annotation files. Using these locations, features were extracted from heartbeats using the DWT method. Min-max normalization was applied to the extracted features. The significance levels of the features were found using the SelectKBest method and chi2 score function. In this study, two data sets consisting of 170 features and 27 selected features were obtained. The classifiers were fed by these data sets. Random Forest, Extra Trees and Decision Tree Classifiers were used as machine learning methods. Among these methods, Random Forest classifier showed the best performance result. Thus, arrhythmic heartbeat detection method was proposed to assist healthcare professionals in the diagnosis of arrhythmia using the data set containing 10-class arrhythmic heartbeats and 27 DWT-based features. In future studies, decision support systems that diagnose arrhythmia with other machine learning methods can be developed using different ECG databases containing arrhythmic beats.

REFERENCES

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Denil, M., Matheson, D., & De Freitas, N. (2014, January). Narrowing the gap: Random forests in theory and in practice. In International conference on machine learning (pp. 665-673).

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.

Hsu, P. Y., & Cheng, C. K. (2020, July). Arrhythmia Classification using Deep Learning and Machine Learning with Features Extracted from Waveform-based Signal Processing. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 292-295). IEEE.

Liu, Z., Yao, G., Zhang, Q., Zhang, J., & Zeng, X. (2020). Wavelet Scattering Transform for ECG Beat Classification. Computational and Mathematical Methods in Medicine, 2020.

Mishra, G., Sehgal, D., & Valadi, J. K. (2017). Quantitative structure activity relationship study of the anti-hepatitis peptides employing random forests and extra-trees regressors. Bioinformation, 13(3), 60.

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.

Nisbet, R., Elder, J., & Miner, G. (2018). Handbook of statistical analysis and data mining applications. Elsevier Academic Press.

Ramírez, E., Melin, P., & Prado-Arechiga, G. (2020). Hybrid Model Based on Neural Networks and Fuzzy Logic for 2-Lead Cardiac Arrhythmia Classification. In Hybrid Intelligent Systems in Control, Pattern Recognition and Medicine (pp. 193-217). Springer, Cham.

Sahoo, S., Subudhi, A., Dash, M., & Sabut, S. (2020). Automatic Classification of Cardiac Arrhythmias Based on Hybrid Features and Decision Tree Algorithm. International Journal of Automation and Computing, 1-11.

Sathyadevan, S., & Nair, R. R. (2015). Comparative analysis of decision tree algorithms: ID3, C4. 5 and random forest. In Computational intelligence in data mining-volume 1 (pp. 549-562). Springer, New Delhi.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

Ullah, A., & Anwar, S. (2020). One Dimensional Convolution Neural Network Model for ECG Arrhythmia Classification. Technical Journal, 25(02), 85-94.

Uri1, Cardiovascular diseases (CVDs), (24.11.2020), Available: https://www.who. int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

Uri2, MIT-BIH Arrhythmia Database, (15.11.2020) Available: http://www.physionet.org /physiobank/database/mitdb/.

Uri3, Sklearn feature selection SelectKBest, (18.11.2020), Available:https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html#sklearn.feature_selection.SelectKBest.

Uri4, Sklearn feature selection chi2, (17.11.2020), Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html#sklearn.feature_selection.chi.

Uri5, Extra Tree Classifier for Feature Selection, (18.11.2020), Available: https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/#:~:text= Extre

mely%20Randomized%20Trees%20Classifier(Extra,to%20output%20it's%20classification%20result.

Uri6, Scikit-learn Machine Learning in Python, (18.11.2020), Available: https://scikit-learn.org/stable/index.html.

Uri7, Google Colaboratory, (10.11.2020), Available: https://colab.research.google.com/.

Yakut, O., Solak, S., & Bolat, E. D. (2018). IIR Based Digital Filter Design for Denoising the ECG Signal. Journal of Polytechnic, 21(1), 173-181.

Yakut O., (2018). Classification of Arrhytmias in ECG Signal Using Soft Computing Algorithms. PhD, Kocaeli University, Kocaeli, Turkey.