Exploring Feature Selection and Classification Algorithms For Cardiac Arrhythmia Disease Prediction

RAVINDER AHUJA[1], SC SHARMA[2]

[1] Galgotias University Greater Noida and IIT Roorkee Saharanpur Campus, INDIA
[2] IIT Roorkee Saharanpur Campus, INDIA

Abstract: Cardiac Arrhythmia is the disease in which heartbeats abnormally due to which death of a person may occur if not diagnosed on time. Timely and accurate detection of cardiac arrhythmia can save the life of the patient. In this study fourteen classification algorithms and six feature selection algorithms are explored to find the best combination which can accurately detect cardiac arrhythmia. On the features selected through feature selection techniques fourteen classification algorithms are applied to classify cardiac arrhythmia. The random forest algorithm for feature selection and random forest classification algorithm found best among all the models applied with an accuracy of 86.57%, precision 79.12%, recall 79.12%, and f1-score 79.12%.

Keywords: Cardiac arrhythmia, Confusion Matrix, Disease prediction, Ensemble Algorithms, Feature Selection, multiclass Classification.

Received: June 14, 2021. Revised: June 10, 2022. Accepted: July 8, 2022. Published: August 5, 2022.

1. Introduction

Heart disease have affected the significant amount of the population, not only in India but all over the world. Any sort of irregular behavior in the heart may be life-threatening. A tool can be used for diagnosing the proper activity of heart, known as ElectroCardioGram (ECG), which produces a graph for electrical pulses [1]. Specific parameters are taken into account, for the proper examination of the heart. If there is a slight change in these parameters, it results in the ailment of heart, and it may occur due to several reasons [2]. An arrhythmia is a form of irregularity detected in the electronic pulses generated by the heart, and it may occur due to several reasons. If left untreated for a long time, it poses a threat to human life and may lead to a cardiac arrest. Therefore, accurate detection and classification of arrhythmia are essential. These conditions may cause the heart to beat fast or slow, skipping some beats. Because of these types of behaviors, ECG will form different graphical patterns, and therefore, arrhythmia can be easily detected [3]. There are generally two different categories, one which causes the heart to beat too slowly, usually below 60 rpm known as Bradycardia, and another one causes the heart to beat at 100 rpm, known as Tachycardia [4]. There are other types of arrhythmia too. Arrhythmias can be identified by the location point of its occurrence in heart and by the change in the rhythms generated by the heart. Supraventricular arrhythmia starts in atria; therefore, it is also known as a trial arrhythmia. But sometimes, these ECG recordings are of long duration, and it raises difficulties for a doctor to look at those and find irregularities [5]. Therefore, Machine Learning can be used
for the automation of arrhythmia diagnosis, and it will be quite helpful. In this paper, six feature selection techniques are applied to reduce the dimension and further fourteen classification algorithms, and their ensemble has been applied to classify into one of the sixteen classes. The major points of the work done in this study is as follows:

1. Six feature selection and fourteen classification techniques are explored for finding the optimized combination which can give better performance.

2. Due to small size of dataset cross validation technique is applied with different values of k (2, 5, and 10). Best results are reported at K=10.

The rest of the paper is divided into the following sections: section 2 contains related work, section 3 contains dataset description, preprocessing techniques used, feature selection techniques used, and classification algorithms, and methodology used, section 5 contains experimental results, and section 6 contains conclusion.

2. Related Work

So many techniques in the past have been developed for cardiac arrhythmia detection. Principal Component Analysis approach was applied on the ECG dataset to reduce the features, and further six neural networks were used to classify the records into normal or having cardiac arrhythmia [6]. The development in the field of automation of ECG analysis and recording the patient's health status was reported by [7]. Some of these methods that have helped in the development of ECG analysis are neural networks [8], and self-organizing map [9]. The first action that anyone has to take in saving the patient's life is the proper diagnosing of arrhythmia [10]. The missing values in the data set also play an important role. In Standard ECG (12-lead), missing value is replaced by the nearby value in the attribute, and a multilayer perceptron algorithm is applied for the classification of arrhythmia [11]. In paper [12], authors have used a correlation-based feature selection technique to select essential features from the UCI repository dataset. The neural network algorithm is applied along with the Levenberg-Marquardt method for the classification of arrhythmia. The random forest classification algorithm is implemented with resampling technique is used to classify cardiac arrhythmia [13]. In paper [14], firstly, various preprocessing techniques are applied first. Various feature selection techniques are applied, and last Different classification algorithms, including neural network, random forest, gradient boosting, are used on the ECG dataset. In paper [15-17], authors have applied various classification algorithms on the dataset to classify into 16 classes. In paper [18] PCA (Principal Component Analysis) feature selection technique is used to extract essential features, and further SVM classification algorithm is applied to classify arrhythmia. In paper [19], authors have applied neural networks on the dataset for prediction of cardiac arrhythmia and achieved an accuracy of 76.67%. In paper [20], authors have applied the Naïve Bayes classification algorithm with the train-test split ratio of 70-30 and achieved an accuracy of 70.50%. In paper [21], authors have applied feature selection techniques in two steps: the wrapper method part and the filtering part. Further SVM and KNN classification algorithms are applied on the dataset, and the best accuracy (73.80%) is achieved with 20-fold cross-validation. In paper [22], authors have firstly replaced the missing value with the closest value and applied feature selection technique, and a total of 198 features were selected. Further, modular Neural Network with three layers was used for classification and achieved an accuracy of 78.89% with a train-test split ratio of 90-10.

3. Materials and Methods

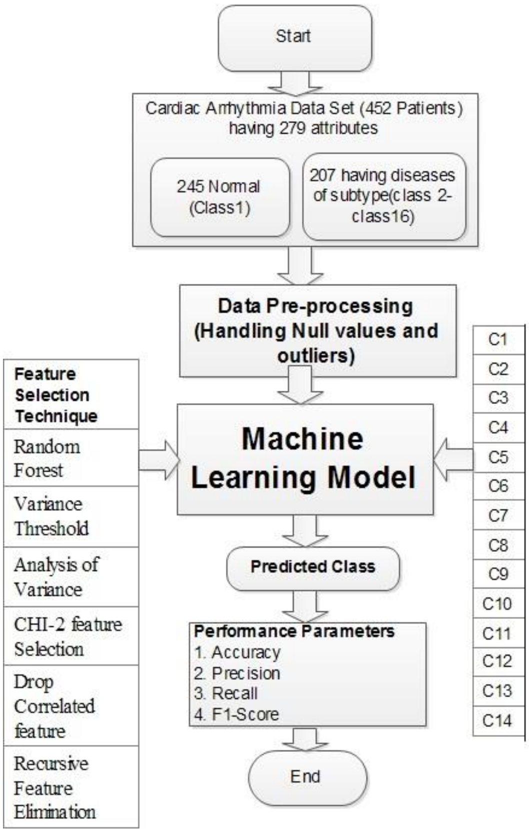
The methodology applied has different components which are as follows:

3.1. Dataset Description: The dataset is taken from the UCI machine learning repository [23]. The dataset consists of 452 rows corresponding to each patient. Two hundred seventy-nine attributes are recorded for each patient like age, weight, ECG related data, heart-related data, QRS duration, etc. There are sixteen classes associated with the dataset. Class 1 is corresponding to no arrhythmia and class 2 to 15 corresponding to different types of arrhythmia. Class 16 is corresponding to the unlabeled patient. Two hundred forty-five rows are corresponding to class 1 label, remaining 185 classes corresponding to 14 classes, and 22 rows are unlabeled.

3.2. Data Pre-processing: The dataset contains missing values and outliers. The missing values are filled with mean. The outliers are detected using the interquartile range and handled with logarithm. A standard scalar is used to normalize the data.

3.3. Feature Selection Methods: The dataset contains 279 features. To reduce complexity and make training faster, six feature selection techniques are applied. The feature selection techniques used in our study are as follows: (i) Random Forest (RF) [24] (ii) Analysis of Variance (ANOVA) [25] (iii) Variance Threshold (VT) [26] (iv) Dropping Highly Correlated Feature [27] (v) Recursive Feature Elimination [28] (vi) Chi-Square Test [29].

3.4. Classification Algorithms: Fourteen classification algorithms are applied which are as follows: (i) Support Vector Machine (SVM) [30] (ii) Bernoulli Naive Bayes (NB) [31] (iii) Random Forest (RF) [32] (no. of estimators as 1200 and random state as 42) (iv) K-Nearest Neighbors (KNN) [33] (Euclidean distance, and the number of neighbors taken is three) (v) Logistic Regression [34] (vi) Decision Tree [35] (vii) Averaging: [36] (viii) Bagging: [37] (linear kernel with the value of C = 1) (ix) Light GBM [38] (x) AdaBoost [39] (xi) XGBoost [40] (learning rate = 0.01) (xii) Stacking [41] (xiii) ID3 [42] (xiv) Majority Voting [43]



C1-C14 are classifiers

Figure 1: The overview of the methodology use

5. Experimental Results

The four performance evaluation parameters namely accuracy, f1-score, precision, and recall are considered. The hyper parameters of classification algorithms are tuned using a grid search approach. The dataset after cleaning process is given as input to feature selection algorithms and further, fourteen classification algorithms are applied. The dataset considered is small so, f-fold cross validation technique is applied considering k to be 2, 4, 5, and 10 which are presented in this section. The best results are obtained with the value of k=10. Corresponding to random forest feature selection and fourteen classification algorithms results are presented in Table 1. The threshold value is set to 0.001, which returns 163 features in random forest feature selection technique. It is observed from the results that random forest classifier has produced better results with an accuracy of 86.57%, recall of 79.12%, precision of 79.12%, and F1-score of 79.12%.

Next to the random forest is the maximum voting classifier giving an accuracy of 71.93, precision 77.12%, recall 77.12%, and f1-score 77.12%. The next to maximum voting is averaging, giving an accuracy of 73.05, precision 76.92%, recall 76.92%, and f1-score 76.92%. The worst performance is given by support vector machine classifier giving an accuracy 55.34%, precision 58.24%, recall 58.24%, and f1-score 58.24%. Corresponding to variance threshold feature Selection and fourteen classification algorithms results are presented in Table 2. Threshold Value of 0.01 is set up for selecting features, and 195 features are returned. It is observed from the results that random forest classifier has produced better results with an accuracy of 80.21%, recall of 78.23%, precision of 78.23%, and F1-score of 78.23%. Next to the random forest is the maximum voting classifier giving an accuracy of 73.22%, precision 77.12%, recall 77.12%, and f1-score 77.12%. The next to maximum voting is ID3 giving an accuracy of 71.23,

Table 1: Results based on Random Forest feature Selection Technique

Algorithm	**Accuracy**	**F1-Score**	**Recall**	**Precision**
Naive Bayes	68.78	72.52	72.52	72.52
Decision Tree	60.55	57.14	57.14	57.14
KNN	57.77	63.73	63.73	63.73
SVM	55.34	58.24	58.24	58.24
Random Forest	**86.57**	**79.12**	**79.12**	**79.12**
Bagging	64.72	69.23	69.23	69.23
Adaboost	61.67	61.43	61.43	61.43
Averaging	71.93	76.92	76.92	76.92
XGBoost	71.38	75.82	75.82	75.82
Light GBM	71.33	76.24	76.24	76.24
Max Voting	73.05	77.12	77.12	77.12
Logistic Regression	62.67	65.93	65.93	65.93
Stacking	71.38	75.82	75.82	75.82
ID3	65.89	70.55	70.55	70.55

Table 2: Results based on Variance Threshold feature Selection Technique

Algorithm	**Accuracy**	**F1-Score**	**Recall**	**Precision**
Naive Bayes	66.72	72.52	72.52	72.52
Decision Tree	59.72	58.24	58.24	58.24
KNN	58.05	63.73	63.73	6373
SVM	68.72	71.42	71.42	71.42
Random Forest	**80.21**	**78.23**	**78.23**	**78.23**
Bagging	54.72	63.73	63.73	63.73
Adaboost	61.67	63.73	63.73	63.73
Averaging	70.67	74.35	74.35	74.35
Gradient Boosting	71.11	74.72	74.72	74.72
Light GBM	58.22	61.31	61.31	61.31
MaxVoting	73.22	79.12	79.12	79.12
Stacking	71.11	74.72	74.72	74.72
Logistic Regression	67.58	69.23	69.23	69.23
ID3	71.23	74.85	74.85	74.85

precision 74.85%, recall 74.85%, and f1-score 74.85%. The KNN classifier has produced poor results among all, giving an accuracy of 58.05%, precision 63.73%, recall 63.73%, and f1-score 63.73%. Corresponding to ANOVA feature selection and fourteen classification algorithms results are presented in Table 3. The number of selected features is set up manually as 55, so it will return 55 features. It is observed from the results that random forest classifier has produced better results with an accuracy of 76.87%, recall of 78.29%, precision of 78.29%, and F1-score of 78.29%. Next to the random forest is the maximum voting classifier giving an accuracy of 75.27%, precision 77.21%, recall 77.21%, and f1-score 77.21%. The next to maximum voting is stacking, giving an accuracy of 70.83, precision 75.82%, recall 75.82%, and f1-score 75.82%. The light GBM classifier has produced poor results among all with an accuracy 59.55%, precision 62.22%, recall 62.22%, and f1-score 62.22%.

Corresponding to CHI-2 feature selection and fourteen classification algorithms results are presented in Table 4. We set the number of features manually to be 69, so all the algorithms will work on the best 69 features. It is observed from the results that random forest classifier and maximum voting has produced better results among all the classification algorithms and giving an accuracy of 74.22%, recall of 77.70%, precision of 77.70%, and F1-score of 77.70%. The next best performing classifier is stacking, giving an accuracy of 72.36%, precision 75.39%, recall 75.39%, and f1-score 75.39%. The worst performance is given by SVM classifier, giving an accuracy 55.34%, precision 58.24%, recall 58.24%, and f1-score 58.24%.

Table 4: Results based on CHI-2 feature Selection Technique

Algorithm	Accuracy	F1-Score	Recall	Precision
Naive Bayes	64.72	67.77	67.77	67.77
Decision Tree	63.33	66.63	66.63	66.63
KNN	61.38	65.93	65.93	65.93
SVM	55.34	58.24	58.24	58.24
Random Forest	**74.22**	**77.70**	**77.70**	**77.70**
Bagging	59.55	62.22	62.22	62.22
Adaboost	60.83	63.33	63.33	63.33
Averaging	63.64	67.29	67.29	67.29
Gradient Boosting	65.82	68.30	68.30	68.30
Light GBM	58.24	61.62	61.62	61.62
MaxVoting	**74.22**	**77.70**	**77.70**	**77.70**
Stacking	72.36	75.39	75.39	75.39
Logistic Regression	65.21	69.87	69.87	69.87
ID3	67.85	72.54	72.54	72.54

Table 3: Results based on the ANOVA Technique

Algorithm	Accuracy	F1-Score	Recall	Precision
Naive Bayes	64.44	67.62	67.62	67.62
Decision Tree	63.33	66.63	66.63	66.63
KNN	61.38	65.93	65.93	65.93
SVM	67.77	69.30	69.30	69.30
Random Forest	**76.87**	**78.29**	**78.29**	**78.29**
Bagging	64.72	67.77	67.77	67.77
Adaboost	63.05	63.05	63.05	63.05
Averaging	68.14	71.81	71.81	71.81
Gradient Boosting	69.83	74.28	74.28	74.28
Light GBM	59.55	62.22	62.22	62.22
Max Voting	75.27	77.21	77.21	77.21
Stacking	70.83	75.82	75.82	75.82
Logistic Regression	65.93	66.63	66.63	66.63
ID3	59.99	63.83	63.83	63.83

Corresponding to dropping correlated feature selection feature selection and fourteen classification algorithms results are presented in Table 5. The correlation factor of 0.25 is used through which 221 features were dropped. It is observed from the results that random forest classifier has produced better results with an accuracy of 80.22%, precision of 79.12%, recall of 79.12%, and F1-score of 79.12%. Next to the random forest is maximum voting and XGBoost classifier giving an accuracy of 72.83%, precision 76.92%, recall 76.92%, and f1-score 76.92%. The KNN classifier has produced poor results among all with an accuracy of 58.05%, precision 60.73%, recall 60.73%, and f1-score 60.73%.

Corresponding to recursive feature elimination feature selection and fourteen classification algorithms results are presented in Table 6. It is observed from the results that random forest classifier has produced better results with an accuracy of 74.72%, precision of 66.63%, recall of 66.63%, and F1-score of 66.63%. Random forest is a collection of decision tree methods. Each decision tree constructs a classifier based on a random data sample, and several classifiers are integrated to generate a single classifier known as random forest.

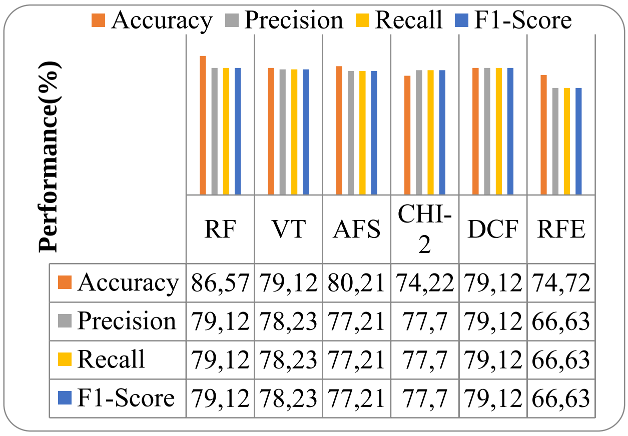
Table 6: Results based on Recursive Feature Elimination Feature Selection Technique

Algorithm	Accuracy	F-Score	Recall	Precision
Naive Bayes	67.77	57.14	57.14	57.14
Decision Tree	68.61	54.94	54.94	54..94
KNN	59.16	53.80	53.80.	53.84
SVM	55.55	49.45	4945	49.45
RF	**74.72**	**66.63**	**66.63**	**66.63**
Bagging	66.38	57.14	57.14	57.14
Ada Boost	66.94	58.24	58.24	58.24
Averaging	74.07	64.69	64.69	64.69
XGBoost	72.77	63.73	63.73	63.73
Light GBM	67.50	58.94	58.94	58.94
Max Voting	72.77	64.63	64.63	64.63
LR	66.38	56.74	56.72	56.74
Stacking	72.77	64.63	64.63	64.63
ID3	64.32	57.14	57.14	57.14

Table 5: Results based on Drop Correlated feature Selection Technique

Classification Algorithm	Accuracy	F1-Score	Recall	Precision
Naive Bayes	67.16	71.42	71.42	71.42
Decision Tree	62.77	61.53	61.53	61.53
KNN	58.05	60.73	60.73	60.73
SVM	66.72	71.42	71.42	71.42
Random Forest	**80.22**	**79.12**	**79.12**	**79.12**
Bagging	66.31	75.82	75.82	75.82
Ada Boost	62.22	61.53	61.53	61.53
Averaging	72.10	75.65	75.65	75.65
XGBoost	72.83	76.92	76.92	76.92
Light GBM	68.37	71.11	71.11	71.11
Max Voting	72.83	76.92	76..92	76.92
Logistic Regression	64.11	69.23	69.23	69.23
Stacking	70.83	76.92	76.92	76.92
ID3	67.77	71.54	71.54	71.54

Comparison of different approaches reported in literature on the same dataset are compared with our approach and presented in Table. The results showed that the methodology designed in this work performed around 3% better. It is found from the results presented in Tables 1 to 6 that random forest classifier has performed better in comparison with fourteen classifiers concerning six feature selection techniques. Random forest is a robust classifier since it produces its result by combining the outputs of many decision trees (created by dividing the training data into subsets). Furthermore, the random forest feature extraction technique assisted us in selecting features from a large category with high weightage in categorising the cardiac arrhythmia. Figure 2 gives a comparison of six different feature selection techniques used in our study. The table provides the performance parameters reported in each of the feature selection techniques. It has been observed that the random forest feature selection technique is performing better as compared to all other feature selection techniques applied.



	RF	VT	AFS	CHI-2	DCF	RFE
■ Accuracy	86,57	79,12	80,21	74,22	79,12	74,72
■ Precision	79,12	78,23	77,21	77,7	79,12	66,63
■ Recall	79,12	78,23	77,21	77,7	79,12	66,63
■ F1-Score	79,12	78,23	77,21	77,7	79,12	66,63

Figure 2: Comparison of performance concerning feature selection techniques

Kruskal-Wallis Test: This is a statistical test. H_0 (Null Hypothesis): There is no significant difference between the performances of models. Alpha taken is 0.05 and from the p-value observed from the test, which is less than alpha, so we can say that the null hypothesis is rejected. This test is performed with the scipy package in python. Table 8 shows the p-value and test statistic observed for the accuracy variable.

Table 8: Kruskal-Wallis Test Statistics

Test Statistic	Variable	P-Value
59.571	Accuracy	p=0.000000062666

6. Conclusion

The paper presents a machine learning framework for the detection of a multiclass arrhythmia. Firstly dataset is preprocessed to make it suitable for further processing, and then six feature selections were applied to extract the essential features from the dataset having 279 attributes. Fourteen classification algorithms are applied, including some of the ensemble techniques also to detect the presence of cardiac arrhythmia and classify them into different sixteen categories. All the fourteen classification algorithms were compared based on accuracy precision, recall, and f-score. It is observed that the best performance is reported in random forest feature selection with random forest classification algorithms from all the combinations of six feature selection techniques and fourteen classifiers. The highest performance achieved with this combination is the accuracy of 86.57%, 79.12%. It is also observed that a random forest classifier is performing better among all fourteen classifiers in all the six feature selection techniques applied. Thus it can be concluded from this study that random forest is the best performer classifier in this dataset, and along with Random forest classifier, it performs better among all the combinations of feature selection and classifiers applied.

Table 7: Comparison with Existing approaches

Paper	Accuracy (%)	Classification Algorithms
A. Batra and V. Jawa[15]	84%	SVM
E.Namsrai et al. [20]	70.50%	NB
K. A., K.Niazi e. al. [21]	73.80%	KNN
D. Gao et al. [19]	76.67%	ANN
S.M. Jadhavet. al. [22]	78.89%	Modular NN
Vasu Gupta et. al.[5]	77.4%	RF+SVM
Our Approach	86.57%	Random Forest Feature Selection + Random Forest Classifier

References

[1] J Zuo, W. M., Lu, W. G., Wang, K. Q., & Zhang, H. (2008, September). Diagnosis of cardiac arrhythmia using kernel difference weighted KNN classifier. In *2008 Computers in Cardiology* (pp. 253-256).IEEE.

[2] Alickovic, E., &Subasi, A. (2016).Medical decision support system for diagnosis of heart arrhythmia using DWT and random forest classifier. *Journal of medical systems*, *40*(4), 108.

[3] Kumar, S. U., &Inbarani, H. H. (2017). Neighborhood rough set based ECG signal classification for diagnosis of cardiac diseases. *Soft Computing*, *21*(16), 4721-4733.

[4] Özbay, Y., &Karlik, B. (2001). *A recognition of ECG arrhythmias using artificial neural networks*. SELUK UNIV KONYA (TURKEY) ELECTRICAL AND ELECTRONICS ENGINEERING.

[5] Gupta, V., Srinivasan, S., & Kudli, S. S. (2014). Prediction and Classification of Cardiac Arrhythmia.

[6] A. M. Elsayad, "Classification of ECG arrhythmia using learning vector quantization neural networks," in *Proceedings of the 2009International Conference on Computer Engineering and Systems, ICCES'09*, pp. 139–144, egy, December 2009.

[7] Raut, R. D., &Dudul, S. V. (2008, July).Arrhythmias classification with MLP neural network and statistical analysis. In *2008 First International Conference on Emerging Trends in Engineering and Technology* (pp. 553-558). IEEE.

[8] Yu, S. N., & Chen, Y. H. (2007). Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. *Pattern Recognition Letters*, *28*(10), 1142-1150.

[9] Hussain, H., &Fatt, L. L. (2007, December). Efficient ECG signal classification using a sparsely connected radial basis function neural network. In *Proceeding of the 6th WSEAS International Conference on Circuits, Systems, Electronics, Control, and Signal Processing* (pp. 412-416).

[10] Bhardwaj, P., Choudhary, R. R., &Dayama, R. (2012). Analysis and classification of cardiac arrhythmia using ECG signals. *International Journal of Computer Applications*, *38*(1), 37-40.

[11] S. M. Jadhav, S. L. Nalbalwar, and A. A. Ghatol, "Arrhythmia disease classification using Artificial Neural

Network model,” in Proceedings of the 2010 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2010, pp. 653–656, India, December 2010.

[12] M. Mitra and R. Samanta, "Cardiac Arrhythmia Classification Using Neural Networks with Selected Features, "Procedia Technology, vol. 10, pp. 76–84, 2013.

[13] A.Ozc¸ift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis, "*Computers in Biology and Medicine*, vol. 41, no. 5, pp. 265–271,2011.

[14] A. Batra and V. Jawa, “Classification of Arrhythmia Using ConjunctionofMachine Learning Algorithms and ECG DiagnosticCriteria,” *Training Journal*, 1975.

[15] T. Soman and P. O. Bobbie, "Classification of arrhythmia using machine learning techniques," *WSEAS Transactions on Computers*, vol. 4, no. 6, pp. 548–552, 2005.

[16] A. Fazel, F. Algharbi, and B. Haider, *Classification of CardiacArrhythmias Patients*, Haider B Classification of CardiacArrhythmias Patients.

[17] S. Samad, S. A. Khan, A. Haq, and A. Riaz, "Classification of arrhythmia," *International Journal of Electrical Energy*, vol. 2, no.1, pp. 57–61, 2014.

[18] N. Kohli and N. Verma, "Arrhythmia classification using SVM with selected features," *International Journal of Engineering, Science and Technology*, vol. 3, no. 8, pp. 22–31, 2012.

[19] D. Gao, M. Madden, D. Chambers, and G. Lyons, "Bayesian ANN Classifier for ECG arrhythmia diagnostic system: A comparison study," in Proceedings of the International Joint Conference on Neural Networks, IJCNN 2005, pp. 2383–2388, Canada, August 2005.

[20] E.Namsrai, T.Munkhdalai, M. Li, J. Shin, O.Namsrai, and K.H.Ryu, "A feature selection-based ensemble method for arrhythmia classification," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 31–40, 2013.

[21] K. A. K.Niazi, S. A. Khan, A. Shaukat, and. Akhtar, "Identifyingbest feature subset for cardiac arrhythmia classification," in *Proceedings of the Science and Information Conference, SAI 2015*, pp. 494–499, UK, July 2015.

[22] S.M. Jadhav, S. L.Nalbalwar, and A. A. Ghatol, "Modular neural network-based arrhythmia classification system using ECG signal data," International Journal of Information Technology and Knowledge Management1, vol. 4, no. 1, pp. 205–209, 2011.

[23] https://archive.ics.uci.edu/ml/datasets/arrhythmia

[24] Díaz-Uriarte, R., & De Andres, S. A. (2006).Gene selection and classification of microarray data using the random forest. *BMC bioinformatics*, *7*(1), 3.

[25] Ding, H., Feng, P. M., Chen, W., & Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Molecular BioSystems*, *10*(8), 2229-2235.

[26] Demir, Ö.,&YılmazÇamurcu, A. (2015). Computer-aided detection of lung nodules using exterior surface features. *Bio-medical materials and engineering*, *26*(s1), S1213-S1222.

[27] Koller, D., &Sahami, M. (1996). *Toward optimal feature selection*. Stanford InfoLab.

[28] Granitto, P. M., Furlanello, C., Biasioli, F., &Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, *83*(2), 83-90.

[29] Jin, X., Xu, A., Bie, R., &Guo, P. (2006, April). Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In *International Workshop on Data Mining for Biomedical Applications* (pp. 106-115).Springer, Berlin, Heidelberg.

[30] Schölkopf, B. (2005). Introduction to Kernel Methods. *The Analysis of Patterns, Erice, Italy*.

[31] Vincent, P., &Bengio, Y. (2002).K-local hyperplane and convex distance nearest neighbor algorithms.In *Advances in neural information processing systems* (pp. 985-992).

[32] Liaw, A., & Wiener, M. (2002).Classification and regression by randomForest. *R news*, *2*(3), 18-22.

[33] Soucy, P., &Mineau, G. W. (2001).A simple KNN algorithm for text categorization.In *Proceedings 2001 IEEE International Conference on Data Mining* (pp. 647-648).IEEE.

[34] Tabaei, B., and Herman, W., A Multivariate logistic regression equation to screen for diabetes. Diabetes Care 25:1999–2003, 2002.

[35] Freund, Y., & Mason, L. (1999, June). The alternating decision tree learning algorithm. In *icml* (Vol. 99, pp. 124-133).

[36] Sanders, S. R., Noworolski, J. M., Liu, X. Z., &Verghese, G. C. (1990). *Generalized averaging method for power-conversion circuits. Technical report* (No. AD-A-221977/2/XAB; LIDS-P--1970).Massachusetts Inst. of Tech., Cambridge, MA (USA).Lab. for Information and Decision Systems.

[37] Dietterich, T. G. (2000, June). Ensemble methods in machine learning.*International workshop on multiple classifier systems* (pp. 1-15).Springer, Berlin, Heidelberg.

[38] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ...& Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree.In *Advances in Neural Information Processing Systems* (pp. 3146-3154).

[39] Korada, N. K., Kumar, N. S. P., &Deekshitulu, Y. V. N. H. (2012). Implementation of naïve Bayesian classifier and ADA-boost algorithm using maize expert system. *International Journal of Information Sciences and Techniques (IJIST)*, *2*.

[40] Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Update*, *1*(1), 2007.

[41] Naess, O. E. (1979). Superstack—an iterative stacking algorithm. *Geophysical Prospecting*, *27*(1), 16-28.

[42] Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1-39.

[43] Umanol, M., Okamoto, H., Hatono, I., Tamura, H. I. R. O. Y. U. K. I., Kawachi, F., Umedzu, S., & Kinoshita, J. (1994, June). Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems. In *Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference* (pp. 2113-2118). IEEE.

[44] Ruta, D., &Gabrys, B. (2005).Classifier selection for majority voting. *Information fusion*, *6*(1), 63-81.

[45] Mukhopadhyay, S., &Sircar, P. (1996).Parametric modeling of ECG signal. *Medical and Biological Engineering and Computing*, *34*(2), 171-174.

[46] Mustaqeem, A., Anwar, S. M., &Majid, M. (2018). Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants. *Computational and mathematical methods in medicine*, *2018*.