Classification of Cardiac Arrhythmia using improved Feature Selection methods and Ensemble Classifiers

Rajat Jain, Pranam R Betrabet, B Ashwath Rao, N V Subba Reddy

Department of Computer Science & Engg., Manipal Institute of Technology, Manipal, Karnataka State, India, PIN-576104.

E-mail: rajat.jain@learner.manipal.edu

Abstract. Arrhythmia is one of the life-threatening heart diseases which is diagnosed and analyzed using electrocardiogram (ECG) recordings and other symptoms namely rapid heartbeat or chest-pounding, shortness of breath, near fainting spells, insufficient pumping of blood from the heart, etc along with sudden cardiac arrest. Arrhythmia records a hasty and aberrant ECG. In this implementation, the arrhythmia dataset is collected from the UCI machine learning repository and then classified the records into sixteen stated classes using multiclass classification. The large feature set of the dataset is reduced using improved feature selection techniques such as t-Distributed Stochastic Neighbor Embedding (TSNE), Principal Component Analysis (PCA), Uniform Manifold Approximation, and Projection (UMAP) and then an Ensemble Classifier is built to analyse the classification accuracy on arrhythmia dataset to conclude when and which approach gives optimal results.

1. Introduction

Today, when the entire world is fighting the pandemic, it is seen that heart diseases have become more common, especially during the second wave of Covid-19. Heart diseases are in light for a long, especially Arrhythmia. It is one of the life-threatening heart diseases which causes serious health problems in patients when left untreated. Accurate treatment along with early detection plays a lifesaving role when it comes to heart diseases. Generally, an electrocardiogram (ECG) is the universal tool to diagnose and study heart functioning. This is done by attaching electrodes on the body of patients which generates his hearth’s graphical scheme of the electrical impulses. Normally, ECG signals consist of the P waves, QRS complex, and T waves. Shape, time span, and the correlation between QRS complex, T wave, P wave, and R-R interval. Any hasty fluctuation in this specification designates that the heart is having an infirmity that may have various reasons behind it [1].

Talking about Arrhythmia, it may not be known to common people but it is a very well-known and understood jargon in the field of medical sciences, specifically cardiology. It is a contortion in heart rhythms that may result in some heart disease whose severity totally depends upon how early it is being detected and how accurate the treatment is. Mismatch of any may lead to severe heart disease and a serious threat to patient’s life. In Arrhythmia, the plain sailing rhythm of an electrical system of the heart gets disturbed and directs the heart to pump too fast or too slow i.e., to skip or to race the beats causing inconsequent gestures of heart signals. Mostly, Arrhythmia is diagnosed and analyzed using electrocardiogram recordings and other symptoms namely rapid heartbeat or chest-pounding, shortness
of breath, near fainting spells, insufficient blood pumping of blood from heart, etc along with sudden cardiac arrest in the worst cases. Thus, arrhythmia reveals an abrupt and off-center ECG signal. Broadly stating, arrhythmia is divided into two categories namely tachycardia and bradycardia.

In tachycardia, the heartbeat shoots up to 100 beats per minute (bpm) and in bradycardia heartbeat usually drops down below the rate of 60 bpm [2]. Both the conditions are risky and life-threatening if not treated in time. Thus, one can easily sense how important it is to get it diagnosed and treated promptly to shrink the risk of sudden death.

With the emergence of a large number of remote healthcare systems for patients suffering from cardiac diseases, the urgency of a robust, systematic, and intelligent system for classifying arrhythmia is being predominately addressed and appreciated. Aiming to develop such a robust and precise diagnosing system, in order to obtain an inch-perfect arrhythmia classification from ECG signals various techniques of machine learning were adopted previously [3]. To select an appropriate and well-suited technique to refine the arrhythmia classification accuracy, one has to strictly consider the requirements of specified patients, the backdrop of the application, earlier experiences, and analysis of data making it a complex task altogether.

In this paper, we aim at stockpiling the dataset on cardiac arrhythmia disease from UCI, a Machine Learning (ML) repository, fabricated in 1987 by some graduate students and their professor David Aha at UC Irvine. After collecting the dataset, this implementation classified the records into sixteen stated classes using multiclass classification. The large feature set of the dataset is reduced using improved feature selection techniques such as t-Distributed Stochastic Neighbor Embedding (tSNE), Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) and then using Ensemble Classifiers analyze the classification accuracy on arrhythmia dataset.

2. Literature Survey

Sensing the urgency of ways to accurately classify the arrhythmia precisely in different classes, many proposals were introduced and worked upon in the past with an aim to evolve a sharp classification model for the detection of arrhythmia. To categorize cardiac arrhythmia patients, a neural network based on learning vector quantization has been pertained on the electrocardiogram dataset. In order to categorize instants having arrhythmia or as normal, PCA is used heed neural networks, being six in number, in order to reduce the features [4].

To categorize the patients of arrhythmia using pendant twelve lead electrocardiogram data recording, research is conducted, applying artificial neural networks (automated). If there is some missing data, it is being figured out by restoring the values of attributes with the nearest attribute values from the referred class. After doing this, for arrhythmia classification, a multilayer perceptron was used with the static backpropagation method [5]. Generated feed-forward neural network is also used for doing the similar work [6]. Along with this, some other approaches like Bayesian artificial neural networks [7], modular neural networks [8] and perceptron having multiple layers with one-against-all method [9] are also used in a similar fashion for categorizing cardiac arrhythmia into sixteen different classes. Apart from these, a contemporary approach for categorizing cardiac arrhythmia is lodged in [10] whose operating principle revolves around picking up the most appropriate characteristics from the UCI electrocardiogram data by using correlation-based feature selection technique. Levenberg-Marquardt is waged for a prior and precise arrhythmia along with incremental backpropagation neural network.

Also, to outline a computer-aided diagnostic system, decision trees are used for categorization of cardiac arrhythmia successfully. The use of this design can be of huge advantage at the end of hospitals as these decision and diagnosis support systems can turn out to be a helping hand to physicians for disease diagnosis efficiently and accurately. Furthermore, to refine the classification system for the diagnosis of arrhythmia, a random forest ensemble method has been proposed which is based on a resampling strategy [11].

Various ML algorithms like gradient boosting, neural networks, random forests, decision trees, and support vector machines are used, after applying feature selection strategies and rigorous pre-processing on ECG data, for arrhythmia classification [12]. Likewise, medical ECG dataset has been treated by
algorithms like J48, OneR, Naïve Bayes, support vector machine, K-nearest neighbor, random forest, logistic regression, and decision trees to categorize arrhythmia into sixteen different classes [13]. Noteworthy work is conducted on ECG data furnishing support vector machine-based methods for arrhythmia diagnosis and using PCA for the selection of significant features [14]. An efficient model is proposed which uses k-nearest neighbour and a support vector machine for classifying arrhythmia patients in which a refined precision estimate is accomplished using an amalgam of sequential forward search and F-score for feature selection [15]. In data mining, pattern recognition, and machine learning, feature selection techniques prove to be a war winning weapon, especially against large datasets. On unsupervised data only, in [16] a feature selection procedure is proposed which is constructed over dynamic mutual information. Various experiments are being conducted on the UCI ECG dataset in order to validate the coherence and productivity of the method using typically four algorithms. In [17], a novel ensemble-based technique is introduced for choosing the features from the electrocardiogram dataset that are relevant by selecting the subset of features from the ECG dataset, and on each selected subset, classifiers are being trained. The classifiers used namely support vector machines, decision trees and naive bayes. In all the above-mentioned works, the choice of the most relevant features remains a heavy stone to move and the accuracy of classification has room left for further refinement and improvement. The censorious nature of disease diagnosis and prognosis calls for a highly accurate, efficient, and robust system, which is to be deployed for supporting decision-making clinically. In this paper, our put forth approach focuses upon filling these gaps.

3. Methodology

In this implementation we look forward to a fresh approach of classifying the UCI ECG dataset and finding the training and tests score, basis on which we would be in a position to compare the results obtained using four different way outs, that is described in the later part of this section, and come up with a conclusion which would be presented in the later parts of this paper. Broadly speaking, our implementation design consists of three phases or stages:

- Data collection phase
- Data pre-processing phase
- Data classification phase

3.1 Data collection phase

This is the first phase which is generally not being focused properly, but we need to be aware of the sensitivity and importance of inch-perfect data especially when we are dealing with medical science computations. More the data is precise, more accurate results can be obtained which would play a game-changer role in saving lives of the patients or even reducing the pain and duration of treatment in general cases by assuring early and accurate diagnosis. For our implementation, we have also selected the same data set on which the above-described models, described in section 2 of this paper, put forth their work i.e., dataset of arrhythmia collected from UCI ML repository. Below is a snipped snapshot of the used dataset just for the eyes of the reader.
3.2 Data pre-processing phase
This is the second phase in the queue of our implementation. It is being well known and observed that features with a smaller numeric range will have less bump on the classification accuracy than the features having large numeral values. Features will large numeric values will be more precise. Attributes used in this implementation have a wide numeral range. In such a scenario, as stated above, features or attribute values with a high numeric range will have a greater impact on accuracy than rest of the features. In order to surpass this impact of the response variable for these attributes or features data normalization is being performed [18].

The premier grail of this technique is to refine the performance by restricting the footprints of huge data value features on the model. For normalizing the data, scaling and centring approach is embraced for the real values of the dataset, which refines the numeral stability of the put-forth classification system. To limit the mean value of features, centring transformation is applied, so that it turns zero. For an observation set s, such that \( s = \{s_1, s_2, s_3, \ldots, s_n\} \), we obtained a fresh set of transformed observations after applying centering transformation.it is given as

\[ \text{Cen}_s = \{s_i - \mu(s)\} \]

where \( \text{cen}_s = \{\text{cen}_{s_1}, \text{cen}_{s_2}, \text{cen}_{s_3}, \ldots, \text{cen}_{s_n}\} \), where total observations is denoted by n and the final mean \( \mu(\text{cen}_s) \) equates to 0. Once centering is done, in order to ensure that cen, has a standard deviation (SD) equals one, a scaling transformation is applied. This is done by dividing all the observations in cen, by the SD of all samples in set s. It is given by

\[ \text{sc}_{s_i} = (\text{cen}_{s_i} / \sigma(s)) \]

such that \( \sigma(\text{sc}_s) \) equals one.

3.3 Data classification phase
This phase is the heart of implementation. It is the third phase in the queue which comes after the completion of the data pre-processing phase. This phase focuses on the classification of arrhythmia data into different class labels. Firstly, our dataset is bifurcated into training data and testing data. In our approach, we have considered five classifiers namely, linear SVM, logistic regression, decision tree, random forest and kernelized SVM. We will train the above mentioned five classifiers on the training data and combine these five classifiers to build an ensemble classifier. This ensembled classifier takes the majority class label from the class labels produced by the five classifiers independently. Now using this ensembled classifier, we will calculate the training score and testing score using for different approaches mentioned below.
3.3.1 First Approach: Without Feature Selection
In our first approach, we will build the ensembled classifier, by majority vote, without performing any feature selection process on the dataset and check the training and testing data accuracy and record in form of training and testing scores for comparison.

3.3.2 Second Approach: PCA Feature Selection
Our second approach focuses on performing the feature selection process on the arrhythmia dataset by using PCA technique, choosing 100 principal components. Now we will again build the ensemble classifier on this processed data, processed using PCA technique, and check the training and testing data accuracy and record in form of training and testing score for comparison. Below Figure 2 depicts the scatter of the data points in two-dimensional space after apply PCA feature selection.

![Figure 2. Scatter plot after PCA.](image)

3.3.3 Third Approach: TSNE Feature Selection
In our third approach, we will perform feature selection on the dataset using the t-Distributed Stochastic Neighbor Embedding (TSNE) technique, reducing the dimensionality to two dimensions. Now, following the similar process as in the above two approaches, we will again build the ensemble classifier on this processed data and check the training and testing data accuracy, and record in form of training and testing score for comparison. Below Figure 3 depicts the scatter of the data points in two-dimensional space after apply TNSE feature selection.

![Figure 3. Scatter plot after TSNE.](image)
In our fourth and final approach, we will perform feature selection on the dataset using Uniform Manifold Approximation and Projection (UMAP) technique, reducing the dimensionality to two dimensions and again build the ensemble classifier on this processed data and record the training and testing score. Below Figure 4 depicts the scatter of the data points in two-dimensional space after apply UMAP feature selection.

Below Figure 5 represents the overall process flow of this implementation in form of a block diagram, starting from the collection of arrhythmia dataset to all the way to derive results and have comparison.
4. Results

After completion of all the three phases mentioned in section 3 of this paper, we have with us the training data and testing data accuracy in form of training and testing scores which are obtained using four different approaches mentioned in the 3.3 section of this paper i.e., training and test scores obtained from ensembled classifier along with the training and test scores obtained from individual classifiers. Below we present the results obtained, in form of training and testing scores, compiled in form of a result Table 1 and presented graphically in Figure 6 along with the cross-validation results, with number of folds equal to 5, depicted in Table 2 and presented graphically in Figure 7.

Table 1. Performance comparison of feature selection methods

| Score Category | Without Feature Selection | With PCA Feature Selection | With TSNE Feature Selection | With UMAP Feature Selection |
|----------------|---------------------------|----------------------------|----------------------------|-----------------------------|
| Train Score    | 0.96                      | 0.92                       | 0.60                       | 0.56                        |
| Test Score     | 0.69                      | 0.65                       | 0.59                       | 0.60                        |

Table 2. Results of cross validation

| Score Category | Without Feature Selection | With PCA Feature Selection | With TSNE Feature Selection | With UMAP Feature Selection |
|----------------|---------------------------|----------------------------|----------------------------|-----------------------------|
| Mean Score     | 0.72                      | 0.71                       | 0.59                       | 0.55                        |
| Mean SD Score  | 0.03                      | 0.03                       | 0.02                       | 0.01                        |

Below presented is the graphical representation of the contents of Table 1 and Table 2 in form of a graph for a better pictorial view and ease of comparison.
Figure 6. Performance comparison of feature selection methods.

Figure 7. Results of cross validation.

Below, Figure 8 and Figure 9 are the snippets from our implementation that consolidatedly compares the test and train recall scores of individual classifiers without and with PCA.
| Model                                      | Train Recall Score | Test Recall Score |
|--------------------------------------------|--------------------|-------------------|
| KNN Classification                         | 0.687500           | 0.647059          |
| Logistic Regression                        | 0.841146           | 0.676471          |
| Linear SVM                                 | 0.934896           | 0.705882          |
| Kernelized SVM                             | 0.976562           | 0.676471          |
| Naive Bayes                                | 0.687500           | 0.632353          |
| Decision Tree                              | 0.846354           | 0.720588          |
| Random Forest                              | 0.992188           | 0.750000          |
| KNN Classification with PCA                | 0.695312           | 0.632353          |
| Logistic Regression with PCA               | 0.825521           | 0.676471          |
| Linear SVM with PCA                        | 0.916667           | 0.705882          |
| Kernelized SVM with PCA                    | 0.968750           | 0.676471          |
| Decision Trees with PCA                    | 0.617188           | 0.617647          |
| Random Forest with PCA                     | 0.963542           | 0.647059          |

**Figure 8.** Individual comparison of test and train recall scores without and with PCA.

**Figure 9.** Graphical representation of individual comparison of test and train recall scores without and with PCA.
5. Conclusion and Future Scope

Comparing the test and train scores of individual classifiers, we could conclude that random forest without feature selection produced the best scores while KNN and Naïve Bayes produced the least training and test scores. Thus, when it comes to an individual classifier, one can choose to go with random forest without feature selection when inch-perfect accuracy is the primary focus while one can opt for linear SVM with PCA feature selection approach if reducing the computational cost is to be taken care of.

Analyzing the future scope, one can build the classification system using various deep learning models like Artificial Neural Network (ANN) or one can build an ensemble classifier taking some different combination of individual classifiers as in this implementation we could observe that individual performance of random forest classifier, in terms of accuracy, is better than the build ensembled classifier. This may be because in the ensembled classifier, the accuracy achieved by random forest is getting suppressed by using the other four classifiers. This can be improved by taking some other combination of classifiers with random forest.

References

[1] Alickovic, Emina, and Abdulhamit Subasi. "Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier." Journal of medical systems 40.4 (2016): 108.
[2] Shi, Haotian, et al. "Automated heartbeat classification based on deep neural network with multiple input layers." Knowledge-Based Systems 188 (2020): 105036.
[3] Al Rahhal, Mohamad Mahmoud, et al. "Deep learning approach for active classification of electrocardiogram signals." Information Sciences 345 (2016): 340-354.
[4] Mustaqeem, Anam, et al. "Wrapper method for feature selection to classify cardiac arrhythmia." 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017.
[5] Pandey, Saroj Kumar, and Rekh Ram Janghel. "ECG arrhythmia classification using artificial neural networks." Proceedings of 2nd International Conference on Communication, Computing and Networking. Springer, Singapore, 2019.
[6] Khan, Ali Haider, Muzammil Hussain, and Muhammad Kamran Malik. "Arrhythmia Classification Techniques Using Deep Neural Network." Complexity 2021 (2021).
[7] Ahmed, Aya F., Mohamed I. Owis, and Inas A. Yassine. "Novel Bayesian classifier discriminant function optimization strategies for arrhythmia classification." IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2014.
[8] Ochoa, Alberto, et al. "Noise-tolerant modular neural network system for classifying ECG signal." Informatica 43.1 (2019).
[9] Niazi, Khalid Ahmed Khan, et al. "Identifying best feature subset for cardiac arrhythmia classification." 2015 Science and Information Conference (SAI). IEEE, 2015.
[10] Mitra, Malay, and R. K. Samanta. "Cardiac arrhythmia classification using neural networks with selected features." Procedia Technology 10 (2013): 76-84.
[11] Mustaqeem, Anam, Syed Muhammad Anwar, and Muahammad Majid. "Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants." Computational and mathematical methods in medicine 2018 (2018).
[12] Goel, Supriya, Pradeep Tomar, and Gurjit Kaur. "A fuzzy based approach for denoising of ECG signal using wavelet transform." International Journal of Bio-Science and Bio-Technology 8.2 (2016): 143-156.
[13] Firyulina, M. A., and I. L. Kashirina. "Classification of cardiac arrhythmia using machine learning techniques." *Journal of Physics: Conference Series*. Vol. 1479. No. 1. IOP Publishing, 2020.

[14] Salem, Milad, Shayan Taheri, and Jiann–Shiun Yuan. "ECG arrhythmia classification using transfer learning from 2-dimensional deep CNN features." *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2018.

[15] Shimpi, Prajwal, et al. "A machine learning approach for the classification of cardiac arrhythmia." *2017 international conference on computing methodologies and communication (ICCMC)*. IEEE, 2017.

[16] Liu, Huawen, Yuchang Mo, and Jianmin Zhao. "Conditional Dynamic Mutual Information-Based Feature Selection." *Comput. Informatics* 31.6 (2012): 1193-1216.

[17] Namsrai, Erdenetuya, et al. "A feature selection-based ensemble method for arrhythmia classification." *Journal of Information Processing Systems* 9.1 (2013): 31-40.

[18] Price, E. Magda, and Wendy P. Robinson. "Adjusting for batch effects in DNA methylation microarray data, a lesson learned." *Frontiers in genetics* 9 (2018): 83.