International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

CARDIAC ARRHYTHMIA CLASSIFICATION USING SVM, KNN AND NAIVE BAYES ALGORITHMS

Raghavendra M Devadas

Assistant Professor, School of Engineering, Computer Science Department, Presidency University, Bangalore, India

---

Abstract - *) A normal human being heart beat is somewhere between 60 to 100 beats per minute, any deviation from normal heart beat is a medical condition referred as Cardiac Arrhythmia. The purpose of our work is to classify Arrhythmia into 16 different types using three machine learning algorithms viz SVM, KNN and Naïve Bayes. This work considers UCI Arrhythmia dataset which has huge number of features which are reduced using feature selection technique. This work shows improvement in accuracy of 9.9 % in SVM model, 3.3 % in KNN model and 24.2 % in Naïve Bayes model when we consider relevant features only. This study performs multiclass classification considering all the original features in the dataset and also leaving out the irrelevant features and compares the performance of three algorithms using Accuracy and Kappa scores. The results show comparative performance among SVM (accuracy 71.4 %) and Naïve Bayes (accuracy 70.3 %) and least accuracy of 62.6 % for KNN algorithm after applying feature selection.*

Key Words: Classification, KNN, Machine Learning, Naïve Bayes, SVM.

1.INTRODUCTION

A Cardiac Arrhythmia is defined as when the heart beat deviates from the normal rhythm. A normal heart beat is somewhere between 60 – 100 beats/minute. Whenever a heart beats below 60 beats or above 100 beats per minute (BPM) we refer as a heart is beating out of rhythm. Following are the requirements for a normal heart beat.

- Heart rate between 60-100BPM.
- It is very important that a heartbeat should originate from SA node only.
- Cardiac impulse should propagate through normal conduction pathway with normal velocity.

In case if any of the above said requirements is not satisfied, we can infer as Cardiac Arrhythmia. For example, if heart rate below 60 BPM we refer it as Bradycardia arrhythmia and if heart rate is above 100 BPM, it is called as Tachycardia arrhythmia. Following are some types of arrhythmias caused due to different reasons, Sinus arrhythmia, Atrial arrhythmia, Junctional arrhythmia, Ventricular arrhythmia etc. Electrocardiogram (ECG) is a tool used to visualize the electricity that flows through the heart. It uses flat metal electrodes placed on person's chest to notice the electrical charges generated by the heart as it beats, which are then graphed. Doctors can analyze the patterns generated in graph to get an understanding of heart rhythm and evaluate cardiac health.

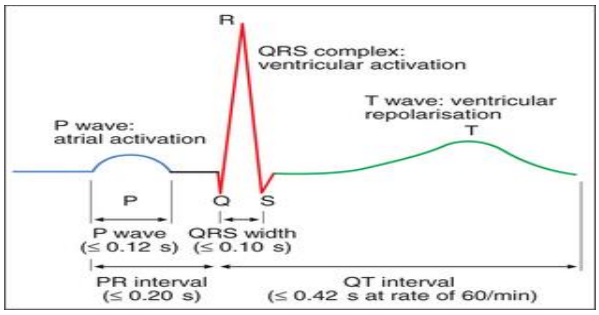

Fig. 1. Components of ECG graph

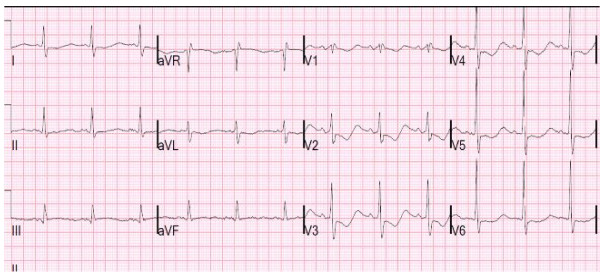

Fig. 2. Example of Abnormal ECG

Due to high mortality rate of heart diseases [1], early detection and precise discrimination of ECG signal is essential for the treatment of patients. In past many researchers have worked on data mining techniques to predict and classify diseases like diabetes, heart disease [2-3].

In recent years machine learning techniques have proved its ability to correctly classify and predict results. Hence, this work uses machine learning algorithms to classify arrhythmia types which can give vital information to Cardiologist to confirm diagnosis. This work uses SVM, KNN and Naïve Bayes algorithms. The aim of this study is to compare the efficiency of these algorithms using Accuracy and Kappa scores for evaluation in classifying arrhythmia

types. This paper is divided into following sections, Section II discusses related work done over arrhythmia dataset and classification models. Section III Theoretical background of algorithms. Section IV will show experimentation and results. Finally, conclusion is provided in section V

2. RELATED WORK

T. Soman and P. O. Bobbie [4], have applied Naïve Bayes, J48 and OneR algorithms for classifying arrhythmia using ECG dataset and have evaluated the performance of the algorithms and they have found that OneR and Naïve Bayes show the stable accuracy rate, this is not true for J48 algorithm.

Namrata Singh and Pradeep Singh [5], authors present model for diagnosis of cardiac arrhythmias. They have used three types of machine learning methods, namely, linear SVM, random forest, and JRip, and analyzed the performance of the feature selection methods. Their experimental results show that highest accuracy of 85.58% was obtained with random forest classifier using gain ratio feature selection method with a subset of 30 features.

Vasu Gupta, Sharan Srinivasan, Sneha S Kudli [6], they have implemented neural networks, random forest, svm and naïve bayes on arrhythmia dataset. Also, they propose novel approach by combining random forest and linear kernel svm and showed a classification error of 77.4%.

Batra, A., Jawa, V [7], authors have worked using fusion of machine learning methods and ECG diagnostic criteria which improved the accuracy of detecting arrhythmia disease using electrocardiogram (ECG) data. They evaluated classification performance using parameters such as confusion matrix, kappa score, confidence interval, area under ROC curve (AUC), and overall accuracy. Authors neural networks, decision trees, random forest, gradient boosting, and support vector machines were applied. Combining SVM and Gradient Boosting show 84.82% overall accuracy was achieved.

Yeniterzi, S., Yeniterzi, R., Kücükural, A., Sezerman, U [8], they have used genetic algorithms and found that classification accuracy that was achieved from 278 features, only 7 features is enough to get the same classification accuracy. Different features they used are linear existence, of ragged R wave, and other five measurements that were taken from ECG. GA's can be used very efficiently in combination with SVMs to find relatively important features in cardiac arrhythmia database.

Coast, D.A., Stern, R.M., Cano, G.G., Briller, S.A [9], propose novel approach to ECG cardiac arrhythmia analysis using Hidden Markov Models (HMM). Their approach is a blend of structural and statistical knowledge of the ECG signal. They found that QRS complexes and R-R intervals can used to predict ventricular arrhythmias.

Jadhav, S.M., Nalbalwar, S.L., Ghatol [10], they classify arrhythmia into normal and abnormal groups using Modular neural network (MNN), 82.22% was the classification accuracy. Six measures were used to evaluate the classification performance.

Luz, E.J.D.S., Schwartz, W.R., Cámara-Chávez, G., Menotti D [11], have done a comprehensive review of different methods of classifying heart beat abnormalities using the heartbeat segmentation methods, ECG signal processing, data mining algorithms and feature description techniques.

3. MACHINE LEARNING CLASSIFICATION MODELS

3.1 Support Vector Machines (SVM)

The classification and regression problems can be effectively solved using SVM which is supervised machine leaning algorithm. Suppose there are two features x1 and x2, with x1 be a set all square boxes and let x2 be as set with all circles plotted on 2-dimensional coordinate system. The goal is to design a hyperplane that classifies all the training data in two classes. The best choice will be the hyperplane that leaves the maximum margin from both classes. When the number of training data is small, SVMs outperform conventional classifiers [12].

3.2 K- Nearest Neighbors (KNN)

KNN is a fundamental machine learning classification and regression technique. KNN are been used in data mining in many application fields [13]. The algorithm works as follows. A case is separated by a greater vote of its neighbors, with the case being allocated to the class greatest mutual amongst its K nearest neighbor calculated by a distance function. If K = 1, the instance is simply allocated to the class of its nearest neighbor [14].

3.3 Naïve Bayes (NB)

NB works on Bayes formula to envisage the class of indefinite data sets. It is very easy to construct and modify large datasets by using Naïve Bayes model [15]. It considers the assumption that features of a measurement are autonomous of each other. It takes each feature distinctly

and identify proportion of preceding measurements that fit to class A that have the equivalent value for this feature.

4. EXPERIMENTATION AND RESULTS DISCUSSION

This work uses UCI Arrhythmia dataset [16] and table 1 provides the details and Fig 3 shows the flow chart of our proposed method.

Table. 1. Characteristics of dataset

Number of Attributes	280
Number of Instances	452
Missing Values	Present
Attribute Characteristics	Categorical, Integer, Real
Outcome attribute	1 to 16 values, with 1 as normal and 2-16 refers to various arrhythmia types

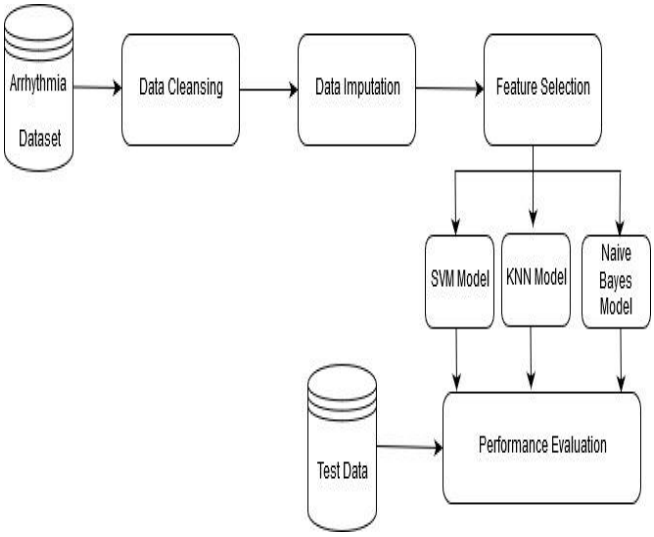

Fig. 3. Proposed system of Classification

4.1 Data Cleansing

- Identify features with just one value and eliminate them.
- Eliminate features which has all same values.
- Identify features with missing data, and eliminate if it has greater than 20 values missing and two features were eliminated, hence our total features are now 278 as compared to 280 in original dataset.

4.2 Data Imputation

Data Imputation is a process of replacing missing values using predicted values. In general, many methods exist viz, Mice, Amelia, missForest, Hmisc and mi which is provided by R in its CRAN repository. This word uses Mice package, and mice works in the following way. Suppose a feature has missing values mice takes the regression of other features of dataset, the missing feature values will be replaced by the predicted values. Predictive Mean Matching (PMM) method is used in mice package. Totally five imputation were predicted. One value which is closest to the concerned feature was imputed in the missing feature column. Missing values were imputed in 3 features. The performance of mice imputation is shown in Fig 4, red spot are the predicted values computed from the existing values which are in blue color.

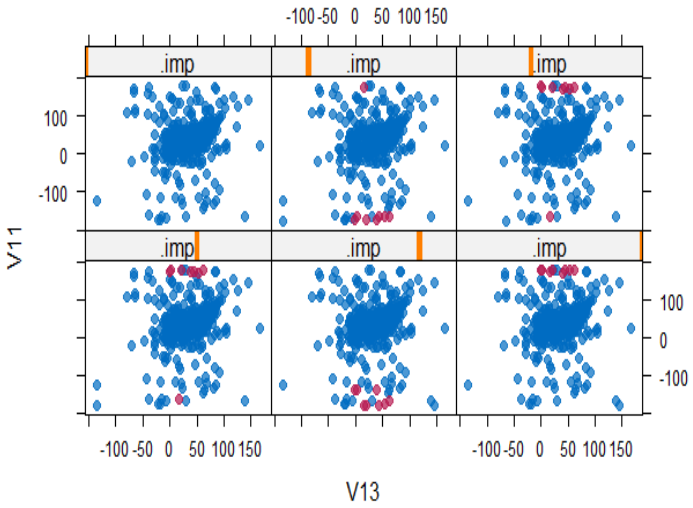

Fig. 4. Predicted values using Mice

4.3 Feature Selection

Since the dataset has huge number of features it is important to consider only the relevant once and drop other features which does not contribute much in getting the machine learning models accurate. This study employed Boruta feature selection technique as it is mostly applicable when we have data set made of several features. Boruta algorithm uses a wrapper approach built around a random forest [17]. After applying Boruta technique for feature selection, got total 88 features out of 278 features, i.e. only 31.6 % of total features are selected to build the machine learning models. Fig 5. describes the feature selected.

```
 [1] "V1"   "V5"   "V7"   "V8"   "V10"  "V11"  "V13"  "V15"  "V17"  "V18"  "V28"  "V30
"  "V33"
[14] "V65"  "V69"  "V76"  "V81"  "V88"  "V90"  "V91"  "V93"  "V100" "V102" "V103" "V10
5"  "V112"
[27] "V113" "V114" "V117" "V124" "V125" "V126" "V137" "V141" "V149" "V160" "V163" "V16
7"  "V168"
[40] "V169" "V170" "V171" "V173" "V177" "V178" "V179" "V181" "V190" "V192" "V197" "V19
9"  "V200"
[53] "V207" "V211" "V217" "V220" "V222" "V224" "V226" "V227" "V228" "V230" "V231" "V23
3"  "V234"
[66] "V237" "V238" "V239" "V240" "V241" "V242" "V243" "V247" "V248" "V249" "V250" "V25
2"  "V257"
[79] "V258" "V259" "V260" "V267" "V269" "V270" "V277" "V279"
```

Fig 5. Feature selection using Boruta method

The training and test data were divided into 80% and 20%, with 10 cross validation and 3 repeats. The performance metrics used for model evaluation are Accuracy and Kappa scores. The work used 3 machine learning models before feature selection and after feature selection and found that accuracies of SVM, KNN and Naïve Bayes before feature selection were 61.5 %, 59.3% and 46.1 % and after feature selection the accuracies of SVM, KNN and Naïve Bayes are 71.4 %, 62.6 % and 70.3 %. It can be seen from table 2, why feature selection is important when modeling machine learning algorithms.

This study shows improvement of 9.9 % in SVM model, 3.3 % in KNN model and 24.2 % in Naïve Bayes model. One surprising result was with Naïve Bayes model which shows 24.2 % and 0.50 improvement in both accuracy and kappa values and we infer that Naïve Bayes performs well when we have a smaller number of features. The performance of SVM and Naïve Bayes are comparatively similar, SVM outperforms with just 1.1 % more accuracy compared to Naïve Bayes as shown in table 2. KNN algorithm does not perform well in classification of Arrhythmia dataset.

Table 2. Performance Evaluation

Model	Performance before feature selection		Performance after feature selection	
	Accuracy (%)	Kappa	Accuracy (%)	Kappa
SVM	61.5	0.42	71.4	0.56
KNN	59.3	0.19	62.6	0.31
Naïve Bayes	46.1	0.14	70.3	0.50

In Fig. 6 and Fig. 7 shows the accuracy and kappa scores of the 3 algorithms.

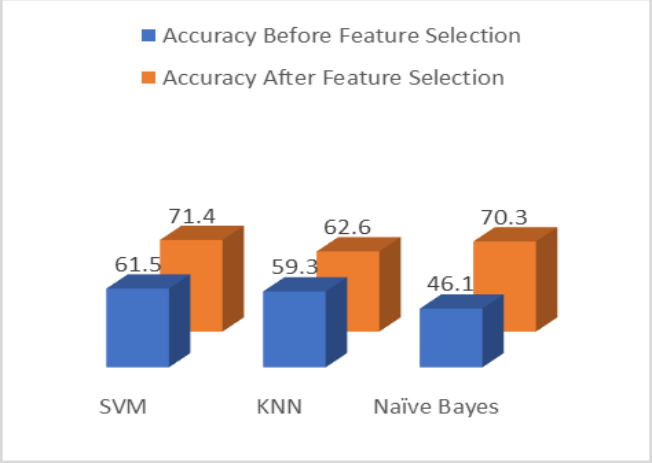

Fig. 6. Accuracy Comparison

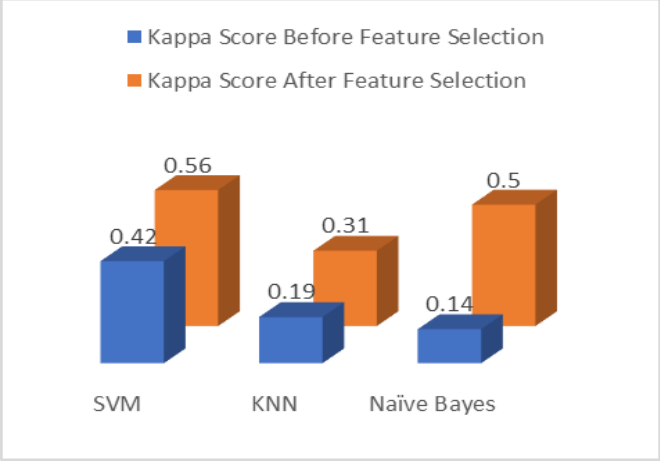

Fig. 7. Kappa Score Comparison

5. CONCLUSION AND FUTURE WORK

This paper considers Arrhythmia instances of different patients and performs multiclass classification using three machine learning algorithms SVM, KNN and Naïve Bayes. The original data was normalized by replacing missing values using Mice technique and eliminating irrelevant features using Boruta approach. After feature selection, data is split into 80 % training and 20 % test data with ten-fold cross validation. The observation shows improvement in accuracy and kappa scores when relevant features are taken into account as compared to considering huge number of features, especially with Naïve Bayes algorithm. Performance among three algorithms depicts similar results among SVM and Naïve Bayes algorithms, with KNN least performance. The future work is to study whether a particular algorithm can be generalized as a best algorithm in all given classification situations or it performs well for only particular datasets

REFERENCES

[1]. S. H. Jambukia, V. K. Dabhi and H. B. Prajapati.: Classification of ECG signals using machine learning techniques: A survey. 2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad, pp. 714-721. (2015)

[2]. Songthung, P., Sripanidkulchai, K.: Improving type 2 diabetes mellitus risk prediction using classification. In: 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–6 (2016)

[3]. Mohan, K.R., Paramasivam, I., Narayan, S.S.: Prediction and diagnosis of cardio vascular disease–a critical survey. In: 2014 World Congress on Computing and Communication Technologies (WCCCT), pp. 246–251 (2014)

[4]. 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)

[5]. Singh N., Singh P. Cardiac Arrhythmia Classification Using Machine Learning Techniques.: In: Ray K., Sharan S., Rawat S., Jain S., Srivastava S., Bandyopadhyay A. (eds) Engineering Vibration, Communication and Information Processing. Lecture Notes in Electrical Engineering, vol 478. Springer, Singapore (2019)

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

[7]. Batra, A., Jawa, V.: Classification of arrhythmia using conjunction of machine learning algorithms and ECG diagnostic criteria. Int. J. Biol. Biomed. 1, 1–7 (2016)

[8]. Yeniterzi, S., Yeniterzi, R., Kücükural, A., Sezerman, U.: Feature selection with genetic algorithms on cardiac arrhythmia database. In: The 2nd International Symposium on Health Informatics and Bioinformatics (HIBIT) (2007)

[9]. Coast, D.A., Stern, R.M., Cano, G.G., Briller, S.A.: An approach to cardiac arrhythmia analysis using hidden Markov models. IEEE Trans. Biomed. Eng. 37, 826–836 (1990)

[10]. Jadhav, S.M., Nalbalwar, S.L., Ghatol, A.A.: ECG arrhythmia classification using modular neural network model. In: IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), pp. 62–66 (2010)

[11]. Luz, E.J.D.S., Schwartz, W.R., Cámara-Chávez, G., Menotti, D.: ECG-based heartbeat classification for arrhythmia detection: a survey. Comput. Methods Programs Biomed. 127, 144–164 (2016)

[12]. Abe, S.: Support vector machines for pattern classification (Vol. 53). London: Springer (2005)

[13]. S. B. Imandoust And M. Bolandraftar.: Application of K-Nearest Neighbor (KNN) Approach for predicting Economic Events: Theoretical Background. S B Imandoust et al. Int. Journal of Engineering Research and Application, vol.3, Issue 5, pp.605-610 (2013)

[14]. R. A. Nugrahaeni and K. Mutijarsa.: Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification. International Seminar on Application for Technology of Information and Communication (ISemantic), Semarang, 2016, pp. 163-168 (2016)

[15]. T. Mahboob, S. Irfan and A. Karamat.: A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms," 2016 19th International Multi-Topic Conference (INMIC), Islamabad, pp. 1-8 (2016)

[16.] Dua, D. and Graff, C.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[17]. Miron Kursa and Witold Rudnicki.: Feature Selection with the Boruta Package. Journal of Statistical Software, Articles, vol.36, no.11, year (2010)