International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

Prediction and Classification of Cardiac Arrhythmia

Kanishka Thakur[1], Mukul shishodia[2] and Kamal Singh Bisht[3]

[1,2,3]*Department of Computer Science, IMS ENGINEERING COLLEGE, Ghaziabad ,Uttar Pradesh, India*

---***---

Abstract- Rapid advancements in technology have facilitated early diagnosis of diseases in the medical sector. One of the most prevalent medical conditions that demands early diagnosis is cardiac arrhythmia. ECG signals can be used to classify and detect the type of cardiac arrhythmia. This paper uses a novel approach to classify the ECG data into one of the sixteen types of arrhythmia using Machine Learning. The proposed method uses the UCI Machine Learning Repository dataset of cardiac arrhythmia to train the system on 279 different attributes. In order to increase the accuracy, the method uses Principal Component Analysis for dimensionality reduction, Bag of Visual Words approach for clustering and compares different classification algorithms like Support Vector Machine, Random Forest, Logistic Regression and K-Nearest Neighbor algorithms, thus choosing the most accurate algorithm, Support Vector Machine.

Key Words *:*Arrhythmia, ,Principal Component analysis, Support vector machine, K-NN, ,logistic regression.

1.INTRODUCTION

In India, a death is recorded every 33 seconds due to heart attack. In the past few decades, coronary heart disease, hypertension and other cardiovascular disease have become a global threat to human life. In our country, this phenomenon is getting increasingly severe due to the aging of population, living environment and unhealthy food consumption.

ECG provides the information which is needed to identify the problems and hence it becomes important when developing an advanced diagnostic system.

1.1Objective:

OurmotiveistoclassifyapatientintooneoftheArrhythmiaclasseslikeTachycardia and Bradycardia based on his ECG measurements and help us in understanding the application of machine learning in medical domain. After appropriate feature selection we planto solve this problem by using Machine Learning Algorithms namely KNearest Neighbour, Logistic Regression, Naïve Bayes and SVM and compared the results in order to get the suitable algorithm for correctly predicting the class cardiacarrhythmia.

1.2 About Model:

Inthispaper,the aim is to develop a hybrid model which uses various machine learning techniques like principal component analysis, Bag of Words model and various classification algorithms. Using this model, it is possible to classify an ECG signal to one of the 16 classes of arrhythmia, where class 1 means normal ECG signal, classes 2 to 15 are different types of arrhythmia and class 16 refers to the rest of unclassified ones. The use of machine learning will help in greater accuracy and high potential to detect severe cardiac arrhythmia possibilitie.

2. RELATED WORK

Our proposed model makes use of this

concept in cardiology—

2.1 A Performance Analysis of Artificial Neural Networksfor :

Cardiac Arrhythmia Detection

The paper takes in an ECG signal and converts the analog signal to a digital signal.The system has extracted 8beats from each ECG signal sampled at 2223 samples per second and classified these beats.The next step was signal pre-processing which was denoising of loaded raw ECG signal. The system then extracts just three features from the signal;QRS complex duration, RR interval both normal and the one averaged over 8 beats. These features were further used by ANN classifiers such as Naive Bayes and Multi- class SVM to predict the class of the arrhythmia. The results were compared and the accuracy of each of the algorithm iscalculated.

2.2 .Identifying Best Feature Subset For Cardiac ArrhythmiaClassification:

This paper presents a model which is divided into two parts - filter part and wrapper part. The filter part deals with feature selection from the cardiac arrhythmia dataset of the UCI machine learning repository.These help in identify ing the best features without taking any assistance of a classification algorithm, but rather, just using a set of presumedcriteria.

The features election model present ed make suseo f both,filter and wrap per techniques of feature selection. For judging the relative importance of each feature, an improved F-score is calculated for each and every feature, which produces a superset of features that can be used. Sequential Forward Search is then used for finding the final subset ofmost important features. Following this, SVM and KNN are used for classification of cardiac arrhythmia using the new list offeatures

Data set

The dataset for the project is taken from UCI Repositoryhttps://archive.ics.uci.edu/ml/datasets/ Arrhythmia There are (452) rows, each representing medical record of a different patient.279 attributes like age, weight and patient's ECG related data are there. General attributes like age and weight have discrete integral values while other ECG features like QRS duration have real values. The variable Class is our target variable. There are in total 13classes.

TABLE 1:CLASSES OF CARDIAC ARRHYTHMIA

NO.	CLASS	INSTANCES
1	Normal	245
2	Ischemic changes(Coronary Artery)	44
3	Old Anterior Myocardial Infarction	15
4	Old Inferior Myocardial Infarction	15
5	Sinus tachycardia	13
6	Sinus bradycardia	25
7	Ventricular Premature Contraction	3
8	Supraventricular Premature Contraction	2
9	Left bundle branch block	9
10	Right bundle branch block	50
11	Left ventricle hypertrophy	4
12	Atrial Fibrillation or Flutter	5
13	Others	22

3. OURAPPROACH

3.1 Feature Selection:

From the dataset, out of the 279 features present, it was infeasible to extract all the features. This is because many features used some information that is not accessible to the doctors while analysing ECG reports of patient. Hence, the dataset was narrowed with the help of Principal Component Analysis (PCA).

3.2 Principal component analysis:

Principalcomponentanalysisisamethodofextractingvariablesthatinfluencethefinal decision the most and provide as much as information as possible. The aim of PCA in this paper is to reduce the dataset containing large amount of dimensions and find out features with low dimensions. A principal component is a combination of the normalized linear original predictors in adataset. Let us assume a

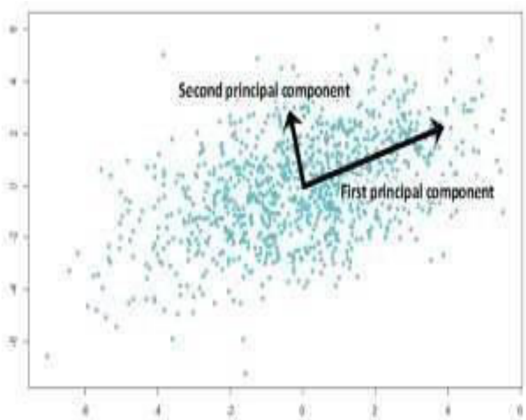
Predictor set as: Y^1, Y^2 ,……, Y^n

The principal component can be written

as: $Z^1= \Phi^{11}Y^1+ \Phi^{21}Y^2+ \Phi^{31}Y^3 +\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \Phi^{n1}Y^n$

Z^1 is first principal component Φn^1 is the loading vector that comprises of loadings (Φ^1,Φ^2..) of first principal component. The loadings are restricted to a unit sum of square.There as on beingt ha tlarge variance an becaused due to high magnitude of the loadings. Φn^1 defines the direction of the principal component (Z^1) along which maximum variance of the data is observed.It give srisetoa line in ndimensional space which is in close proximity to the m observations. Average squared Euclidean distance is used to measure thecloseness.

X^1...Xn are normalized predictors; that have zero mean and unit standard deviation.



The variability captured by the first component is directly proportional to the information captured by that component.

The first principal component results in a line which is nearest to the data i.e. the minimumsumofsquareddistancebetweena datapointandtheline.Thefirstprincipal component outputs a line which is nearest to the data i.e. the minimum value obtained by summing the squared distance between a data point and theline.

Ascreeplotisdevelopedtofindfactorswhichcapture mostofthedatavariability.The values are represented in decreasing order. By plotting a cumulative variance plot, we get a further clearer picture of the number of componentsrequired.

The plot in Fig. shows 150 components depicting around 99% variance in the dataset. Therefore, using PCA the 279 predictors were reduced to 150 with the same explained variance.

3.2.1KNN (K-NearestNeighbours):

$$\square(\square,\square)=\sqrt{\sum_{i}(\square(\square)-\square(\square))^{2}}$$

Here we used KNN because it is simple to implement & very straight forward. Here, an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. This is done by measuring distances between the object and its neighbours. The following formula shows a representation of simple Euclidian distance, where ‘a’ and ‘b’ are the respective positions of the object and one of its neighbours. KNN is very sensitive to irrelevant or redundant features as all features contribute to the similarity and thus to the classification. This was improve by careful feature selection described previously. The results are summarized below.

TABLE 2:KNN CLASSIFICATION WITH PCA

Training-Testing Size	K Neighbours	Training Accuracy	Test Accuracy
70%-30%	6	100 %	55.14

3.2.2LogisticRegression:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}(1-y^{(i)})\log(1-h_\theta(x^{(i)})) + y^{(i)}\log h_\theta(x^{(i)})\right]$$
$$= -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=0}^{1} 1\left\{y^{(i)}=j\right\}\log p(y^{(i)}=j|x^{(i)};\theta)\right]$$

Logistic regression hypothesis gives the output as a estimated probability .A threshold value is set to and based upon a this threshold a estimated probability can be classified intoclass.Forex.letusthresholdvalueis0.5thena0.6estimatedprobabilitytheninput is considered as in class 1 whereas a input with estimated probability 0.3 is considered as in class 0.

Logistic regression hypothesis uses a sigmoid function. We need to maximize the probability by minimizing the loss function .Decreasing the cost will increase the maximum likelihood. Values of Coefficients (beta) that minimize the error in the probabilities predicted by the model to those in data.

TABLE 3:LOGISTIC REGRESSION CLASSIFICATION WITH PCA

Training-Testing Size	Training Accuracy	Test Accuracy
70%-30%	88.92 %	72

3.2.3 Naïve – BayesClassifier

$$P(x,y) = \prod_{i=1}^{m}\left(\prod_{j=1}^{n}\phi_{j,x_j^{(i)}|y=y^{(i)}}\right)\phi_{y^{(i)}} \quad (5)$$

InNaiveBayesalgorithmweassumesthatpredictorsareindependentandusesthebays theorem for classification purpose. We calculate posterior probability and a class with highest posterior probability isoutcomes.

Here

P(c|x)=(P(x|c)P(c))/P(x)

Where

P(c|x) is the posterior probability of class x

P(x) is the prior probability of predictors

P(c) is the prior probability of class.

P(x|c) is likelihood which is the probability of predictor given class.

this algorithms convert input into frequency tables and then with the help of calculated prior probabilities , we calculate posterior probabilities and consider a highest among them as outcome. The results are summarised below –

TABLE 4:NAÏVE BAYE CLASSIFIACATION

Training-Testing Size	Training Accuracy	Test Accuracy
70%-30%	70.56 %	55.88

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

3.2.4 SVM (Support VectorMachines)

In SVM we find a Hyperplane for N dimensional space where N is the number of features, this hyper plane is a line for 2 dimensional and a plane is considered as a hyperplane for 3 Dimension. Points falling in same side of a hyperplane is considered as in a same classes.

In SVM our aim is to maximize the margin of hyperplane from the points. For this purpose we have Support vectors, this are the points closer to hyperplane and influence its position.

With PCA

Training-Testing Size	Training Accuracy	Test Accuracy
70%-30%	95.88 %	**72.05%**

dimensional space where N is the number of features, this hyper plane is a line for 2 dimensional and a plane is considered as a hyperplane for 3 Dimension. Points falling in same side of a hyperplane is considered as in a same class.by plugging input values into the equation of hyperplane, we can predict the class of a input

3.3 .Results:

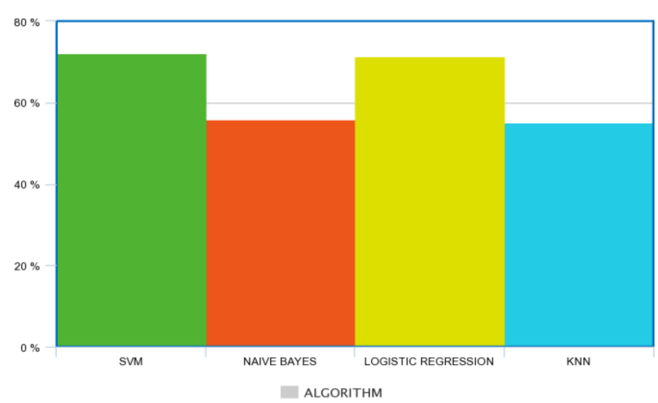
FIG 1: COMPARISON GRAPH BETWEEN DIFFERENT ALGORITHMS

The main objective of this project was to develop a system that could robustly detectan arrhythmia. The second objective of this project was to develop a method to robustly classify an ECG trace into one of 13 broad arrhythmia classes. We report our performance for each of the five methods using two different methodologies. Weshow results for each algorithm, as well as vary other parameters for betterresults.

We obtained best result in SVM with 72.05 accuracy, other algorithms Logistic regression gives accuracy 71.42, KNN gives accuracy 55.14 and naïve Bayes gives accuracy 55.88.

4.CONCLUSION

It is clear from the above data that the SVM and Logistic Regression algorithms are capable of automatically detecting arrhythmias with reliable accuracy(Training Data = 88.9% and Testing Data=72%). Our general approach in this project was as follows. We started with KNN and we tried to obtain maximum accuracy for different valuesofK ranging from 3 to 13. Then we used Logistic Regression which uses the sigmoid



functionandweranitusingGradientdescentandNewton'smethod.Logisticregression gave comparatively better results with average accuracy around 73 %. Naïve-Bayes classifier gave poor results due to problem of lack of enough training examples (452) and excessive number of features. SVM using linear kernels gave the best results with averageaccuracyofclassificationaround96%fortrainingsetand73%fortestingset.

We used 4 classifiers for the classification of cardiac arrhythmia. These were Naive Bayes Algorithms, Support Vector Machine, Logistic Regression and KNN classifier.

When the dataset was cross-validated and tested, the maximum accuracy was found to beobtainedbySupportVectorMachineClassifier.The accuracyobtainedwas72.05%.

Thus in our approach, we have used the Support Vector Machine Classifier to obtain the best possible results for classifying arrhythmia.

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to our faculty, Prof. Mukesh Kumar Singh Department of computer science and engineering, IMS Engineering College for helpingus and guiding us in the execution of this project and also for his consistence support and valuable suggestion throughout our research work for this project.

REFERENCES

[1] "UCI machine learning repository: Arrhythmia data set," 1998.[Online]. Available: https://archive.ics.uci.edu/ml/datasets/Arrhythmia . Accessed:Feb. 10,2017.

[2] "Heart attack kills one person every 33 seconds in India - Times of India", The Times of India, 2017. [Online]. Available:http://timesofindia.indiatimes.com/life-style/health-fitness/healthnews/Heart-attack-kills-one-person-every-33-seconds-in-India/articleshow/52339891.cms. [Accessed: 09-Mar- 2017].

[3] S. Xue, X. Chen, Z. Fang, and S. Xia, "An ECG arrhythmia classification and heart rate variability analysis system based on android platform," 2015 2nd International Symposium on Future Information Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC) and Communication Technologies for Ubiquitous HealthCare (Ubi-HealthTech), May2015.

[4] Nir Kalkstein, YaronKinar, Michael Na'aman, Nir Neumark, and PiniAkiva, "Using Machine Learning to Detect Problems in ECG Data Collection," in Computing in Cardiology, IEEE,2011.

[5] O. Valenzuela, F. Rojas, L. J. Herrera, F. Ortuno, H. Pomares, and I. Rojas, "Comparison of different computational intelligent classifier to autonomously

detect cardiac pathologies diagnosed by ECG," 2013 13th International Conference on Intelligent Systems Designand

Applications, Dec. 2013.

[6] Desai, Usha et al. "Machine Intelligent Diagnosis Of ECG For Arrhythmia Classification Using DWT, ICA And SVM Techniques". 2015 Annual IEEE India Conference (INDICON) (2015): n. pag. Web.4 Sept. 2016.

[7] Deselaers, Thomas, Lexi Pimenidis, and Hermann Ney. "Bag-of- visual words models for adult image classification and filtering." *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*.IEEE,

2008.

[8] https://ieeexplore.ieee.org/document/8282537