A Machine Learning Approach for the Classification of Cardiac Arrhythmia

Prajwal Shimpi
Student, Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India
prajwalshimpi@gmail.com

Sanskruti Shah
Student, Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India
sanskrutirahulshah@gmail.com

Maitri Shroff
Student, Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India
maitrishroff1995@gmail.com

Anand Godbole
Professor, Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India
anand_godbole@spit.ac.in

***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 introduces 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 [1] 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.**

Keywords—Machine Learning, Artificial Intelligence, Classification, Bag of Visual Words, Clustering, Cardiac Arrhythmia

I. INTRODUCTION

In India, a death is recorded every 33 seconds due to heart attack [2]. 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. [3]

Electrocardiogram (ECG) data can be easily converted into digital format with methods that are out of scope of this paper. The digitized data ranges from two extremes - small number of data-points on a large group of people to large amount of data on a small group of people. This new scenario in the medical world calls for application of machine-learning approaches to analyze the data - the analysis may be directed at providing better medical treatment (personalized medicine), at identifying emergencies and future outcomes as early as possible [4].

ECG provides the information which is needed to identify the problems and hence it becomes important when developing an advanced diagnostic system [5]. An irregular heartbeat i.e. arrhythmia can be detected from an ECG signal, which can be used to detect if a person might suffer cardiovascular problems in the future. During visual interpretation of ECG by physicians, miscalculation of beats may give rise to inaccuracies. This is because the interpretation is complicated and time consuming for large datasets. Also, good discrimination between normal and abnormal classes is not guaranteed by features pertaining to time domain. These difficulties can be solved using machine intelligent classification systems. [6]

Past work in this arena includes frameworks meant for experimentation, which use machine learning techniques to predict a disease and suggest remedies for the classified disease. This work is majorly based on Neural Networks, Markov chain models and SVM.

In this paper, 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 possibilities.

II. LITERATURE SURVEY

A. Bag-of-Visual-Words Models for Adult Image Classification and Filtering[7]

This paper presents a Bag of Visual Words approach for the filtering of pornographic images and classifying them into different classes of explicit content. Here, the images are partitioned into patches which represent the local features, which are found using the difference of Gaussian interest points. These are used as the visual words in the histogram. The unsupervised training algorithm then classifies the histograms into the probable classes using Support Vector Machines and Log-Linear Models. A filter system is then designed, with five classes, with class 0 as the least and class 4 as the highest level of explicit content. Binary filtering is then performed and the image is filtered if the value returned is +1, and allowed if the value is -1. The analysis is performed on approximately 8500 images, 1700 images belonging to each class for the testing of the proposed algorithm.

In the past papers, the concept of 'bag of visual words' has not been used in any other arena except for this and for the

classification of food images. Thus, our proposed model makes use of this concept in cardiology.

B. *Performance Analysis of Artificial Neural Networks for Cardiac Arrhythmia Detection[8]*

The paper takes in an ECG signal and converts the analog signal to a digital signal. The system has extracted 8 beats from each ECG signal sampled at 2223 samples per second and classified these beats. The next step was signal preprocessing 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 is calculated.

C. *Identifying Best Feature Subset For Cardiac Arrhythmia Classification [9]*

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 [1]. These help in identifying the best features without taking any assistance of a classification algorithm, but rather, just using a set of presumed criteria. The feature selection model presented makes use of both, filter and wrapper 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 of most important features. Following this, SVM and KNN are used for classification of cardiac arrhythmia using the new list of features.

III. MODEL OF OUR APPROACH

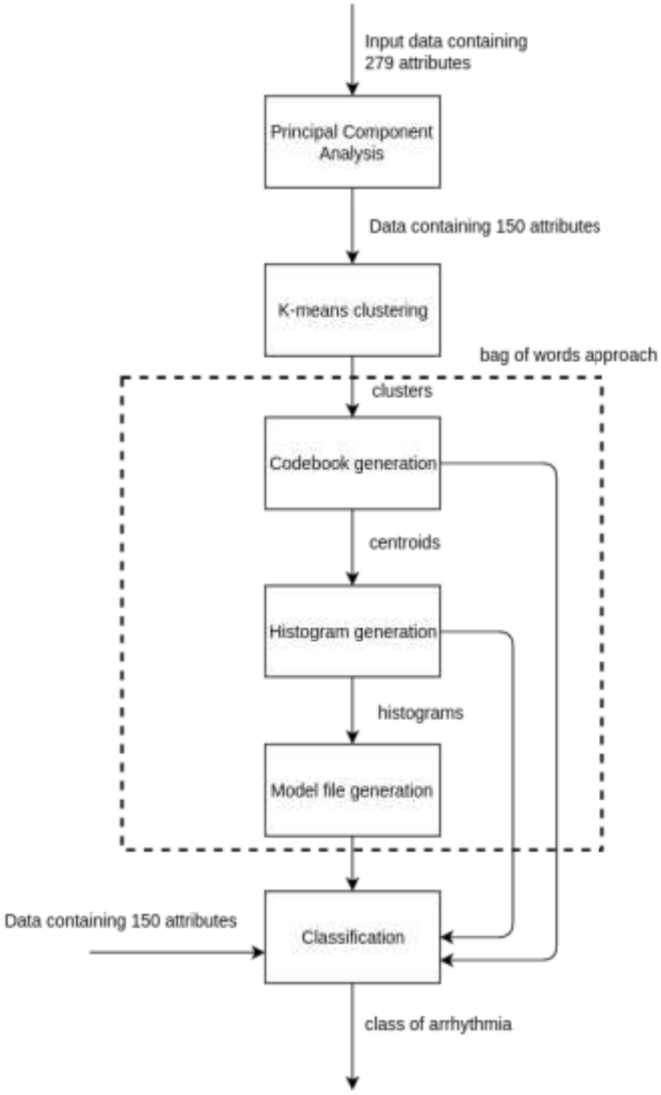

Fig. 1. *Model of our approach*

The flow of the algorithm used is displayed in Fig. 1.

IV. PRINCIPAL COMPONENT ANALYSIS

A. *Dataset*

The dataset was obtained from the UCI - Machine Learning Repository [1] which contains the patient ECG data for 472 patients. Each record contains 279 attributes.

B. *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 analyzing ECG reports of patient. Hence, the

dataset was narrowed with the help of Principal Component Analysis (PCA).

Principal component analysis is a method of extracting variables that influence the final decision the most and provide as much as information as possible [10]. PCA is simple and efficient. Along with simplicity and efficiency, data compression can be achieved by representing subspace in low dimension. Apart from dimensionality reduction, it finds its application in statistical pattern recognition in image processing. [11] Principal Component Analysis can be applied to practically any dataset or scenario, like reducing the number of features obtained from the feature extraction of an image of a user's iris, as shown in [12]. 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 a dataset.

Let us assume a predictor set as:

$Y^1, Y^2,...Yn$

The principal component can be written as:

$Z^1 = \Phi^{11}Y^1 + \Phi^{21}Y^2 + \Phi^{31}Y^3 + ... + \Phi n^1 Yn$

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. The reason being that large variance can be caused 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 gives rise to a line in n dimensional space which is in close proximity to the m observations. Average squared Euclidean distance is used to measure the closeness.

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

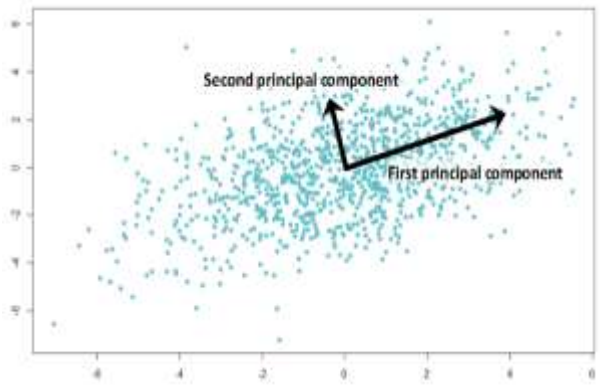

Fig. 2. *The first two principal components in two dimensional graph [10]*

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 minimum sum of squared distance between a data point and the line. The first principal 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 the line.

A scree plot is developed to find factors which capture most of the data variability. The values are represented in decreasing order. By plotting a cumulative variance plot, we get a further clearer picture of the number of components required. [10]

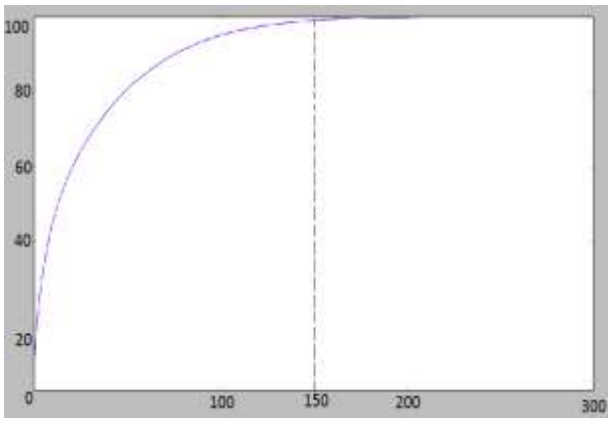

Fig. 3. *Scree plot of number of necessary attributes*

The plot in Fig 2. 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.

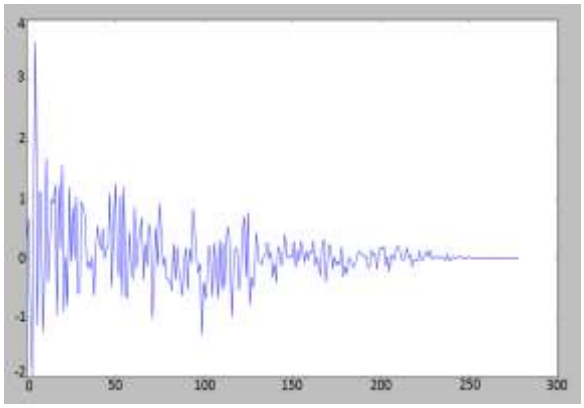

Fig. 4. *Cumulative variance graph*

The plot shows the variance of the 279 predictors, we chose 150 predictors with the least variance.

V. Bag-of-Words Approach

BOW was originally introduced as an approach used for NLP and retrieval of information of documents which are treated as orderless collection of words, word order, etc. In the recent years, this model has been used in the fields of computer vision like image classification and object recognition.

All the features extracted from an image are used as a pattern in this model. BOW is used for prototyping complex

objects, such as cloud with diverse parts in a concise feature vector.

Once the descriptors for every superpixel in the training set are obtained, K-Means is used as non-supervised clustering on these descriptors. "K" stands for the number of clusters that have to be created, "means" is the the average value of each class. Due to its speed and efficiency, it is popular in large data clustering. [13] This in turn generates a codebook, A=[a1,a2,…,aC]∈RD×C with C codewords (C=k).

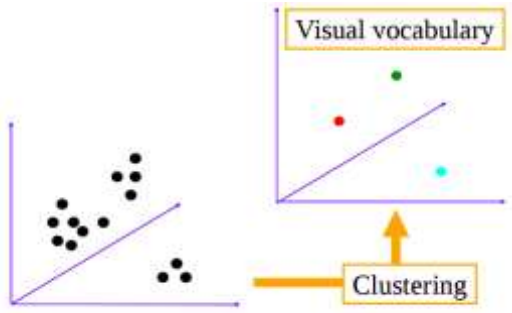

Fig. 5. *Clustering [14]*

V=[v1,v2,…,vN]∈RC×N can be obtained as the respective representation of those N features in F using the codebook. In feature coding, each fi is denoted by the codebook A by designating SIFT descriptor of the image patch to the nearest visual words calculated by L2 norm. This produces values on C codewords; with a coding vector vi with C elements. [8]

For each superpixel a histogram can be produced. Next, two operations are performed on the histogram; it is measured using the weighting plan and normalized with the L2 norm to produce a frequency vector of length C. [14]

Similarly, using this approach for cardiac arrhythmia, we perform K-Means clustering on the UCI dataset [1], since we do not need feature extraction here. We have 16 classes of cardiac arrhythmia, but due to the lack of data of classes 11, 12 and 13, we consider just the 13 classes. The final centroids of these newly formulated clusters are then stored in a codebook file. This file is used during the computation of histograms. A histogram is a frequency chart of every feature in an image. Histograms for every image are subsequently computed using the codebook file and then, these too, are written to a file.

Next, these histograms provide the necessary features and class IDs which are essential for the formulation of the model files. Model files for the each classification algorithms are created, which are used during testing.

VI. Classification

Reading and understanding different machine learning algorithms, four of the algorithms were shortlisted.

1) *Support Vector Machines (SVM)*
2) *Logistic Regression Algorithm*
3) *K-Nearest Neighbors (KNN) Algorithm*
4) *Random Forest Algorithm*

SVM can be used in a variety of domains such as face and speech recognition, face detection and image recognition. It's simple algorithm, as outlined in [15], makes it an easy to use classification technique. Model files created at the training phase along with the input values are used for testing.

VII. Result and Analysis

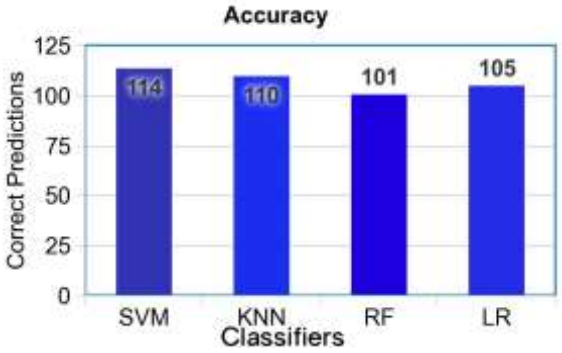

Fig. 6. *Accuracy of the four classifiers*

The test data contained 125 samples out of which the correctly classified samples are illustrated in the Fig 6.

VIII. Future Scope

The UCI dataset [1] used consisted of already extracted features. The future scope of this paper involves directly extracting features from an ECG signal. Apart from the 16 possible classes that were taken in consideration in this paper, the ECG signal can be classified into a different set of cardiac arrhythmias.

IX. Conclusion

We used 4 classifiers for the classification of cardiac arrhythmia. These were Random Forest Algorithm, Support Vector Machine, Logistic Regression and KNN classifier. When the dataset was cross-validated and tested, the maximum accuracy was found to be obtained by Support Vector Machine Classifier. The accuracy obtained was 91.2%. Thus in our approach, we have used the Support Vector Machine Classifier to obtain the best possible results for classifying arrhythmia.

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/health-news/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

and Communication Technologies for Ubiquitous HealthCare (Ubi-HealthTech), May 2015.

[4] Nir Kalkstein, Yaron Kinar, Michael Na'aman, Nir Neumark, and Pini Akiva, "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 Design and 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] N. Sultana, Y. Kamatham, and B. Kinnara, "Performance Analysis of Artificial Neural Networks for Cardiac Arrhythmia Detection," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, IEEE, 2016. Accessed: Aug. 3, 2016.

[9] Niazi, Khalid Ahmed Khan et al. "Identifying Best Feature Subset For Cardiac Arrhythmia Classification". 2015 Science and Information Conference (SAI) (2015): n. pag. Web. 28 Sept. 2016.

[10] A. V. C. Team, S. Ray, F. Shaikh, D. Gupta, and S. Kaushik, "Practical guide to principal component analysis (PCA) in R & python," in Python, Analytics Vidhya, 2017. [Online].Available:https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/. Accessed: Feb. 10, 2017.

[11] Agarwal, Sugandha, Priya Ranjan, and Amit Ujlayan. "Comparative Analysis Of Dimensionality Reduction Algorithms, Case Study: PCA". *2017 11th International Conference on Intelligent Systems and Control (ISCO)* (2017): n. pag. Web. 19 Dec. 2016.

[12] Dewi, Aisyah Kumala, Astri Novianty, and Tito Waluyo Purboyo. "Stomach Disorder Detection Through The Iris Image Using Backpropagation Neural Network". 2016 International Conference on Informatics and Computing (ICIC) (2016): n. pag. Web. 13 Dec. 2016.

[13] Jia Qiao, and Yong Zhang. "Study On K-Means Method Based On Data-Mining". 2015 Chinese Automation Congress (CAC) (2015): n. pag. Web. 17 Nov. 2016.

[14] Y. Yuan and X. Hu, "Bag-of-words and object-based classification for cloud extraction from satellite imagery," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 8, pp. 4197–4205, Aug. 2015.

[15] Karamizadeh, Sasan et al. "Advantage And Drawback Of Support Vector Machine Functionality". 2014 International Conference on Computer, Communications, and Control Technology (I4CT) (2014): n. pag. Web. 5 Dec. 2016.