ECG Signal Classification using K Nearest Neighbors

Bishal Malla
Dept. of Computer and Electronics Engineering
Kantipur Engineering College
Lalitpur, Nepal
bishalmalla16@gmail.com

Jenish Pokharel
Dept. of Computer and Electronics Engineering
Kantipur Engineering College
Lalitpur, Nepal
jenish.pokharel99@gmail.com

Manish Paudel
Dept. of Computer and Electronics Engineering
Kantipur Engineering College
Lalitpur, Nepal
manish.paudel09@gmail.com

Mimansha Khadka
Dept. of Computer and Electronics Engineering
Kantipur Engineering College
Lalitpur, Nepal
mimanshkdk07@gmail.com

Bishal Thapa
Project Assistant, Lecturer
Dept. of Computer and Electronics Engineering
Kantipur Engineering College
Lalitpur, Nepal
thapamarine8@gmail.com

***Abstract*—Electrocardiogram(ECG) is used by doctors as an important diagnosis tool and in most cases, which is recorded and analyzed at hospital after the appearance of first symptoms or recorded by patients using a device named ECG and analyzed afterward by doctors.This paper provides an approach to detect various kind of arrhythmia based on the data provided by the patient (user). The paper uses K Nearest Neighbors (KNN) Classification approach to accurately identify the disease. At first, the data sets are provided to the system and most important features are selected using ExtraTree Classier. Secondly, the model is trained using KNN Classification algorithm. Next stage involves collecting the data based on the ECG checkup provided by the user which is subject to testing procedure with the help of Euclidean distance.**

***Index Terms*—Cardiovascular, ECG (Electrocardiogram), ExtraTree, K Nearest Neighbors**

I. Introduction

The electrocardiogram (ECG) signal contains lots of pathological information about patient's heart processes. One important analysis in the ECG is the classification of heartbeats, which is important for the detection of arrhythmia. Arrhythmia can be split into life-threatening and non-life-threatening arrhythmia. Heartbeat classification of long-term ECG recordings required for the diagnosis of non-life-threatening arrhythmia could be time-consuming and impractical, thus automatic algorithms could exhibit a great aid. Therefore, the automatic heart beat classification of the latter arrhythmia is worth studying. ECG which reveals the rhythm and activity of the heart, is an important non-invasive clinical tool for cardiologists to diagnose various heart diseases. Many automatic ECG arrhythmia classification systems have been investigated using computational intelligence. A successful ECG arrhythmia classification sometimes contains three important procedures: feature extraction, feature selection, and classification. Feature extraction is an important procedure that usually influences the classification performance of any ECG arrhythmia classification system. Therefore, to extract relevant features and reduce their dimensions to achieve the best classification results with higher accuracy have become the primary stage for the ECG arrhythmia classification problems.

In today's world, an optimal and intelligent problem solving approaches are required in every field, regardless of simple or complex problems. Researches and developers are trying to make machines and software more efficient, intelligent and accurate. This is where the Artificial Intelligence plays its role in developing efficient and optimal solutions. Data mining techniques are used to explore, analyze and extract data using complex algorithms in order to discover unknown patterns in the process of knowledge discovery. Prediction is done with the help of available knowledge or previous values so accuracy in prediction is the main challenge.

With the available dataset, ECG classification techniques is partitioned into three parts, namely; preprocessing, feature selection and classification.

- In the pre-processing stage, several missing values are replaced with an average value.
- In the feature selection stage, only selection of

relevant features is done.
- In classification stage, the classifier is applied for final-heartbeat classification.

II. Related Works

AliveCor's FDA-cleared Heart Monitor and AliveECG app for the android is for use by medical professionals and patients to record and review single-channel electrocardiogram (ECG) rhythm strips [1]. This app requires the AliveCor Heart Monitor device that attaches to your mobile phone for taking ECG recordings. AliveCor turns your smart phone into an electrocardiogram by snapping onto the back of any iPhone. Now you can record your ECG in seconds. Just start the app and place your fingers on the metal sensors. To take cardiac measurements, the user presses the device against the skin found near the heart. The device has received much recognition for its mobility and the anticipation of its capability to catch irregular heart rhythms earlier.

In [2], Jovic et al. proposed a classification method based on analysis of a combination of Heart Rate Variability and approximate entropy features in order to classify ECG in four classes (normal heart rhythm, arrhythmia, supra-ventricular arrhythmia, and congestive heart failure). ECG records from online databases were analyzed by seven clustering and classification algorithms. Results show that the top 3 accurate classification methods are Random Forest(RF) with 99.6%, Bayesian Networks with 99.4% and Support Vector Machine (SVM) with 98.4%.

Electrocardiogram (ECG), non-stationary signals, is extensively used to evaluate the rate and tuning of heartbeats. In [3], the main purpose is to provide an overview of utilizing machine learning and swarm optimization algorithms in ECG classification. Furthermore, feature extraction is the main stage in ECG classification to find a set of relevant features that can attain the best accuracy. Published literature presented in this paper indicates the potential of Artificial Neural Network(ANN) and Support Vector Machine(SVM) as a useful tool for ECG classification.

Zhu et.al [4], presented the investigation of the use of artificial neural networks for ECG. Abnormality detection in an ECG monitoring scheme was done. Simulations are performed to explore how well neural networks can work after a short learning phase. The method used aims to produce a system which will perform well practically. It uses the waveform slope to locate the QRS complex for each ECG cycle which allows a small training data set to be formed for each individual patient. The performance of three neural network models is compared and results of the simulation show that the correct recognition of normal cycles and VPB cycles is typically greater than 95%.

In [5], N. Deepika et al. proposed Association Rule for classification of Heart-attack patients. The extraction of significant patterns from the heart disease data warehouse was presented. The heart disease data warehouse contains the screening clinical data of heart patients. Initially, the data warehouse pre-processed to make the mining process more efficient. Later, the significant items were calculated for all frequent patterns with the aid of the proposed approach. The frequent patterns with confidence greater than a predefined threshold were chosen and it was used in the design and development of the heart attack prediction system.

In [6], Chazal and Reilly presented a method for the automatic processing of the ECG for the classification of heartbeats. Feature sets were based on ECG morphology, heartbeat intervals, and RR-intervals. Artificial Neural Network (ANN) has been used to classify the ECG arrhythmias. Different structures of ANN have been trained by arrhythmia separately and also by mixing these 10 different arrhythmias. These patterns were tested with the most appropriate ANN structures of single classification case and mixed classification cases. Classifications were found to be 4.3% and the average error of mixed classification 2.2%.

III. Proposed Method

In the field of medical case for ECG, the authors seek for a classifier that is good enough to classify the patients ECG data with minimal dataset. The proposed system architecture is well described by the figure below:

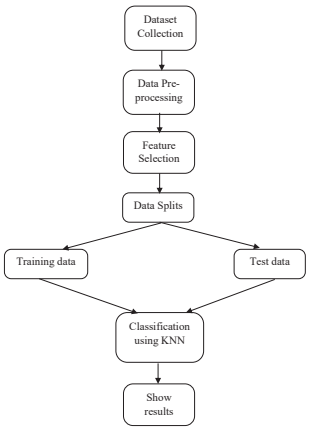

Fig. 1. ECG Classification model architecture

For this ECG Classification, dataset is downloaded from UCI repository. The training dataset contains a total data of 452 patients consisting 278 features each. This Cardiac Arrhythmia Database contains 279 attributes, 206 of which are linear valued and the rest are nominal. [7]

The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 13 groups. Class 01 refers to Normal, ECG classes 02 to 12 refers to different classes of arrhythmia and class 13 refers to the unclassified ones.

TABLE I
OUTPUT CLASS DISTRIBUTION

Class Code	Class	No. of instances
1	Normal	245
2	Ischemic changes (Coronary Artery Disease)	44
3	Old Anterior Myocardial Infarction	15
4	Old Inferior Myocardial Infarction	15
5	Sinus tachycardy	13
6	Sinus bradycardy	25
7	Ventricular Premature Contraction (PVC)	3
8	Supraventricular Premature Contraction	2
9	Left bundle branch block	9
10	Right bundle branch block	50
11	Left ventricule hypertrophy	4
12	Atrial Fibrillation or Flutter	5
13	Others	22

A. Data Pre-processing

Data pre-processing generally includes error correction such as normalization, noise elimination, filling up the missing values, etc. Pre-processing includes correction of noisy data, filling up the missing values, removing the outliers and thus generating a refined dataset for further processing. In this paper, data pre-processing involves simply filling up missing values with an average value.

B. Feature Selection

Feature selection is a process where it automatically select those features in the dataset that contribute most to the prediction variable or output. Having irrelevant features in the data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

This paper uses ExtraTree Classifier for important feature selection.

1) ExtraTree Classifier: ExtraTrees Classifier is an ensemble learning method fundamentally based on decision trees. It is like Random Forest, randomizes certain decisions and subsets of data to minimize over-learning from the data and over-fitting.

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. Information gain is used to decide which feature to split on at each step in building the tree. Simplicity is best, so to keep our tree small, at each step it is needed to choose the split that results in the purest daughter nodes.

During feature selection from the pre-processed ECG datasets, it consist of decision trees, each of them built over a random extraction of the observations from the data-set and a random extraction of the features. At each node, the tree divides the data-set into 2 buckets, each of them hosting observations that are more similar among themselves and different from the ones in the other bucket. Therefore, the importance of each feature is derived from how pure each of the buckets is and the important ECG attributes are thus separated.

C. K-Nearest Neighbors(KNN)

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given unseen observation. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by

$$d\left(x, x'\right) = \sqrt{(x_1 - x_1{}')^2 + (x_2 - x_2{}')^2 + ... + (x_n - x_n{}')^2}$$

More formally, given a positive integer K, an unseen observation x and a similarity metric d, KNN classifier performs the following two steps:

- It runs through the whole data set computing d between x and each training observation. Well call the K points in the training data that are closest to x the set A. Note that K is usually odd to prevent tie situations.
- It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label. (Note I(x) is the indicator function which evaluates to 1 when the argument x is true and 0 otherwise)

$$P\left(y = j \mid X = x\right) = \frac{1}{K} \sum_{i \varepsilon A} I\left(y^{(i)} = j\right)$$

Finally, the users input x gets assigned to the class with the largest probability.

1) Pseudo-code of KNN:

1) Load the data
2) Initialize the value of k
3) For getting the predicted class, iterate from 1 to total number of training data points
   - Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since its the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
   - Sort the calculated distances in ascending order based on distance values
   - Get top k rows from the sorted array
   - Get the most frequent class of these rows
   - Return the predicted class

RESULT AND DISCUSSIONS

After applying extra tree classification algorithm for feature selection on pre-processed datasets, 16 most important features are derived which are to be considered for training and testing the ECG model. The dataset is divided into a standard form where 80% of data is used for training and and other 20% for testing data.

Among 16 different features, some of them with their distribution pattern are shown below:

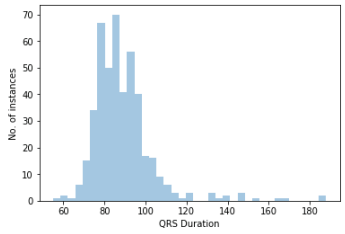

Fig. 2. Distribution of QRS Duration

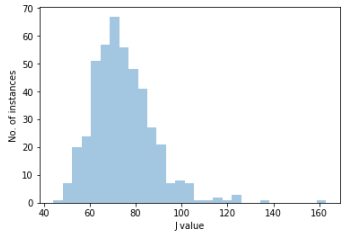

Fig. 3. Distribution of J value

D. ECG Classification Evaluation

KNN is non parametric(it means that it does not make any assumptions on the underlying data distribution) and lazy learning algorithm(it does not use the training data points to do any generalization). KNN is less computationally intensive, its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

As this paper uses dataset of multiple-classes which is not exponentially large and cannot be generalized, KNN comes to provide all these dimensions efficiently. Hence KNN algorithm.

This paper uses Confusion Matrix as an evaluation method which contains the information about actual and predicted classification done by the classifier. It simply allows visualization of the performance of any algorithm.

	NOR	CAD	OAMI	OIMI	ST	SB	PVC	SPC	LB	RB	LVH	AF	Oth
NOR	231	12	0	9	3	6	2	1	0	8	2	2	11
CAD	4	29	0	2	0	0	0	0	0	0	0	0	0
OAMI	0	0	15	0	0	1	1	0	0	0	0	1	1
OIMI	1	0	0	3	0	0	0	0	0	0	0	0	2
ST	0	0	0	0	9	0	0	0	0	0	0	0	0
SB	1	1	0	0	0	17	0	0	0	0	0	0	1
PVC	0	0	0	0	0	0	0	0	0	0	0	0	0
SPC	0	0	0	0	0	0	0	1	0	0	0	0	0
LB	0	0	0	0	0	0	0	0	9	0	0	0	0
RB	6	2	0	1	1	1	0	0	0	42	1	0	1
LVH	0	0	0	0	0	0	0	0	0	0	1	0	0
AF	1	0	0	0	0	0	0	0	0	0	0	2	0
Oth	1	0	0	0	0	0	0	0	0	0	0	0	6

Fig. 4. Confusion matrix of ECG Classification model

Under this classification model, the accuracy of the system is 76.92%.

E. Comparison with KNN via scikit-learn library

The supervised machine learning algorithm, KNN under consideration via scikit-learn library gave a comparable accuracy of 76.67%. In this instances, dataset was divided into 361 training sets and 91 testing set. Comparing again with support vector machine (SVM) algorithm accuracy was found to be 71.428%.

IV. Conclusion

This paper has successfully implemented extra tree classifier to select features from the dataset obtained from UCI Repository. KNN Classification approach has been implemented to train and test the model and classify the user's input data. The classified result is presented in human readable form. The accuracy of system is comparable with the model built via scikit-learn library. However, changing the value of k impacts the overall accuracy of the system. Choosing lower value of k overfits the model whereas higher value of k affects the class with lesser data, so choosing k=4 seems appreciable for the system resulting to the accuracy of 76.92%.

Acknowledgment

We acknowledge Head of Department of Kantipur Engineering College Er. Rabindra Khati & our respected teacher Er. Dipesh Shrestha for their continuous supervision, support and regular feedback. And finally we are grateful to our family and friends for their encouragement in getting us complete this project within the stipulated time.

References

[1] AliveCor, "Alivecor heart monitor and aliveecg app (kardia mobile) for detecting atrial fibrillation," 2015. [Online]. Available: https://www.nice.org.uk/advice/mib35/chapter/technology-overview

[2] A. Jovic and N. Bogunovic, "Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features," Artificial Intelligence in Medicine, vol. 51, no. 3, pp. 175–186, 2011.

[3] E. H. Houssein, M. Kilany, and A. E. Hassanien, "Ecg signals classification: a review," International Journal of Intelligent Engineering Informatics, vol. 5, no. 4, pp. 376–396, 2017.

[4] Y. Zhu, "SVM classification algorithm in ecg classification," in Information Computing and Applications, C. Liu, L. Wang, and A. Yang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 797–803.

[5] N. Deepika, "Association rule for classification of heart-attack," vol. 20, no. 3, pp. 45–50, 2001.

[6] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ecg morphology and heartbeat interval features," IEEE transactions on bio medical engineering, vol. 51, no. 7, pp. 1196–1206, 2004.

[7] D. Dua and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml