OPTIMIZATION OF PREDICTION AND CLASSIFICATION OF CARDIAC ARRHYTHMIA

Submitted By:

NIKHIL ARORA (9916103199)

DIVYANSH SRIVASTAVA (9916103020)

Under The Supervision Of:

RUPESH KUMAR KOSHARIYA

DEC- 2019

Submitted in partial fulfilment of the Degree of

Bachelor of Technology

in

Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

TABLE OF CONTENTS

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Noida Divyansh Srivastava (9916103020)

Date: 16 December 2019 Nikhil Arora (9916103199)

CERTIFICATE

This is to certify that the work titled "**Optimization of Prediction and Classification of Cardiac Arrhythmia"** submitted by Divyansh Srivastava and Nikhil Arora in partial fulfilment for the award of degree of B Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor:

Name of Supervisor : Mr. Rupesh Kumar Koshariya

Designation : Assistant Professor (Grade 1)

Date :

ACKNOWLEDGEMENT

First and foremost, we would like to thank our guide Mr. Rupesh Kumar Koshariya– Assistant Professor (Grade 1) , Jaypee Institute of Information Technology, Noida for guiding us thoughtfully and efficiently throughout this project, giving us an opportunity to work at our own pace along our own lines, while providing us with very useful directions whenever necessary. We would also like to thank our friends and classmates for being great sources of motivation and for providing encouragement throughout the length of this project. We offer our sincere thanks to all other persons who knowingly or unknowingly helped us in this project.

Signatures of Students

Nikhil Arora (9916103199)

Divyansh Srivastava (9916103020)

SUMMARY

It has been Cardiac arrhythmias are disturbances in the normal rhythm of the heartbeat. Any interruption to the electrical impulses that cause the heart to a contract can result in arrhythmia. Now, there are two mostly found types of arrhythmia first is Tachycardia (heartbeats:above 100 per minute) and Bradycardia (heartbeats: below 60 per minute). An arrhythmia occurs when the electrical signals to the heart that coordinate heartbeats are not working properly. For instance, some people experience irregular heartbeats, which may feel like a racing heart or fluttering. Many heart arrhythmias are harmless; however, if they are particularly abnormal, or result from a weak or damaged heart, arrhythmias can cause serious and even potentially fatal symptoms. Some patients have no symptoms, but a doctor might detect an arrhythmia during a routine examination or on an EKG. Even if the patient notices symptoms, it does not necessarily mean there is a serious problem; for instance, some patients with life-threatening arrhythmias may have no symptoms while others with symptoms may not have a serious problem. Hence it is extremely important to correctly predict and classify arrhythmia for efficient and accurate treatment.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

S.No	TITLE
1	ECG - Electrocardiogram
2	SVM - Support Vector Machines
3	KNN - K-nearest Neighbour
4	DBSCAN - Density Based Spatial Clustering of Applications with Noise

1. Introduction

1.1 General Introduction

Heart disease is one of the major problems world-wide. Timely detection of heart disease along with proper medical help can save lives. ECG or EKG (Electrocardiogram) is one of the most important tools for heart functions diagnosis. It generates a graphic record of heart's electrical impulses.

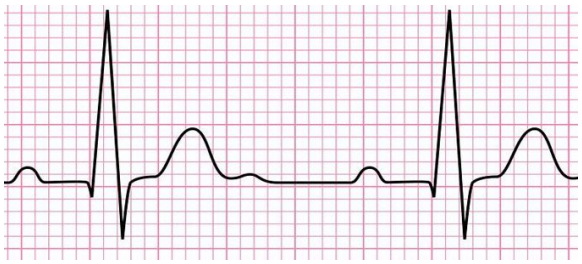


Fig. 1. Normal ECG

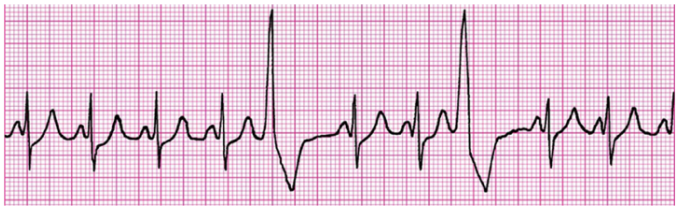


Fig. 2. Abnormal ECG

Arrhythmia describes an irregular heartbeat. Heart arrhythmias, also known as cardiac dysrhythmia, is a collection of conditions in which the heart beats irregularly. Dysfunction of the electrical signals to the heart that coordinates the heartbeats causes arrhythmia. In arrhythmia heart beats irregularly, that is either too fast or too slow. In both cases it can be harmful. For instance, some people experience irregular heartbeats, which may feel like a racing heart or fluttering. Most of the arrhythmias are harmless to the human body, however, if they are particularly abnormal, or are a result of a weak or damaged heart , arrhythmias can cause serious and even potentially fatal symptoms. The heart rate of a healthy person, when resting should be between 60 to 100 beats per minute. The more fit a person he lower their resisting heart rate. The diagnosis of arrhythmia involves handling of huge

amount of ECG data which poses the risk of human error in the interpretation of the data. If a health care provider failed to diagnose a severe case of arrhythmia, they can even be held liable for medical malpractice. Hence, computer assisted analysis of the ECG data and arrhythmia detection and classification can play a huge role as a decision support system to the doctors. Below are some facts which shows how important the problem is:

- A study showed that up to 23% of misdiagnosed heart attacks were due to the improper reading of a patient's ECG.
- A study showed that as compared to electrocardiographs, the physicians misread as normal about 36% of abnormal T waves.
- A study showed that out of 1.5 million heart attacks occurring every year in United States, 11000 cases are because of misdiagnoses.
- About 4.7 billion dollars paid in 2008 to resolve all malpractice claims nationwide

1.2 Problem Statement

Diagnosis of arrhythmia can be some times difficult. Some arrhythmias are harmless while others can prove to be fatal. Doctors study the ECG of a patient to find whether a patient is suffering from any kind of arrhythmia or not. So the aim is to make this process more swift and smooth by making use of machine learning. But without good accuracy prediction and classification will do no good. Hence it is extremely important to correctly classify arrhythmias as it can be a matter of life and death.

1.3 Significance of the Problem

Prediction of cardiovascular diseases is tricky, it should be accurate. There is no space for error be it human or machine error. Many intelligent automated decision support system has been implement to tackle the problem. Most of them are based on contemporary machine learning techniques which provide below average performances. Which is not

tolerable in the field of medical where it can be a matter of life and death. Thus better solution approaches are need of the hour.

1.4 Empirical Study

Studying ECG can be time-consuming, minute details are easy to be missed. Here comes machine learning into the picture. We can collect relevant data about the patient such as age, weight, height, sex similarly we can extract useful features from the patient's ECG. All this information will be used for prediction and classification of cardiac arrhythmia. Some of the machine learning algorithms already applied are Naive Bayes, SVM, Random Forests and Neural Networks.

These machine learning models have been implemented and tested on the UCI machine learning repository's arrhythmia dataset. Python being the most common tool for creating the models. R and Matlab have also been used in some cases.

1.5 Brief Description of the Solution Approach

A good dataset is the key to a good prediction model. As seen in our study much of the importance is given to the models and their comparison. Contrary to which we have focused on both the model and the dataset. A good amount of time has been spent to correctly preprocess the dataset.

Preprocessing the dataset was the first step of our solution. Unnecessary features and missing data are eliminated. Feature selection algorithms namely recursive feature elimination is applied to obtain most important features. These features are then used to train the models for prediction and classification.

1.6 Comparison of Existing Approaches to the Problem Framed

Many decision support systems have been implemented for prediction and classification of cardiac arrhythmia using different learning models. We

have implemented few of the supervised machine learning techniques resulting in astonishing outcomes in some cases.

In paper 3 accuracies obtained for SVM and random forest are 66% and 72% respectively whereas for the same algorithms we were able to get accuracies of 70% and 76% respectively.

2. Literature Survey

2.1 Summary of Papers Studied

Name : Cardiac Arrhythmia Classification Using Neural Networks with Selected Features

In this paper classification of arrhythmia is done using three standard machine learning algorithms namely OneR (One Rule), J48 and Naive Bayes. A detailed comparison between the three mentioned techniques is carried out. The comparison is based on mainly three principles predictability, accuracy and ease-of-learning of these algorithms.

WEKA (Waikato Environment for Knowledge Analysis) software environment for machine learning has been used for data analysis and classification of cardiac arrhythmia. The dataset has been obtained from the machine learning dataset archives at the University of California, Irvine.

A comparison of learning time i.e the time taken to build the algorithm and the accuracy i.e the number of correctly classified instances on the dataset between the three machine learning algorithms is illustrated in Fig 3 and Fig 4 respectively.

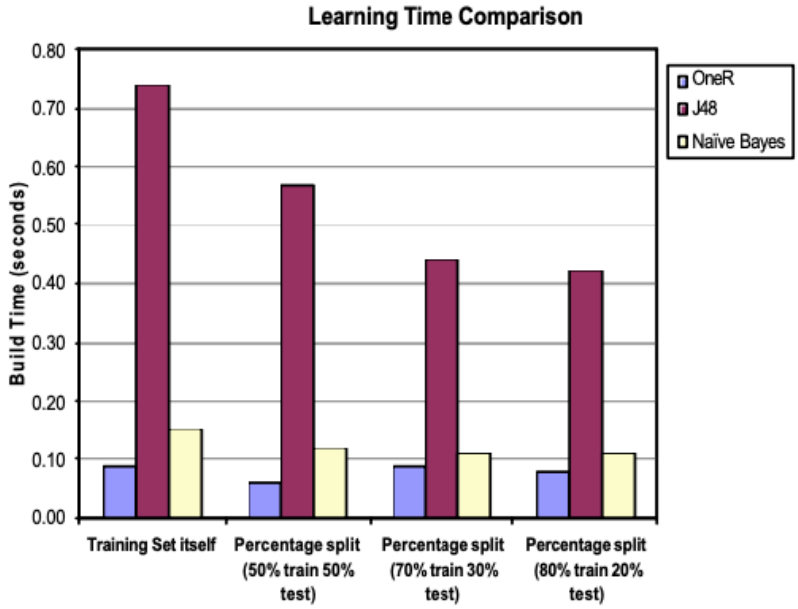


Fig. 3. Learning time comparison

It is concluded that the accuracy rate of OneR and Naive Bayes are more stable that J48. A proposal has been made, repeating the experiment using other machine learning algorithms such as support vector machines, as a future scope.

Name : Classification of Arrhythmia Using Machine Learning Techniques

Implementation of neural networks with feature selection for classification of cardiac arrhythmia is presented in this paper. Incremental Back propagation Neural Network (IBPLN) and Levenberg-Marquardt (LM) classification has been used. The classifiers are tested on the UCI arrhythmia data base.

Correlation based feature selection (CFS) technique is used for feature extraction and reduction. As a result of which a reduced set of 18 features is obtained. The data is partitioned into three sets training, validation and test as a data preprocessing step. Classification accuracy, specificity and sensitivity were used as the parameters to compare classification results. Worst simulation results, best simulation results and an average of 100 simulations were the main focus of the study. Fig. 5 shows a block diagram of the steps involved in forming the decision system for arrhythmia classification using neural networks.

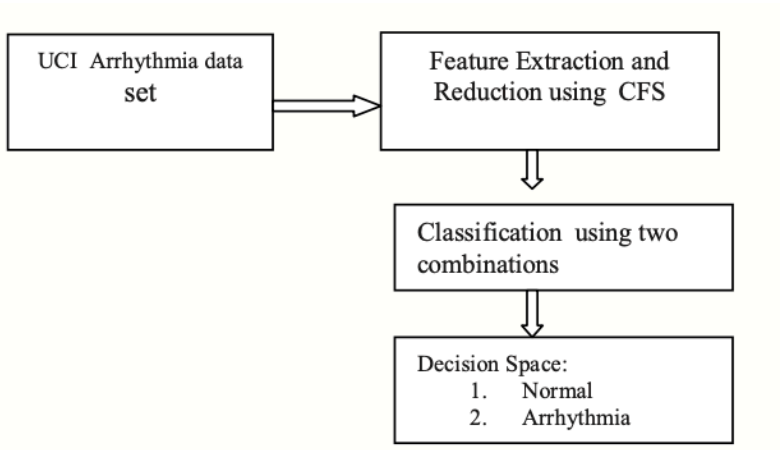


Fig. 4. Block diagram of arrhythmia classification

Name : Prediction and Classification of Cardiac Arrhythmia

In this paper the authors have used dataset from the UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Arrhyth-mia (1 CSV file, 1 information file). They extracted different parameter values from ECG waveforms and used other information about the patient like age, medical history, etc to detect arrhythmia. They have implemented a few popular prediction and classification techniques namely Naive Bayes, SVM, Random Forests and Neural Networks. The dataset used is highly biased towards no arrhythmia. Feature reduction is also done as a part of data preprocessing. A combination of SVM and Random Forest has also been used. A serial classifier consisting of RF and linear kernel SVM are fused to get better accuracy.

Name : Radial basis function Neural Network for Prediction of Cardiac Arrhythmias based on Heart rate time series

In this paper author has used dataset from MIT–BIH arrhythmia database as standard database which contains 48 half-hour ECG recording of 47 ambulatory subjects. The samples are recorded with sampling frequency of 360 Hz and digitized by analog to digital converter with 11 bit resolution. Now, this paper classifies arrhythmia into one of the eight classes namely Left bundle branch block (LBBB), Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Right bundle branch block (RBBB), Sinus bradycardia (SBR), Atrial flutter (AFL), Premature Ventricular Contraction (PVC), and Second degree block (BII). Time series analysis of heart rate is done for the prediction of the disease. ECG data is preprocessed to detect QRS complexes and RR Interval, which is then fed into Radial Basis Function Neural Network. The four non-linear parameters studied for cardiac arrhythmia prediction in this paper are Poincare plot geometry, Spectral entropy, Largest Lyapunov exponent and Detrended fluctuation analysis.

Name : Cardiac Arrhythmia Classification Using Fuzzy Classifiers

In this paper author has used dataset from MIT–BIH arrhythmia database as standard database which contains 48 half-hour ECG recording of 47 ambulatory subjects. The samples are recorded with sampling frequency of 360 Hz and digitized by analog to digital converter with 11 bit resolution. In this paper authors have used fuzzy classifiers to classify arrhythmia using ECG data. Features are extracted from ECG signals using multi-resolution wavelet transform. The authors have computed the non-linear parameters of ECG signals for the entire dataset. Fuzzy classification allows multiple conclusions to exist which is an advantage over more deterministic algorithms. The proposed classifier is made up of two main blocks ECG parameterizer and Fuzzy classifier.

Name : Comparative study of Neural Networks for Prediction of Cardiac Arrhythmias

In this paper author has used dataset from MIT–BIH arrhythmia database as standard database which contains 48 half-hour ECG recording of 47 ambulatory subjects. The samples are recorded with sampling frequency of 360 Hz and digitized by analog to digital converter with 11 bit resolution. In this paper authors have presented a comparison between Radial basis function neural network(RBFN) and Multi-layer perceptron(MLP).

The standard MIT-BIH arrhythmia database has been used. Each ECG signal record has been investigated for irregular rhythms. The outcomes and results show that RBFN is more effective. Its prediction accurate is more than that of MLP.

Name : Implementation of the Objective Clustering Inductive Technology Based on DBSCAN Clustering

In this paper authors have used DBSCAN clustering algorithm within the framework of the objective clustering inductive technology. They have used the data Aggregation and Compound of the Computing school of the East Finland University and the gene expression sequences of lung cancer patients of the database. They have developed the architecture of the

objective clustering algorithm. By this paper they try to solve the problem of reproducibility error, in other words, satisfactory clustering results obtained on one dataset do not repeat while using another similar dataset. They presented the practical implementation of this technology based on the k-means and agglomerative hierarchical clustering algorithms. The key condition for implementing this technology successfully was the careful selection of the external and internal clustering quality criteria, which take into account both the character of the objects grouping within clusters and the character of the clusters distribution in the studied space. Below image is taken from the research paper to show whole process they go through for implementing DBSCAN algorithm on the database.

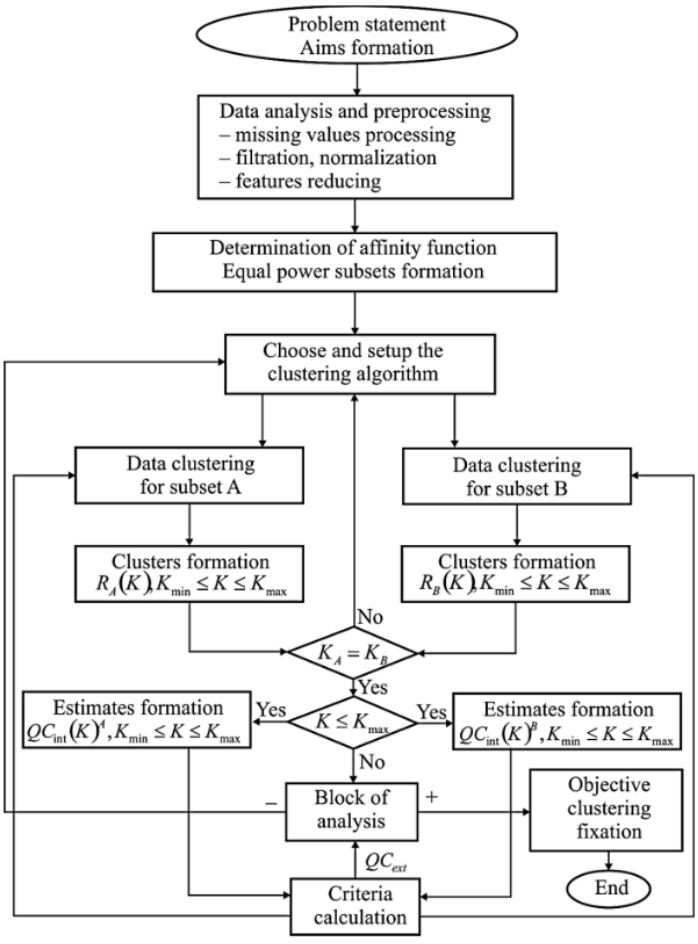


Fig. 5. Implementation Flowchart

Before reaching to the result and conclusion we need to understand some the major definition they told in paper to understand how the DBSCAN is working here.

DEFINITION 1: Eps-neighbourhood of a point p is the neighbourhood, which is defined as following:

EPS(p)={q belongs to D|dist(p, q)<=EPS},

Where dist(p, q) – is the proximity distance between the points p and q.

DEFINITION 2: The point q is directly density-reachable from the point p if

1) q belongs to EPS(p);
2) Neps (p) <=MInPts

Where Neps(p) – is the quantity of points within of Eps-neighbourhood of the point p, MinPts – the least quantity of the points within the Eps-neighbourhood of the point p.

Using above two definition we formed the clusters which can be seen in below images (taken from the research paper.

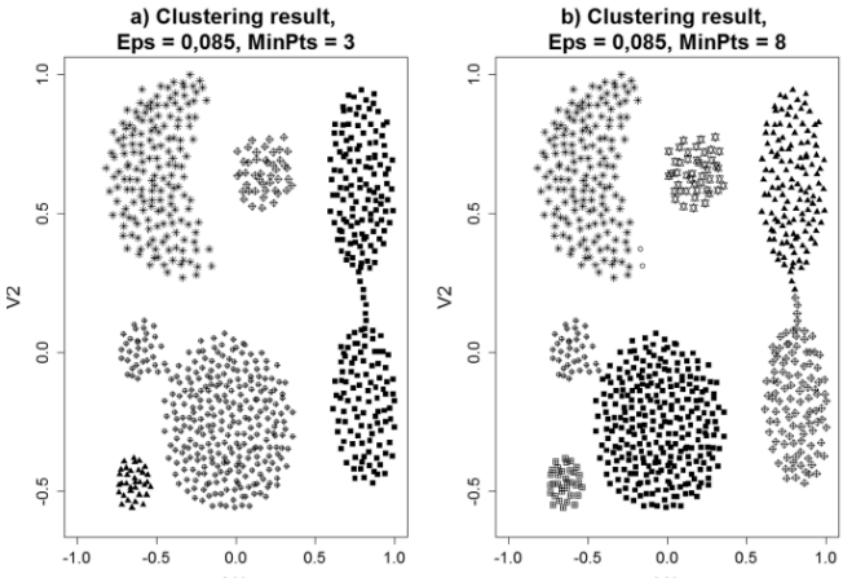


Fig. 6. Results of the Aggregation data clustering in case of Eps = 0,085 and a) MinPts = 3; b) MinPts = 8

Name : Feature identification from imbalanced data sets for diagnosis of Cardiac Arrhythmia

In this paper authors used dataset from UCI machine learning repository. They have tried to select the best feature from the imbalanced dataset. The sample number of this data set is 452, represented by " i ", i= 1,2,...,452. Each sample contains 279 feature attributes, in which the first 4 feature attributes are gender, age, height, weight, and the other feature attributes are from the ECG records. Each attribute is represented by "j ", j= 1, 2,...,279. Rank the attributes by the order of original data set, namely, 1 f means age, 2 f means gender, 3 f means stature, ... , and so on. The data set is divided into 16 classes in the data set, of which the first class refers to healthy patients, classes 2-15 refer to 14 kinds of cardiac arrhythmia, and the rest kinds of cardiac arrhythmia are classified as class 16. Below table is taken from research paper showing different categories and sample size of arrhythmia dataset.

The main thing here done is data preprocessing which leads to optimization of the algorithms. There full methods which are used to preprocess the dataset are:

1) Fill the missing data. The mean value of the attribute in the patient' class is used to fill the missing data.

2) Balance the dataset. According to the sample sizes in the data set, class 1 is taken as the majority group, and the rest classes are regarded as the minority groups. The sample number of the majority group is much larger than the minority groups. The equalization process is as follows:

    a. Handle the majority group by using the random under-sampling method: x samples are removed from the normal sample data set, which are random selected from class 1.

    b. Oversample the minority groups by using SMOTE method. The minority groups are oversampled by taking each

minority group sample and introducing synthetic examples along the line segments joining any of the k minority group nearest neighbors, respectively. According to the sample number of majority class and minority classes in the data set of Arrhythmia. Considering the prediction accuracy and the sample number of minority classes, the amount of each class is set to 60. The majority group is under-sampled by random under-sampling method, namely, 60 samples selected from 245 normal samples randomly, and the minority classes are over-sampled by the SMOTE method, respectively. The obtained samples are recombined to form a new data set and a standardize decision matrix is obtained, after the characteristics value of the new data set are standardized and the classification of disease diagnosis is centralized.

They have also used 10-fold cross validation to train the classifier, which means the dataset is randomly divided into 10 parts. A fold of the data is selected as the test set, and the remaining 9 folds are used as the training set, and subsequently the trained models are employed to forecast the test set. Repeat for 10 times, and the accuracy of classification prediction can be obtained by averaging the predict results.

2.2 Integrated Summary of the Literature Studied

Research Paper	Authors	Summary
Prediction and Classification of Cardiac Arrhythmia	Vasu Gupta, Sharan Srinivasan, Sneha S Kundli	They have implemented a few popular prediction and classification techniques namely Naive Bayes, SVM, Random Forests and Neural Networks.
Radial basis function Neural Network for Prediction of Cardiac Arrhythmias based on Heart rate time series	J. P. Kelwade	Time series analysis of heart rate is done for the prediction of the disease. ECG data is preprocessed to detect QRS complexes and RR Interval, which is then fed into Radial Basis Function Neural Network.
CARDIAC ARRHYTHMIA CLASSIFICATION USING FUZZY CLASSIFIERS	Mrs. B.Anuradha, V.C.Veera Reddy	The authors have computed the non-linear parameters of ECG signals for the entire dataset. Fuzzy classification allows multiple conclusions to exist which is an advantage over more deterministic algorithms.
Comparative study of Neural Networks for Prediction of Cardiac Arrhythmias	J. P. Kelwade	The author have presented a comparison between Radial basis function neural network(RBFN) and Multilayer perceptron(MLP).
Implementation of the Objective Clustering Inductive Technology Based on DBSCAN Clustering Algorithm	S. Babichev, V. Lytvynenko, V. Osypenko	In this paper authors have used DBSCAN clustering algorithm within the framework of the objective clustering inductive technology. They have used the data Aggregation and Compound of the Computing school of the East Finland University and the gene expression sequences of lung cancer patients of the database.

Feature identification from imbalanced datasets for diagnosis of Cardiac Arrhythmia	Liang Lijun, Jin Tingting, Huo Meiya	In this paper authors used dataset from UCI machine learning repository. They have tried to select the best feature from the imbalanced dataset. The sample number of this data set is 452.
Classification of Arrhythmia Using Machine Learning Techniques	THARA SOMAN PATRICK O. BOBBIE	In this paper they have used machine learning schemes, OneR, J48 and Naïve Bayes to classify arrhythmia from ECG medical data sets. The aim of the study is to automatically classify cardiac arrhythmias as a part of our ongoing embedded medical device research, and to study the performance of machine learning algorithms.
Cardiac Arrhythmia Classification Using Neural Networks with Selected Features	Malay Mitra and R. K. Samanta	This research present a new approach for cardiac arrhythmia disease classification. In this connection, intelligent automated decision support systems have been attempted with varying accuracies tested on UCI arrhythmia data base. One of the attempted tools in this context is neural network for classification.

3. Requirement Analysis and Solution Approach

3.1 Overall Description of the Project

The proposed decision support system classifies a person into one of the arrhythmia or non arrhythmia classes. Logistic regression, SVM, Random forest and KNN are used as the tools for classification of arrhythmia data. Dataset preprocessing is done by appropriately cleaning the dataset and then selecting only the most important features for further process. A shallow comparison between prediction and classification accuracies of the mentioned algorithms is done.

3.2 Requirement Analysis

Functional Requirements

1. Software Requirements
    - Python 3.6
    - Jupyter Notebook
    - Anaconda Navigator
    - Operating System : Windows/MacOS/Linux
2. Hardware Requirements
    - Processor intel i3 or higher
    - RAM 4GB or more
    - Disk Space 1GB or more

Non-Functional Requirements

- Reliability: The system should provide high accuracy and performance without failures.
- Usability: The system setup should be easy to implement. Codes written for the processing of the system should have comments for increasing readability for future development.
- Performance: The system should work accurately on all cases. The system should result in minimum efficient run time.

3.3 Solution Approach

The proposed solution aim at providing a good cardiac arrhythmia classification by training and testing various machine learning algorithms namely Logistic Regression, SVM, Random Forest Classifier and K Nearest Neighbours on the dataset obtained form UIC machine learning repository. Feature selection is an important part of the solution. Raw data needed to be preprocessed before any of the feature selection or learning models can be applied.

Chapter - 4 Modeling and Implementation Details

4.1 Design Diagrams

4.1.1 Use Case Diagram

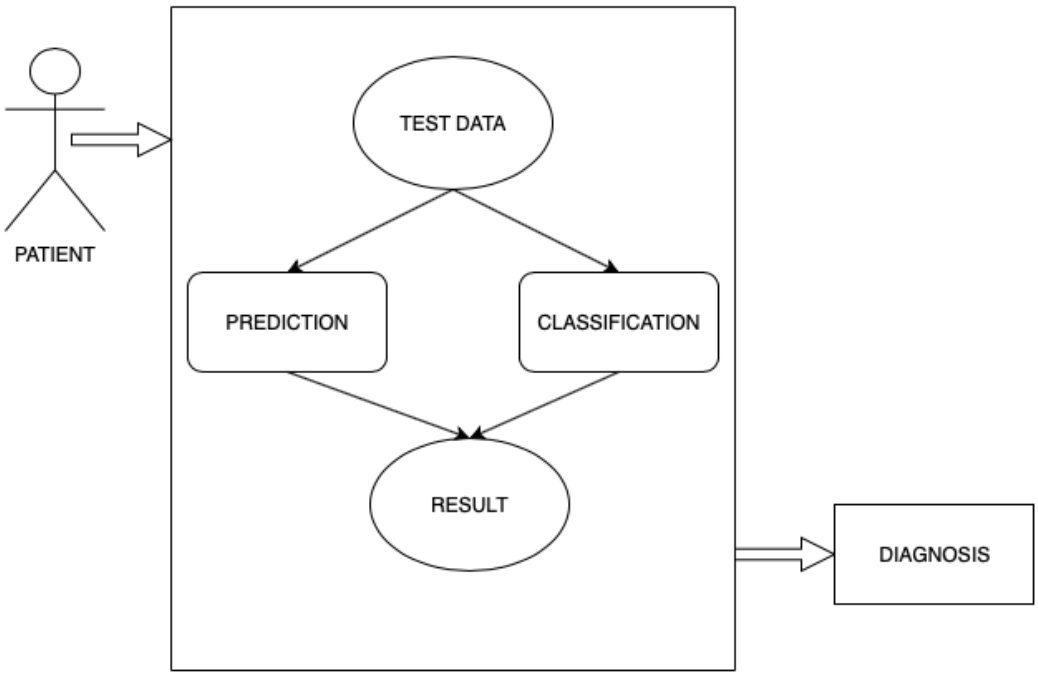
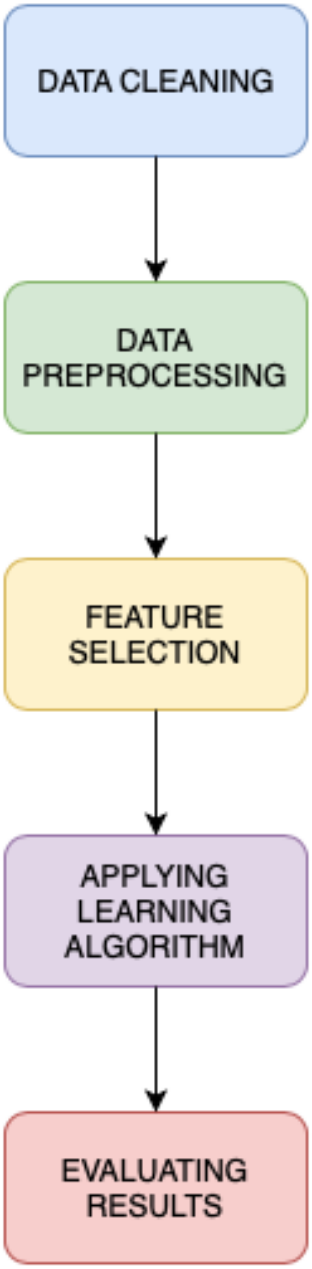


Fig. 7. Use case diagram

4.1.2 Flow Diagram



•Unnecessary attributes such as single valued columns are eliminated.

•Missing data is replaced with the column mean.

•Features are normalised.

•Dataset is spliced into train and test sets.

•Feature selection algorithms are applied.

•A vote for the best feature is taken from the algorithms used.

• Learning models are trained on the dataset.

•Accuracy of the trained model is calculated on the test set.

Fig. 8. Flow diagram

4.2 Implementation Details and Issues

The dataset consists of ECG readings and a few physical description of 451 patients. There are a total of 279 attributes, 206 linear and 73 nominal and a single output which is categorical. Each patient is classified into one of the 16 classes based on their attributes.

Sr. No.	Attribute	Type	No. of attributes	Remark
1	Age, Weight, Height	Linear	3	Physical description of patient
2	Sex	Nominal	1	Sex of the patient
3	Duration of waves	Linear	5	Measured in milliseconds
4	Vector Angles	Linear	5	Measured in degrees
5	Heart Rate	Linear	1	No. of heart beats per minute
6	Average width of wave	Linear	60	Measured in milliseconds
7	Number of intrinsic deflections	Linear	12	
8	Existence of ragged wave	Nominal	36	
9	Existence of biphasic derivation	Nominal	36	
10	Amplitude of wave	Linear	96	Multiplied with 0.1 x multi-volt
11	QRSA, Areas of all segments divided by 10	Linear	12	Area = width x height/2
12	QRSAT	Linear	12	QRSA + 0.5 x width of T wave x 0.1 x height of T wave

Table. 1. Dataset description

In table 2 describes all the class labels and their number of instances present in the original dataset. Form the table it can be inferred that number of instances for non-arrhythmia classes is very less as compared to the arrhythmia class.

Class Code	Class	Number of instances
1	Normal	245
2	Ischemic changes (Coronary Artery Disease)	44
3	Old Anterior Myocardial Infraction	15
4	Old Interior Myocardial Infraction	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	1. degree AtrioVentricular block	0
12	2. degree AV block	0
13	3. degree AV block	0
14	Left ventricule hypertrophy	4
15	Atrial Fibrillation or Flutter	5
16	Others	22

Table. 2. Class distribution

ECG Wave

ECG wave, its components and the meaning of each component of the wave is shown below.

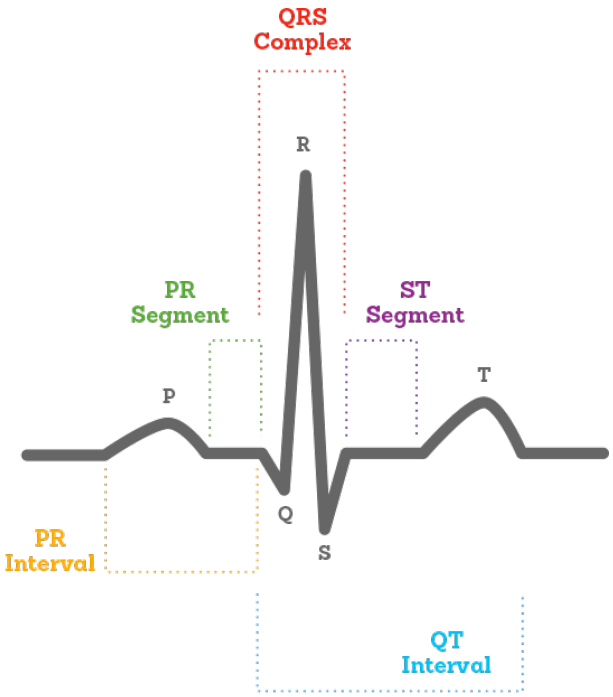


Fig. 9. ECG wave

Sr. No.	Predictor	Description
1	P Wave	The P wave represents atrial depolarization.
2	PR interval	The PR interval is measured from the beginning of the P wave to the beginning of the QRS complex.
3	QRS complex	The QRS complex represents ventricular depolarization.
4	J-point	The J-point is the point at which The QRS complex finishes and the ST segment begins.
5	ST segment	It represents the period when the ventricles are depolarized.
6	T wave	The T wave represents ventricular repolarization.

Table. 3. ECG wave description

Steps taken for the implementation of the proposed solution are as follows:

STEP 1:

Dataset is imported using the pandas library. As the first step of data cleaning all the single valued columns are found and removed. These columns are removed as they will provide no variation to the output as the remain same for each and every instance. 17 attributes were found to have same value for all the data points.

STEP 2:

In the next step, missing values from the dataset were identified. 5 columns were observed to have missing values.

Attribute	Attribute	No. of missing values
1	J_Angle	375
2	P_Angle	22
3	T_Angle	8
4	QRST_Angle	1
5	Heart	1

Table. 4. Attributes with missing values

The columns J_Angle and P_Angle were deleted as the had significantly large number of missing values.

STEP 3:

The remaining missing values in the columns T_Angle, QRST_Angle and Heart were filled by the mean of the respective columns.

STEP 4:

The dataset is then split into train and test sets with a ratio of 80 to 20. The features are scaled using Standard Scaler module of Scikit-learn library. The standard score of a sample x is calculated as:

$$z = (x - u) / s$$

where, u is the mean of the training samples and s is the standard deviation f the training samples.

STEP 5:

It is observed that classes 7, 8, 9, 14 and 15 had extremely low instances which are not enough for a good classification. Thus these cases are deleted. Class 16 is also deleted as it contains unlabelled classes.

Class	No. of instances
7	3
8	2
9	9
14	4
15	5

Table. 5. Classes with extremely few instances

STEP 6:

In this step recursive feature elimination feature selection algorithms is applied to the dataset namely to select most important features for prediction. 60 features are selected in this step.

STEP 7:

Logistic Regression, SVM, Random Forest and KNN algorithms are applied to the processed dataset in this step. Prediction and classification models are learned separately.

STEP 8:

Accuracies of the models are analysed to find the best algorithm for prediction and classification.

Implementation Issues

- Low accuracy on unprocessed data.
- Many irrelevant attributes, explaining no variation between data points and response.
- Few columns with extremely large umber of missing values.
- Number of attributes being very large, it was difficult to plot relation between them.

5. Testing

5.1 Testing Plan

Type of Test	Has it been performed	Explanation	Software Components
Requirement	Yes	Requirement specification must contain all the requirements that are needed by our system.	Manual work need to plan out all the software requirements, time needed to develop, technology to be used etc.
Unit	Yes	Testing technology using which individual modules re tested to determine if there are any issues , by the developer itself.	Manual check is required.
Integration	Yes	Testing where individual components are combined and tested as a group.	Compiling full code and testing it together.
Performance	Yes	Testing to evaluate the input where the best and most optimal output is yielded by the system.	Testing results ensure this.

Table. 8. Testing plan

5.2 Component Decomposition and Type of Testing

Sr. No.	List of modules which require testing	Type of test requires	Techniques used for writing test cases
1	Feature Selection	Requirement, Unit, Performance and Integration	White box
2	Logistic Regression		White box
3	SVM		White box
4	Random Forest Classifier		White box
5	KNN		White box

Table. 9. Components and type of testing

5.3 Test Cases

To test how well the prediction and classification algorithms are performing on the cardiac arrhythmia data the dataset is split into train and test sets in the ratio of 8 : 2.

1. Prediction

Sr. No.	Algorithm	Test Accuracy(%)
1	Logistic Regression	74
2	SVM	81
3	Random Forest	81
4	KNN	69

Table. 10. Prediction accuracies

2. Classification

Sr. No.	Algorithm	Test Accuracy(%)
1	Logistic Regression	74
2	SVM	70
3	Random Forest	76
4	KNN	73

Table. 11. Classification accuracies

5.4 Limitations of the Solution

The system performed well for prediction of arrhythmia, and results were above average for classification. But still classification model couldn't classify many of the arrhythmia classes correctly.

6. Findings, Conclusion and Future Work

6.1 Findings

The proposed model design is found efficient in classifying cardiac arrhythmia. Feature selection improved the system accuracy. Most of the misclassified instances are classified into class 1(normal) which is highly undesirable. Since a large number of features are used to fit a model on a very small number of observations, the model suffers from high variance. Feature selection deals with much of the variance.

It is evident that while SVM and Random forest performed extremely well in case of prediction, KNN proved to be better for classification where as Logistic regression's performance was more or less same.

6.2 Conclusion

The aim of this project was to detect and classify cardiac arrhythmia. With the given given data which is highly imbalanced and heavily biased towards class 1, obtaining a very accurate classification of all arrhythmia classes is not feasible. But since we have managed to obtain a descent model with descent accuracy, it can be conclude that he model will definitely perform better if trained on a bigger at a set hat has more number of rare class instances.

It is observed that personal characteristics such as age, weight, height etc did not make any considerable contribution to the classification. As only the ECG readings are used as features, adding other features that a medical professional may utilize n his diagnosis may improve the performance of the model.

6.3 Future Work

More feature, important according to a medical professional, can be added to the dataset to obtain a better accuracy. Other models can also be tested on the dataset, and ensemble of models can be made. As the prediction accuracy is much better than classification accuracy, prediction model can

be used as a decision support system that may guide the professionals in confirming the presence of arrhythmia with a fairly high certainty.

For this project only the physical description and ECG extracted data is used as features. Processing the ECG wave itself can prove to be beneficial to obtain even more features that might add to the accuracy of the existing models.

References

1. Anuradha, B., and V. C. Reddy. "CARDIAC ARRHYTHMIA CLASSIFICATION USING FUZZY CLASSIFIERS." 2008 Journal of Theoretical & Applied Information Technology.

2. Gupta, Vasu, Sharan Srinivasan, and Sneha S. Kundli."Prediction and Classification of Cardiac Arrhythmia." 2014 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT). IEEE.

3. Kelwade, J. P., and S. S. Shankar."Comparative study of neural networks for prediction of cardiac arrhythmias." 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT). IEEE.

4. Kelwade, J. P., and S. S. Shankar. "Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series." 2016 IEEE First International Conference on Control, Measurement, and Instrumentation (CMI). IEEE.

5. Liang Lijun, Jin Tingting, Huo Meiya.“Feature identification from imbalanced datasets for diagnosis of Cardiac Arrhythmia.” 2018 11th International Symposium on Computational Intelligence and Design.

6. S. Babichev, V. Lytvynenko, V. Osypenko **“**Implementation of the Objective Clustering Inductive Technology Based on DBSCAN Clustering Algorithm.**”** 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT).

7. THARA SOMAN PATRICK O. BOBBIE “Classification of Arrhythmia Using Machine Learning Techniques.**”** 2018 Engineering Vibration, Communication and Information Processing.

8. Maya Mitra, R. K. Samanta “Cardiac Arrhythmia Classification Using Neural Networks with Selected Features.**”** 2013 International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) .

NIKHIL ARORA

MAIN OBJECTIVE

Looking for a career that offers innovation, excitement and challenges, where I can best utilize my skills and education.

EDUCATION

Jaypee Institute Of Information Technology, Noida
B.Tech(Computer Science Engineering)
CGPA : 8.4/10
Year : 2016 - 2020

St Xavier's School, Jhansi
12th || Year : 2015-2016
Percentage : 92.4

St Xavier's School, Jhansi
10th || Year : 2013-2014
Percentage : 93.5

GET IN TOUCH!

Mobile: +917309540686
Email: nikhilarora068@gmail.com
Address: Type 3/7 Sector-1 BHEL Township
Jhansi, Uttar Pradesh (284120)

SKILLS

- C/C++
- Data Structures
- Algorithms
- Django
- Html
- AWS

WORK EXPERIENCE

SUMMER TRAINEE

GEEKF FOR GEEKS May 2019 - Jun 2019

- Learnt in detail about various data structures and programming algorithms.
- Implemented the same in C++.

BACKEND DEVELOPER INTERN

IHEAL365 Dec 2018 - Jan 2019

- Worked as a part of the backend development team.
- Tech used - Django, google cloud platform and firebase.

PROJECTS

SANDTEX

An Online Collaborative Latex Editor

- This project is a web application that compiles the user written latex code and generates PDF for the same.
- Tech used - Django, HTML, CSS, Bootstrap and Firebase.
- Project is live at - http://sandtex.pythonanywhere.com

MYBLOGS

A Personal Blogging Website

- Developed a personal blogging web application where users can create, update and delete posts.
- Tech used - Django, HTML, CSS.
- Project is live at - http://mybloggingapp.herokuapp.com

ACTIVITIES

COMPETITIVE PROGRAMMING

- Codechef - https://www.codechef.com/users/nik068
- Hackerrank - https://www.hackerrank.com /nikhilarora068

SPORTS

- Organized basketball tournament at Converge 2018 (JIIT's annual cultural and sports fest).
- Participated in basketball tournament at Students Olympic National Games(2015-2016), affiliated to Students Olympic Association, India.

DIVYANSH SRIVASTAVA

Location: **Noida**
Phone: **8077726604**
Email: **divyansh.srivastavadps@gmail.com**
GitHub Profile Link: **https:/github.com/9916103020**
LinkedIn Profile Link: **https://www.linkedin.com/in/divyansh-srivastava-a2b87a176/**

PROFESSIONAL PROFILE

I am a CSE Undergraduate student from Jaypee Institute of Information Technology, Noida. Recently, I have been working in Guavus Company as an SDET-Intern and also developed many projects on different languages (as you can see using my GitHub Profile link) from scratch and one project which is still in working phase 'SANDTeX'. It is built using Django(Python Web-Framework) as back-end.

EDUCATION & QUALIFICATIONS

- **Jaypee Institute of Information technology, Noida -**(from 2016 – to present)
  - **CGPA (At Present) –** 7.5

- **Delhi Public School, Agra –** (from 2014 – 2016)
  - **10th (CGPA)–** 9.6
  - **12th Percentage –** 89.8

TECHNICAL COMPETENCIES

Operating System	**macOS(Beginner), Linux (Beginner), Windows XP/7/8/10**
Programming &Scripting Language	**C, C++, Python(Beginner), Django (python Web-Framework)**
Editors (Hands-On)	**CodeBlocks, Pycharm**
Working Knowledge	**Jenkins, Apache Hadoop MapReduce, MySQL, Heroku, PythonAnywhere, Microsoft Excel, Microsoft PowerPoint, Microsoft Word**
Languages Known	**Hindi, English**

PROJECTS & WORK EXPERIENCE

Jun 2018 – present Guavus (A Thales Company)
SDET – Intern

1. Migration of code from python2 to python 3.
2. Knowledge about Jenkins and its working.
3. Knowledge about Hadoop Mapreduce and its working.
4. Knowledge about some Python Libraries which are used for migrating.

Jun 2018 – present Project - SANDTeX

This is an website which gives advantage to user for compiling latex file and obtaining output pdf file on same window. This website also removes the headache of finding supported operating system and other specification everything is already covered in website you just have to register and login then, everything will be done by us. This webiste also provide real time collaboration so that more than one user can work on same file simultaneously. This website is built with Django, HTML, CSS, JAVASCRIPT, and MiKTeX COMPILER.

Github Link: **-** https://github.com/9916103020/SandTex

Jul 2017 – Aug 2017 Project-Login/Signup System (Using Django)
This was my first project of Django to build a fully working login/signup system. It is also deployed on Heroku you can check it.
GitHub Link: - https://github.com/9916103020/Internship_Work
Website Link: - https://internwork.herokuapp.com/

ACHIEVEMENTS

- **College Fest (Volleyball Organising Committee head): - 2 years**
- **SNACKDOWN 2019 – Participation Certificate**
- **Independently Migrated Code from Python 2 compatible to Python 3 compatible**
- **Published An Article on Cyber Security on CyberXploit. Link:https://cyberxploits.com/2019/08/16/csrf-token/**
- **I also uploaded one of my video on youtube about how you can create two virtual environment (Python2 and Python3) on 1 OS (Windows). Link: https://www.youtube.com/watch?v=g50WxOUMqQ4&t=356s**

INTERESTS

Co-Curricular Activites: I am captain of our college Volleyball team and also plays in our college cricket team.

Interests: I am good in Algorithm and Data Structure (in C++). I also participated in many online competitions you can see my Codechef or Hackerrank profile.

Codechef Profile Link: - https://www.codechef.com/users/divy_1998

Hackerrank Profile Link: - https://www.hackerrank.com/divyanshsrivast3

REFERENCES

1. Mr. Anubhav Gupta (SDET, Guavus (a Thales Company)) Ph. 9910023237

2. Mr. Mayank Mahajan (Principal SDET, Guavus (a Thales Company)) Ph. 8800796608