Research on the Machine Learning Algorithms on Heart Condition Predictions

Geetha.M, Ganesan.R, Tallam Tharun Sai

***Abstract*— Now-a-days Health care monitoring widely uses Internet of Things (IoT) and big data which is further integrated into wearable bio sensors. This paper is about finding the best algorithm for predicting the heart condition using different machine learning algorithms. In this we have also included the basic Artificial Neural Networks algorithm for predicting the heart condition of an individual.**

In this work we had predicted the persons heat conditions by knowing some key attributes. By increasing the use of machine learning algorithms, the accuracy of each algorithm is calculated and the quality and value of the health services increases efficiency.

This is mainly about how different the algorithms predict and the accuracy of each algorithm. Here the ANN has the height accuracy when compared to all other machine learning algorithms like, SVM-ploy, SVM-RBF, Naïve Bayes, Decision tree, Random Forest, K-Nearest Neighbor.

Keywords — Decision Tree, K- nearest neighbour, Naive Bayes, Random forest, SVM Poly, SVM RBF, ANN (Artificial Neural Networks using multi-layer Perceptron).

I. INTRODUCTION

Machine Learning and analysis is the extraction of new information by examining large amount of data. It can be used to take certain decisions, estimate and predict using different algorithms. By using machine learning algorithms with IoT devices we can predict different diseases like Cardiovascular Diseases and many more by knowing the condition of the heart. These heart conditions can be predicted using different factors which includes family history(genetics), cholesterol level, diabetes, Blood Pressure, heart beat, age, alcoholic or not, exercise etc. So, the person's data is taken and analyzed to predict his/her heart condition.

A. Back-Ground Motivation

The heart is very important organ of our body. So, it's our duty to take care of our heart's condition. The rate of heart malfunctions is increasing enormously now-a-days. This is due to the busy lifestyle that people are facing. Mostly people are eating junk, fast food and doing their work from sitting in one place like software engineers, bankers and etc. And people are doing less exercise and are less active. Physical activities
are reduced. So, they are not at all taking care of their heart condition. These factors cause many people to suffer from heart diseases. People are busier to take the regular check-ups.

B. Objective

The objective of the paper is to collect the heartbeat data and train the data by a trained dataset called "Cleveland" dataset. This dataset is given by Hungarian and Switzerland doctors from different hospitals. It has 14 attributes and 300 data records in it. This project uses supervised machine learning techniques for the predictions. We will predict only the severity of the heart condition which is ranked from 0 to 4. This can be done by the taking the real data from the sources like Fitbit, Google Fit apps, etc. And adding an extra feature to it that, the heart condition.

II. LITERATURE SURVEY

There are many papers on this solution. Here only few are discussed.

In a paper the predictions are done by the algorithm called Decision Tree, one of the best machine learning algorithms. By this algorithm the author will combine different datasets to avoid allowing irregular details into the scenario. So, this is about finding the best data from which the predictions are made correctly. Only some features from the datasets are extracted and used in the prediction algorithm to get the best results.

In the other paper the importance of feature scaling and dimensionality reduction is explained clearly. This is finally shown by the results using simple machine learning algorithms

In other study the dataset is put into the J48 classifier Un-pruned tree, Decision tree, Naïve Bayes classification, Neural Networks. From which they are getting the accuracy of above 90%. Then they have used Sensitivity and Specificity rate for comparing. This model is done by using the Data Mining model called Transthoracic Echocardiography Report dataset.

The other paper, reflected that the complete study of different papers that have done the heart condition is using Data Mining Techniques and Intelligent Fuzzy Approach. This analysis is based on the number of attributes used in the dataset, accuracy and success of the models. All types of algorithms, tools used etc are taken into consideration and brought into the versus platform which tells the best model.

On further analysis and research, a paper showed how the dependencies of the different attributes in the dataset are responsible for the heart condition prediction. In this the author had taken the 4 different datasets of similar kind of

Revised Manuscript Received on April 12, 2019.
Geetha.M, India
Dr.Ganesan.R, India.
Tallam Tharun Sai, India.

attributes. This paper will tell that those attributes are not sufficient for predicting the heart condition. This have been tested by machine learning algorithms like Naïve Bayes, Decision Tree, SVM, Ada-Boost, LR, MLP etc. He also found about the selection of attributes by IG and Genetic algorithms for selecting the main attributes.

In other paper the comparative study about the different machine learning algorithms like SVM, Logistic regression, Neural networks, RBFs, etc. and how accurate their predictions are.

III. PROBLEM FORMULATION AND EXISTING WORK

[13]There are existing works like, comparing the results of different algorithms namely, Decision Tree, Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbor, Single Conjunctive Rule Learner, Radial Basis Function and Support Vector Machine. Here ensemble predictions of classifiers, which includes bagging, stacking, boosting, are applied to the dataset.

There are works like understanding of different datasets predictions. In these, relationships between the input and output attributes had been concentrated more.

[4] In the other paper the heart condition was predicted by the neural networks and neuro-fuzzy system. They have used CRIP-DM methodology to the Cleveland dataset for predictions. This is done by establishing the relationships, patterns connected with heart diseases. By which they are getting an accuracy of 75.93%.

There are many existing works such as the best algorithms, the highly depended attributes, dimensionality reduced models, feature scaling importance once etc. But this project is different because by the usage of machine learning algorithms which are feature scaled, and also dimensionality reduced one.

IV. PROPOSED WORK

In this proposed work the different machine learning algorithms like Decision tree, Random forest, Naive Bayes, SVM kernel (RBF and ploy), K- nearest neighbour are applied directly to the dataset. And then applied after dimensionality reduction technique to the data set by applying PCA (Principle Component analysis) algorithm which takes the highest variant attributes and then predict. This project also shows how the data is distributed graphically by plotting the graph taking 2 attributes to plot linearly.

The proposed work is to take the real-time data by using ECG and BPM sensors and predict that data using different algorithms like Decision tree, Random forest, Naive Bayes, SVM kernel (RBF and ploy), K- nearest neighbour and for dimensionality Reduction PCA (Principle Component analysis) is used. This is useful to find the patients whose has heart disease in an easy way. And we are going to compare these algorithms and find out the best algorithm by its accuracy value.

V. TOOLS/HARDWARE/SOFTWARE REQUIREMENT

The The hardware and software requirements used for this proposed work are Wearable Medical System, ECG and BPM sensors, Spyder Ide-Python anaconda, Jupiter Notebook.

VI. IMPLEMENTATION STRATEGY

A. Algorithms used for Predictions:

These are the algorithms used for prediction of the Heart Condition. These are all machine learning algorithms. For giving the best results PCA (Principle Component Analysis).

1) Decision Tree:

Decision Tree Algorithm is an ensemble learning of methods like classification, regression and other tasks. This is based on the classification of the attributes based on their Info and gain. The height gain attribute is split into its respective values. The new split data is set into new data sets and the applied as new decision tree. This is repeated until we get pure subsets. This is one of the best algorithms for linearly distributed datasets. [14] Tree models with a discrete set of values that the target variable can take are called classification trees; In these tree structures, leaves represent class labels and branches represent combination of features that lead to class labels.

2) Random Forest Algorithm:

Random Forest algorithm or Random forests are the same algorithms These are ensembles of methods for classification, regression and other tasks. These work on the basis of construction of multiple decision trees. This will be better than decision tree because here we are using multiple trees to predict the results, as there are more trees to predict the accurate results. Random decision forests correct the flaw of decision tree algorithm's habit of overfitting to their training set.

Random forest algorithm is more adaptable and easier to use machine learning algorithm that predicts and produces, even without hyper-parameter tuning. It is mostly used algorithm because of its simplicity.

3) Naïve Bayes:

Naïve Bayes classifier is a mere "probabilistic classifier" within the family.This is based on applying the theorem of Bayes with strong independent assumptions among the features of the attributes. Naïve Bayes classifiers are exceptionally adaptable, requiring different parameters in the amount of factors in a learning problem. By evaluating iterative approx the most extreme probability training can be performed.

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Through this Naive Bayes formula, the predictions are made.

4) SVM Kernel (RBF and POLY):

In Machine Learning, kernel methods are a class of algorithms for pattern analysis the Support Vector Machine (SVM) being the best-known member. The pattern analysis task is to study and find the general relationships in clusters, rankings, in any given data sets.

Radial Basis Function kernel, or RBF kernel,is widely used in various learning algorithms. In precise, it is ordinarily used in Support Vector Machine Classification.

The polynomial kernel is a kernel function widely used with support vector machines (SVMs) and other kernelized models, that exemplifies the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

5) K-Nearest Neighbour:

One of the simplest of all Machine learning algorithms is the k - NN algorithm.The algorithm k-nearest neighbour is a method used to classify and regress. In both cases, the input consists in the feature space of the k closest training examples. The output depends on whether classification or regression k-NN is used

K-NN is a lazy learning method in which the function is only locally approximated and the computation is postponed until final classification.

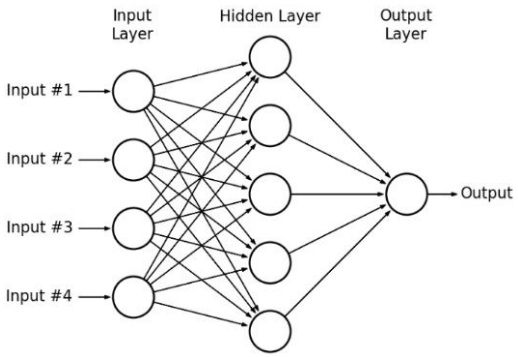
A. Artificial Neural Networks:

Artificial Neural Networks are inspired by the biological neural networks and are the mimic of the human neural networks. This is not itself an algorithm, it is framework of many different machine learning algorithms. There are collection or group of connected nodes called Artificial Neurons.Each neuron is connected like brain synapses used to transmit the signal from neuron to neuron.

It receives, processes then sends the signals to the other artificial neurons in which they are connected. Each neuronal connection is called Edges.

There is a weight in Artificial Neurons and Edges that adjusts as learning progresses.The weights are adjusted to reflect the strength of the signal at a connection.Artificial neurons are typically assembled into layers.

Different layers can transform their inputs in different ways.Signals travel from the first layer (the input layer) to the last layer (the output layer) after multiple crossings of the layers.



B.For dimensionality Reduction:

For the better accuracy dimensionality is reduced so that only the attributes that are responsible for the prediction of the heart condition prediction is taken into consideration and the rest which does not affect the change in the class label is discarded from the dataset. Then the accuracy of the model will be changed appropriately.

1) PCA (Principal Component analysis):

Primary Component Analysis (PCA) is a statistical procedure using an orthogonal transformation to convert a set of observations of potentially correlated variables into a set of values of linearly uncorrelated variables called main components. If variables are observed, the number of separate main components is min (n-1, p). This transformation is defined in such a way that the first main component has the greatest possible variance and, in turn, each successor component has the highest possible variance under the constraint that it is orthogonal to the previous components.The resulting vector is an orthogonal base set that is uncorrelated. PCA is sensitive to the original variables ' relative scaling.**Methodology**

A) The data set used is

- The data set is a Hungarian hospital's data of 400 persons.
- In this there are 14 attributes.
- With 400 data records of different persons in Hungary.
- Here we are using all supervised learning algorithms for prediction.

B) The Attributes are:

These attributes are taken from the UCI machine learning repository where the attributes description is given and explained clearly. [16]

1. Age - age in years
2. sex - sex (1 = male; 0 = female)
3. Cp - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital)
5. cho - serum cholesterol in mg/dl
6. fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. Restecg - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
8. Thalach - maximum heart rate achieved
9. Exang - exercise induced angina (1 = yes; 0 = no)
10. Oldpeak - ST depression induced by exercise relative to rest
11. Slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = down sloping)
12. ca - number of major vessels (0-3) coloured by fluoroscopy
13. Thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
14. Num- the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

3) Method of applying the algorithm:

2) First importing the libraries
3) Reading the dataset, splitting the dataset into training and test of 80% and 20%
4) Applying Feature Scaling
5) Setting the Classifier for all the ML algorithms one at a time
6) Predicating the results
7) Finding the accuracy by using Confusion Matrix
8) Applying the PCA algorithm and taking the new training and test datasets.
9) Setting the Classifier for all the ML algorithms one at a time
10) Predicating the results
11) Finding the accuracy by using Confusion Matrix
12) Comparing the results

VII. ACCURACY OF EACH ALGORITHM

Different types of algorithms like Decision tr.ee, Random forest, Naive Bayes, SVM Kernel (RBF and poly), K-nearest neighbour are used to predict the data of the heart patients and for dimensionality Reduction, PCA (Principle Component analysis) is used using Python anaconda software.

1. Decision Tree

- Accuracy with 14 attributes:
- 44.73%

```
In [4]: cm
Out[4]:
array([[28,  8,  0,  3,  1],
       [ 3,  3,  0,  3,  1],
       [ 1,  5,  1,  1,  1],
       [ 1,  4,  1,  2,  4],
       [ 0,  4,  0,  1,  0]], dtype=int64)
```

- Accuracy after PCA dimensionality reduction:
- 52.63%

```
In [8]: cm
Out[8]:
array([[31,  6,  2,  1,  0],
       [ 8,  1,  1,  0,  0],
       [ 2,  3,  4,  0,  0],
       [ 3,  1,  3,  4,  1],
       [ 1,  0,  2,  2,  0]], dtype
```

2. K- Nearest neighbour

- Accuracy with 14 attributes:
- 52.63%

```
In [10]: cm
Out[10]:
array([[37,  3,  0,  0,  0],
       [ 6,  3,  0,  0,  1],
       [ 2,  7,  0,  0,  0],
       [ 3,  6,  3,  0,  0],
       [ 0,  4,  1,  0,  0]], dtype=int64)
```

- Accuracy after PCA dimensionality reduction:
- 53.94%

```
In [25]: cm
Out[25]:
array([[37,  1,  1,  1,  0],
       [ 7,  2,  1,  0,  0],
       [ 1,  7,  0,  1,  0],
       [ 2,  3,  5,  2,  0],
       [ 0,  1,  3,  1,  0]], dtype=int64)
```

1. Naive Bayes

- Accuracy with 14 attributes:
- 42.10%

```
In [5]: cm
Out[5]:
array([[26,  0,  0,  1, 13],
       [ 2,  1,  1,  0,  6],
       [ 0,  2,  0,  0,  7],
       [ 0,  0,  1,  1, 10],
       [ 0,  0,  0,  1,  4]], dtype=int64)
```

- Accuracy after PCA dimensionality reduction:
- 59.21%

```
In [7]: cm
Out[7]:
array([[38,  2,  0,  0,  0],
       [ 6,  2,  2,  0,  0],
       [ 2,  5,  2,  0,  0],
       [ 1,  3,  5,  3,  0],
       [ 0,  2,  2,  1,  0]], dtype=int64)
```

2. Random forest

- Accuracy with 14 attributes:
- 53.94%

```
In [3]: cm
Out[3]:
array([[37,  1,  1,  1,  0],
       [ 7,  1,  1,  0,  1],
       [ 3,  3,  1,  2,  0],
       [ 3,  3,  4,  2,  0],
       [ 1,  4,  0,  0,  0]],
```

- Accuracy after PCA dimensionality reduction:
- 57.89%

```
In [5]: cm
Out[5]:
array([[35,  2,  2,  1,  0],
       [ 6,  3,  1,  0,  0],
       [ 3,  1,  3,  2,  0],
       [ 3,  2,  4,  3,  0],
       [ 1,  2,  1,  1,  0]], dtype=int64)
```

3. SVM Poly

- Accuracy with 14 attributes:
- 51.31%

```
In [4]: cm
Out[4]:
array([[38,  1,  0,  1,  0],
       [ 8,  0,  2,  0,  0],
       [ 8,  1,  0,  0,  0],
       [ 4,  3,  3,  1,  1],
       [ 1,  3,  0,  1,  0]], dtype=int64)
```

- Accuracy after PCA dimensionality reduction:
- 59.21%

```
In [6]: cm
Out[6]:
array([[40,  0,  0,  0,  0],
       [ 8,  1,  1,  0,  0],
       [ 7,  0,  2,  0,  0],
       [ 4,  2,  4,  2,  0],
       [ 2,  2,  1,  0,  0]], dtype=int64)
```

4. SVM RBF

- Accuracy with 14 attributes:
- 48.68%

```
In [4]: cm
Out[4]:
array([[36,  3,  1,  0,  0],
       [ 7,  1,  2,  0,  0],
       [ 2,  7,  0,  0,  0],
       [ 5,  3,  4,  0,  0],
       [ 1,  3,  0,  1,  0]], dtype=int64)
```

- Accuracy after PCA dimensionality reduction:

```
In [6]: cm
Out[6]:
array([[38,  2,  0,  0,  0],
       [ 6,  3,  1,  0,  0],
       [ 4,  3,  2,  0,  0],
       [ 3,  3,  5,  1,  0],
       [ 0,  3,  2,  0,  0]], dtype=int64)
```

5. Artificial Neural Networks:

- Accuracy – 58.53%

```
Epoch 100/100
217/217 [==============================] - 0s 142us/step - loss: -6.0687 - acc: 0.5853
24/24 [==============================] - 0s 10ms/step
```

After PCA: -

```
Epoch 100/100
217/217 [==============================] - 0s 234us/step - loss: -5.5517 - acc: 0.5530
24/24 [==============================] - 0s 11ms/step
```

VIII. RESULTS AND GRAPHS

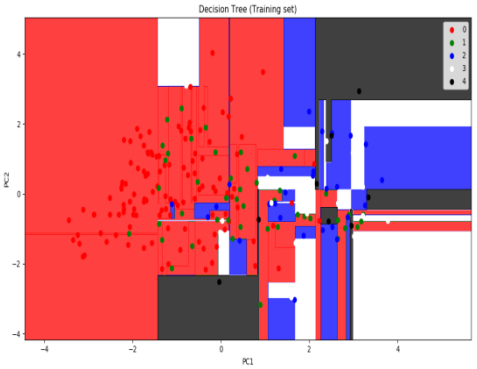


Fig-4.1
Decision Tree (Training set)

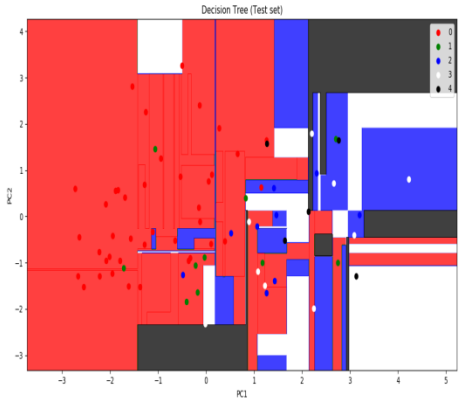


Fig-4.2
Decision Tree (Test set)

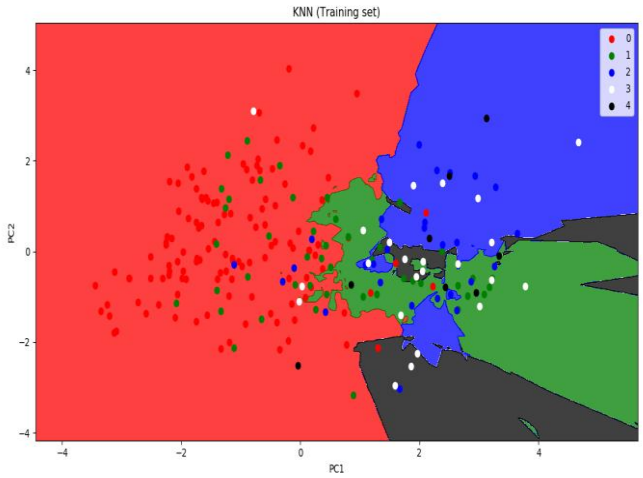


Fig-5.1
Nearest neighbour (Training set)

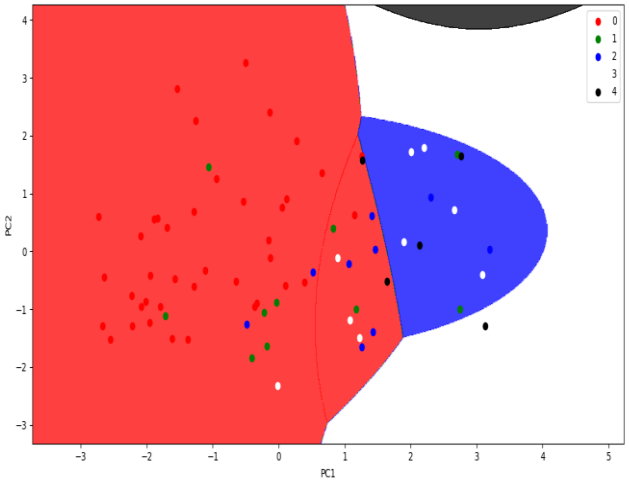


Fig-6.2
Naive Bayes (Test set)

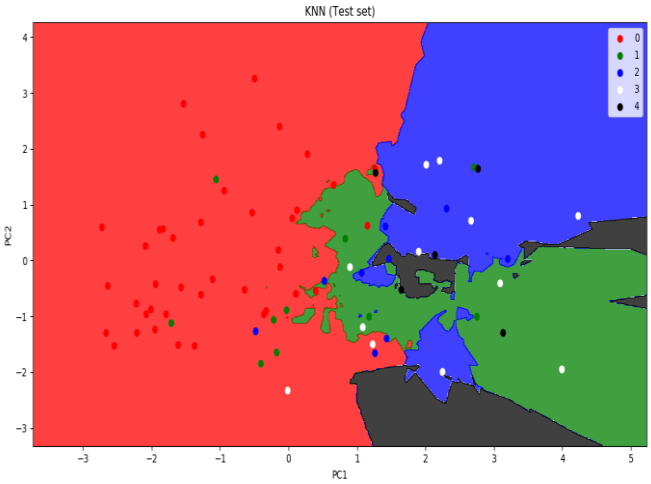


Fig-5.2
K-Nearest neighbour (Test set)

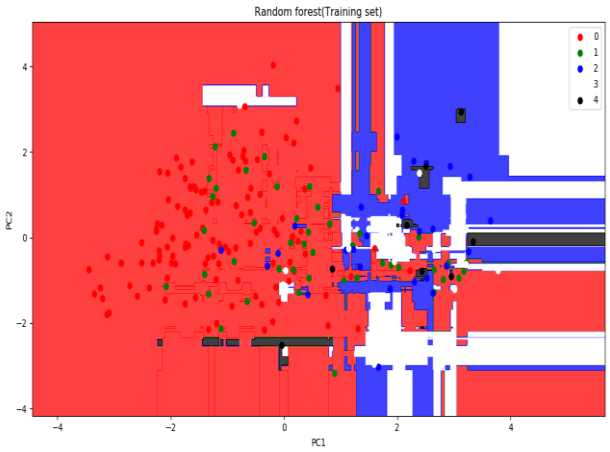


Fig-7.1
Random Forest (Training set)

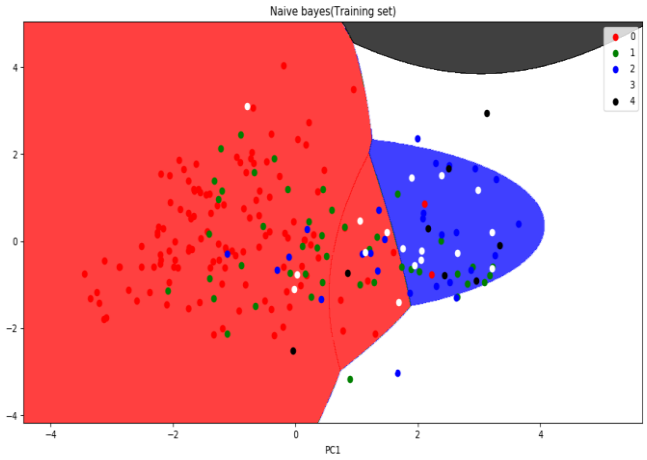


Fig-6.1
Naive Bayes (Training set)

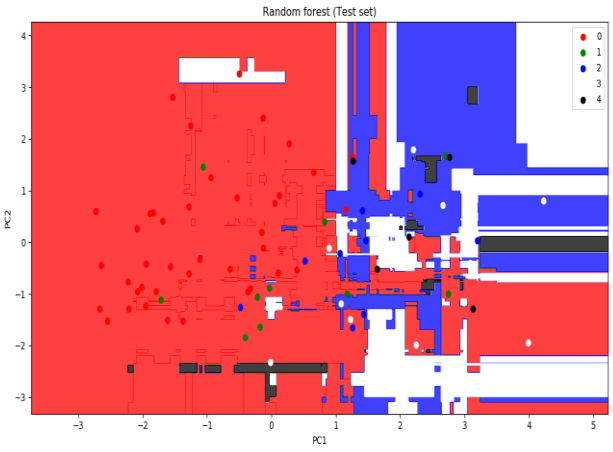


Fig-7.2
Random Forest (Test set)

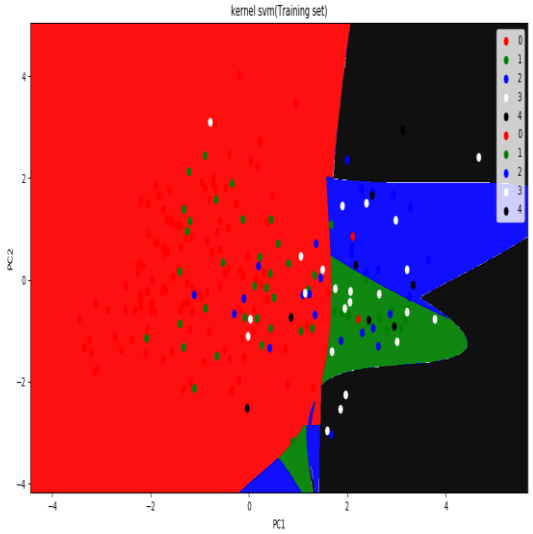


Fig-8.1
SVM Poly (Training set)

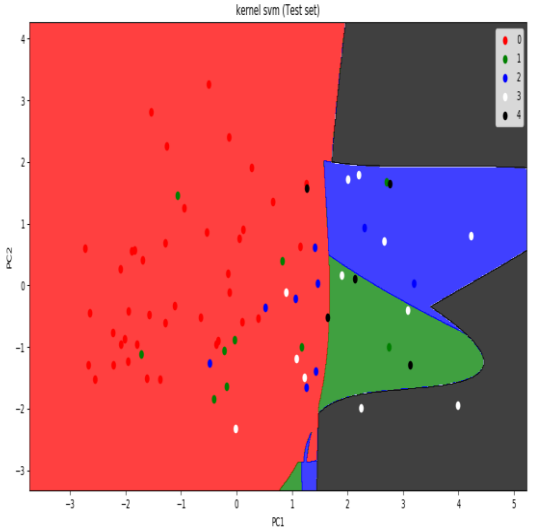


Fig-8.2
SVM Poly (Test set)

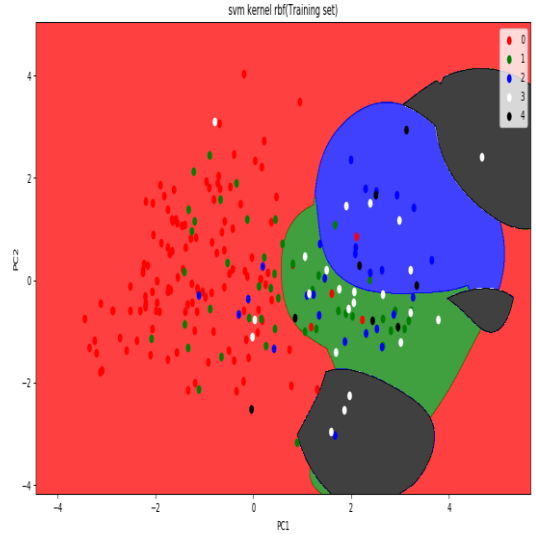


Fig-9.1
SVM RBF (Training set)

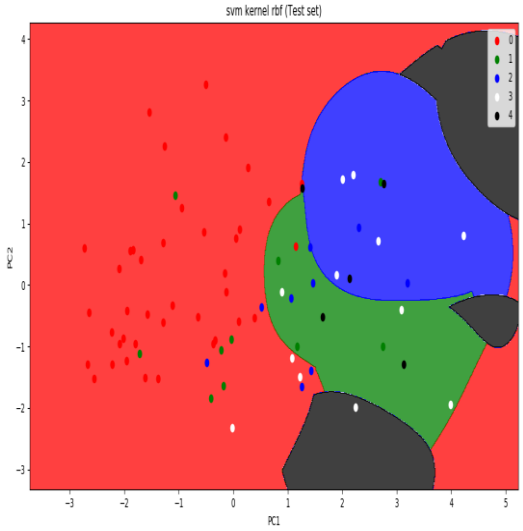


Fig-9.2
SVM RBF (Test set)

IX. CONCLUSION

The conclusion of this paper is the neural networks is the best algorithm for the prediction of the heart condition among the other algorithms like KNN, SVM, Random Forest, Naïve Bayes, and Decision tree. The neural networks results are directly proportional to the no of inputs they are given. So, the more the inputs for them the more the accuracy they show.

REFERENCES

1. *Improvising Heart Attack Prediction System using Feature Selection and Data Mining Methods B. Kavitha* and R. Naveen Kumar Lecturer Department of Computer Applications, Karpagam University Coimbatore, India International Journals of Advanced Research in Computer science, VOLUME 1, No. 4, NOV-DEC 2010*
2. *Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review V. Krishnaiah, G. Narsimha, N. Subhash Chandra, Research Scholar, JNTUH Dept. of CSE, Hyderabad Telangana State, India International Journal of Computer Applications (0975 – 8887) Volume 136 – No.2, February 2016*
3. Heart Disease Prediction System Using Data Mining Techniques Abhishek Tanja Department of Computer Science, S.A. Jain College, Ambala City, India. An International Open Free Access, Peer Reviewed Research Journal (Received: November 15, 2013; Accepted: November 25, 2013)
4. Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) Approaches Mohammad A. M. Abushariah, Assal A. M. Alqudah, Omar Y. Adwan, Rana M. M. Yousef Journal of Software Engineering and Applications, Vol.7 No.12, November 28, 2014

5. An Effective Performance Analysis of Machine Learning Techniques for Cardiovascular Disease Department of Computer Science, Amrita School of Engineering Bangalore Campus, Amrita Vishwa VidyaPeetham, Kasavanahalli, Carmelaram P.O., Bengaluru Applied Medical Informatics Vol. 36, No. 1 /2015, pp: 23-32
6. Feature Analysis of Coronary Artery Heart Disease Data Sets Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam and M. Essam Khalifa British University in Egypt (BUE), Cairo, Egypt International Conference on Communication, Management and Information Technology. Procedia Computer Science 65 (2015) 459 – 468
7. A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez, Department of Computer Science, University of Georgia, Athens, USA Institute of High-Performance Computing and Networking (ICAR - CNR), Naples, Italy
8. Predictive and Descriptive Analysis for Heart Disease Diagnosis František Babič, Jaroslav Olejár Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical university of Košice, Slovakia Proceedings of the Federated Conference on Computer Science and Information Systems pp. 155–163 DOI: 10.15439/2017F219 ISSN 2300-5963 ACSIS, Vol. 11
9. EARLY HEART DISEASE PREDICTIONUSING DATA MINING TECHNIQUES Aditya Methaila, Prince Kansal , Himanshu Arya , Pankaj Kumar Netaji Subhas Institute of Technology,India and 2 Student, B.Tech (CSE), Maharaja Surajmal Institute of Technology New Delhi, India Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014 DOI : 10.5121/csit.2014.4807
10. Effective heart disease prediction system using data mining techniques Poornima Singh, Sanjay Singh, Gayatri S Pandi-Jain L. J. Institute of Engineering and Technology, Gujarat Technological University, Institute of Life Sciences, School of Science and Technology, Ahmedabad University, Ahmedabad, Gujarat, India International Journal of Nanomedicine, Volume 13, T-NANO 2014
11. BACKPROPOGATION NEURAL NETWORK FOR PREDICTION OF HEART DISEASE NABEEL AL-MILLI Financial and Business Administration and Computer Science Department Zarqa University College Al-Balqa' Applied University Journal of Theoretical and Applied Information Technology 10th October 2013. Vol. 56 No.1
12. Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015
13. Seyedamin Pouriyeh, Sara Vahid, Giovanna, Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez. "A comprehensive investigation, and comparison of Machine Learning Techniques in the domain of heart disease", 2017 IEEE Symposium on Computers and Communications (ISCC), 2017
14. En.wikipidia.org
15. Submitted to Texas A&M University, College Station
16. https://archive.ics.uci.edu/ml/datasets/heart+Disease