ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Heart Disease and Alzheimer Prediction based on Hybrid Classification Algorithm

Jayashri J. Patil[1], Nilesh Vani[2]

M. Tech Student, Dept. of Computer Engineering, GF's Godavari of Engineering, Jalgaon, Maharashtra, India[1]

Assistant Professor, Dept. of Computer Engineering, GF's Godavari of Engineering, Jalgaon, Maharashtra, India[2]

ABSTRACT:Nowadays, heart diseases are very common and one of the major causes of death across the world. This calls for an accurate and timely diagnosis of the heart disease. There is abundant data available with the health care systems; however, the knowledge about the data is rather poor. Data scientists have attempted several methods in order to improvise the examination of large data sets. Previously, various data mining techniques have been implemented in the healthcare systems, however, the hybridization in addition to a single technique in the identification of heart disease shows promising outcomes, and can be useful in further investigating the treatment of the heart diseases. The framework enables the representation, extraction, and mining of high order latent structure and relationships within single and multiple disease sequences. This work attempts to survey some recent techniques applied towards knowledge discovery for heart disease prediction and further proposes a novel prediction method with improved accuracy.

Unfortunately, Alzheimer's disease (AD) cannot be slowed or cured with today's medication. The studies have revealed that - a cognition drop is a precursor of AD, the progression of AD is highly correlated to cognition decline, and AD's early detection and intervention becomes increasingly clear to be the best choice of improving quality of life for persons with probable AD. This survey aims to improve the predictive model by focusing on AD early detection. Compared to models built from traditional approaches such as neuron networks, Bayesian networks, we propose a novel prediction method with improved accuracy.

KEYWORDS:Heart Disease Prediction, Machine Learning, Data Mining, Alzheimer's disease (AD), deep learning, AD early detection.

I. INTRODUCTION

Mobile Ad Hoc Networks (MANETs) consists of a collection of mobile nodes which are not bounded in any infrastructure.Nodes in MANET can communicate with each other and can move anywhere without restriction. This non-restricted mobility and easy deployment characteristics of MANETs make them very popular and highly suitable for emergencies, natural disaster and military operations.

1.1 Data mining:

Data mining (sometimes knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

1.2 Basic terms related to data mining:

1.2.1 Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy".

1.2.2 Supervised learning:

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier. The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

1.2.3 Unsupervised learning:

In machine learning, unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning.

The heart attack occurs when the arteries which supply oxygenated blood to heart does not function due to completely blocked or narrowed.

1.2.4 Prediction:

Models continuous-valued functions, that is predicts unknown or missing values.

II. RELATED WORK

2.1 For Heart Disease :

A. Hlaudi Daniel Masethe et al., (2014) used data mining algorithms like J48, NB and REP TREE for predicting heart attacks. The medical database was collected from the doctors in South Africa. The various attributes considered were Gender, age, CPT, ECG, RBP, Thalach, serum cholesterol, alcohol, obesity(diet) and smoking. WEKA - Waikato Environment for Knowledge Analysis tool was used for discovering, analysing and predicting patterns for heart disease. The accuracy obtained were 99.0741, 99.222, 98.148 for J48, REPTREE and NB respectively. [7]

Limitations: Some important attributes such as Col-Ves, Thal, Ex-Ang, etc., are not considered for prediction. Here various data mining algorithms are implemented and compared to find the best method for prediction. They came to a conclusion that algorithms such as J48 and REPTREE are efficient in the prediction of heart disease. This conclusion is derived by not considering some efficient Data mining algorithms such as Regression and Artificial Neural Network algorithms.

B. Theresa Princy R et al., (2016) predicted heart disease using ID3 and KNN algorithm. The ID3 algorithm is used as a classifier, KNN algorithm organizes and pre-processes the incorrect values which are considered as the training set. The basic factors along with some additional factors such as smoking were included. The accuracy level increased to 80.6%. [5]

Limitations: Very few factors are considered for the prediction. Some common influencing factors such as CPT (Chest Pain Type), RECG (Resting electrocardiographic (ECG)) and indirectly influencing factors such as Alcohol and Obesity are not considered. Without using these important factors, the prediction could not be given precisely.

C. VivekanandanTet al., (2017) proposed the challenging tasks of selecting critical features from the enormous set of available features and diagnosing heart disease. DE (Modified Differential Evolution) algorithm is used to perform feature selection. Prediction of heart disease was carried out using Fussy AHP and Feed-Forward neural network. Using 9 attributes an accuracy of 98% was achieved. [2]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Limitations: More number of inputs are not listed and Error minimization is not carried out properly. It could have been carried out by using effective back -propagation model. Without proper testing they have mentioned that large data sets can also be adopted.

D. Aditi Gavhane and GouthamiKokkula (2018) proposed the system that use the neural network algorithm multi-layer perceptron (MLP) to train and test the dataset. Multi-layer perceptron algorithm is a supervised neural network algorithm in which there will be one layer for the input, second for the output and one or more for hidden layers between these two layers.

Limitations: The existing algorithm has been modified into new algorithm which takes only two values for prediction. By considering just two values they have come to a conclusion that their proposed algorithm achieves higher accuracy when compared to other algorithms.

2.2 For Alzheimer Disease:

A. Tingyan Wang andJason L. Qiu (2018) proposed the system that uses deep learning algorithm. In this, they explored how RNN models could be applied to AD early detection modeling (AD-EDM). In particular, using an LSTM RNN, they developed a predictive model based on the NACC's early stages' patient data.

Limitations: The dimensionality of the data is reduced using attribute selection method, since it consumes more time for classification. This decrease in the number of attributes does not give the correct prediction.

B .Veeramuthu et al. (2014) developed a CAD (AR mining algorithm) tool for decision making about the presences of abnormalities in human brain. The author suggested preprocessing of PET dataset for instance, spatial normalization and intensity normalization. Fisher Discriminants ratio (FDR) was used for feature extraction to get ROIs. The instances were classified to normal if the extracted number of verified rules were above the final threshold otherwise image was classified as AD.

Limitations: No dataset details, missing values or any pre-processing steps highlighted

C. Anshul Bhagtani, TanupriyaChoudhuri (2017) proposes a method of selecting features which have more separable value than others and these features are processed and the performance is evaluated. The proposed algorithm will reduce the testing time as it would check for if the person is suffering from Alzheimer's or not.

Limitations: The proposed algorithm which takes only two values for prediction. By considering just two values they have come to a conclusion that their proposed algorithm achieves higher accuracy when compared to other algorithms

III. METHODOLOGY

The Heart Disease and Alzheimer Prediction System is done using C5.0 , SVM and Hybrid Data mining technique.

Record set with medical attributes was obtained from the Cleveland Heart Disease database. With the help of the dataset, the patterns significant to the heart attack prediction and Alzheimer prediction are extracted.

The records are split into two datasets:

Training dataset.

Testing dataset.

A] Attributes Used for Heart Disease :

Sr. No.	Attribute	Meaning
1	AGE	Age
2	SEX	Gender
3	AMPLITUDE	is a measure of its change after a patients exercise
4	RESTING BP	Resting Blood Pressure
5	CAL LEVEL	Level of Calcium

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

6	SPIKE	Narrow Pain occurrence
7	FREQUENCY	cardiovascular disease (CVD) risk factors
8	PULSE RATE	Pulse rate of patient
9	δWAVE FREQUENCY	Alpha waves are neural oscillations in the frequency range of 8–12 Hz
10	HEREDITORI	If hereditary
11	SEIZER ATTACK	If any no. of attack

B] **Attributes Used for Alzheimer Disease :**

Sr. No.	Attribute	Meaning
1	AGE	Age
2	SEX	Gender
3	HYPERTEN	Hypertension
4	RESTING BP	Resting Blood Pressure
5	EDUC. YRS	Years of Education
6	COMFORT	Behavior, comportment, and personality
7	SHOPPING	In the past four weeks, did the subject have any difficulty or need help with: Shopping alone for clothes, household necessities, or groceries
8	GAMES	In the past four weeks, did the subject have any difficulty or need help with: Playing a game of skill such as bridge or chess, working on a hobby
9	STOVE	In the past four weeks, did the subject have any difficulty or need help with: Heating water, making a cup of coffee, turning off the stove
10	MEALS	In the past four weeks, did the subject have any difficulty or need help with: Preparing a balanced meal
11	MEDICATIONS	If any medications

Proposed System :

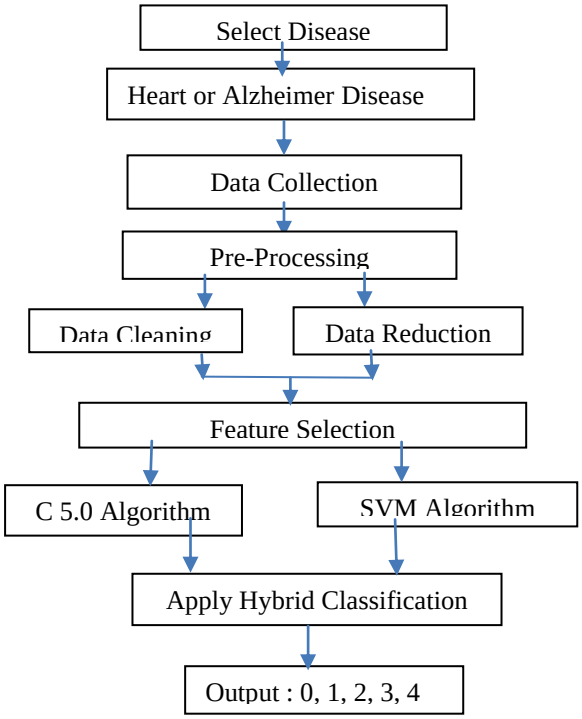
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019



IV. PROPOSED ALGORITHM

1. C5.0 Algorithm:

C5.0 algorithm is a successor of C4.5 algorithm. It gives a binary tree or multi branches tree. It uses Information Gain (Entropy) as its splitting criteria. C5.0 pruning technique adopts the Binomial Confidence Limit method. In a case of handling missing values, C5.0 allows to whether estimate missing values as a function of other attributes or apportions the case statistically among the results.

Algorithm for Experimental Model

Input: dataset.

Output: classified output.

1. Take a data set as input.
2. If that set has more features then apply the feature selection technique (PCA) as pre-processing technique
3. Apply parallelism from step 4 to step 6.
4. Evaluate the entropy value and information gain ratio of all three entropies (Shannon, havrda and Charvat's entropy and quadratic entropy).
5. Construct the models separately using C5.0 algorithm based on various entropies.
6. Find the accuracy and execution time of each model and store the value in array.
7. Find a model that has maximum Accuracy.
8. If two have maximum accuracy then
9. Find a minimum execution time of the model that has maximum accuracy.
10. Classify by that model which has minimum execution time.
11. Else classification done by the model which has maximum accuracy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

12. End

2. SVM Algorithm :

Support Vector machines(SVM) have gained popularity in the machine learning and pattern classification.

The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. For a linearly separable dataset, a classification function corresponds to a separating hyperplane f (x).

SVM guarantees that the best such function is found by maximizing the margin between the two classes. Thus, it is a linear classifier which constructs a separating hyperplane to maximize distance between data.

Input: Training data: X0 =[X1,X2,...,Xk,...,Xl]T
Class labels: y0 =[y1,y2,...,yk,...,yl]T
Initialize: Surviving features: S = [1 ,2,...,n]
RFE sequence: r = []
Repeat until s = []
Restrict training data to good feature indices: X = X0(:,s)
Train the classifier: α = SVM −train(X,y)
Calculate the weight vector of dimension length(s): ω=∑kαkykXk
Calculate the sequence criteria: ci =(ωi)2, for all i
Identify feature with the smallest sequence criterion: f = argmin(c)
Update RFE sequence: r =[s(f),r]
Remove the identified feature: s = s(1 : f −1,f+1:length(s))
End
Output: RFE sequence: r

3. Hybrid Algorithm:

1. Take a data set as input.
2. If that set has more features then apply the feature selection technique (PCA) as pre-processing technique
3. Apply parallelism from step 4 to step 6.
4. Evaluate the entropy value and information gain ratio of all three entropies (Shannon, havrda and Charvat's entropy and quadratic entropy).
5. Construct the models separately using C5.0 algorithm based on various entropies.
6. Find the accuracy and execution time of each model and store the value in array.
7. Find a model that has maximum Accuracy.
8: Initialize an array M[N] for N no. of parameters
N is between 1 to 20 i.e. 1 < N < 20
N is real number.
9: Assume P[N] be array of possible values in M[N]
P[N]= f1; 2; 3; :::g
10: for(i=1; i<Cn ;i++)
for(j=1; j <Cn ; j++)
11: A] Calculate individual probability Pi for all classes
Pi = P (Cn);
12: B] Calculate group probability for all combinations
Pn = P (n | n + c)
where, n and c is no.of classes

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

13: C] Calculate prediction from individual and group
P (Ci | Pn) >P(Cj | Pn)
14: D] Calculate maximum probability from the prediction
P(Ci >Pn) = P(N | Ci) P(Cj) | P(N)
Pmax<P(Ci | Pn)
Pmax = P(Ci | Pn)
end for, end for
15. If two have maximum accuracy then
16. Find a minimum execution time of the model that has maximum accuracy.
17. Else apply Backpropagation and go to step 3
17. Classify by that model which has minimum execution time.
18. Else classification done by the model which has maximum accuracy.
19. End

V. SIMULATION RESULTS

The application of techniques reveals the results of all three applying algorithms C5.0, SVM, and proposed Hybrid Algorithm. In very first step a medical data set has been downloaded from an open source database named UCI repository. In next, irrelevant rows are excluded according to specific requirement of heart disease and Alzheimer i.e. total rows or attributes are more than 300 we reduced or select 11 attributes that are most relevant for the prediction of heart disease and Alzheimer Disease i.e age, sex, amplitude, hypertension etc.

Performance Evaluation

For Performance evaluation of the approach we measure it based on 2 parameters i.e precision and recall . Precision and Recall are defined in terms of a set of retrieved documents (e.g. the list of documents produced for a query) and a set of relevant documents (e.g. the list of all documents that are relevant for a certain topic)

Precision :

Precision is the fraction of retrieved documents that are relevant to the find.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall :

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Below graph highlights the outcome of Hybrid Classification Algorithm implementation. The same prediction model built using C5.0 and SVM is also shown. Figure clearly shows that C5.0 and SVM prediction accuracy does not reach expected accuracy level as compared to Hybrid Classification Algorithm. Bar Chart depicts difference in accuracy performance between C5.0, SVM and Hybrid Classification Algorithm. It is observed that from 500 to 10000 records, the accuracy increases from average 75% to 90%.

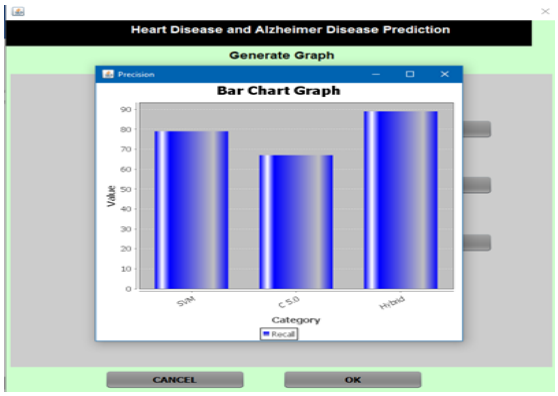
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

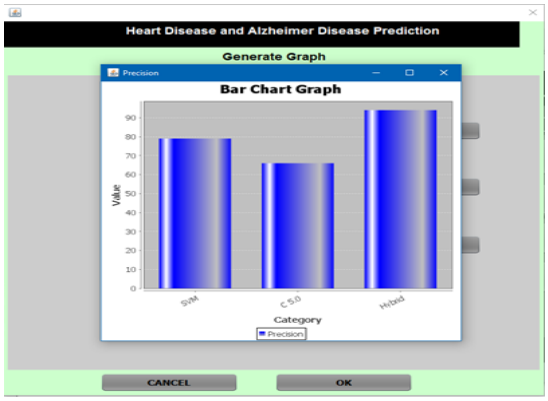
International Journal of Innovative Research in Computer and Communication Engineering

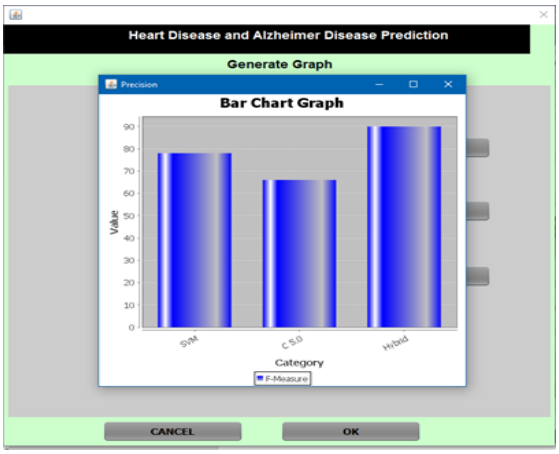
(A High Impact Factor, Monthly, Peer Reviewed Journal)

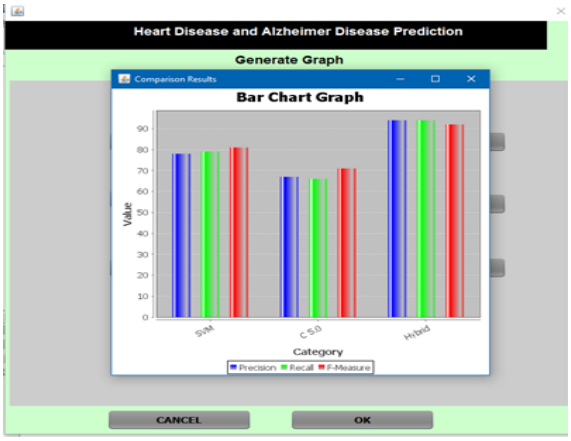
Website: www.ijircce.com

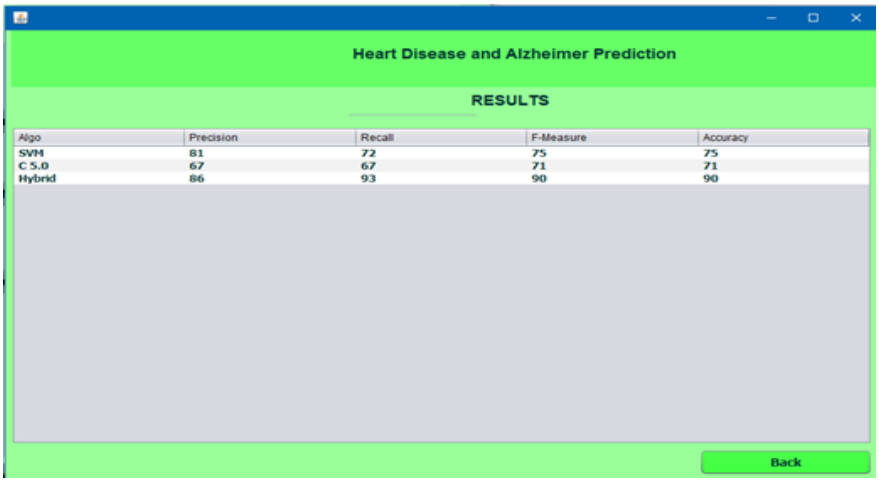
Vol. 7, Issue 5, May 2019











VI. CONCLUSION AND FUTURE WORK

Machine learning systems make medical professionals faster and smarter in their diagnosis. As a result, it reduces uncertainty in their decisions, thereby reducing cost risks and saving valuable time. In this study, the proposed

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

ensemble and hybrid algorithm demonstrate that hybrid machine learning techniques perform better than the individual algorithms on selected medical datasets. The proposed hybrid algorithm composed of C5.0 and SVM algorithm. Selected individual algorithms separately and proposed hybrid algorithm is applied on different datasets using average probability combination rule is used to get better prediction accuracy. The result shows that the hybrid machine-learning algorithm is the key to improve the prediction accuracy of individual machine learning algorithms for Heart disease and Alzheimer. In future, we can apply this hybrid algorithm for the multidisease healthcare solution.

REFERENCES

1. D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.
2. Vivekanandan T et al., "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease", www.elsevier.com/locate/compbiomed, https://doi.org/10.1016/j.compbiomed, Pages: 125-136 (2017)
3. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
4. I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, "Data Mining: Practical machine learning tools and techniques," Morgan Kaufmann, 2016.
5. Theresa Princy R, "Human Heart Disease Prediction System using Data Mining Techniques", International Conference on Circuit, Power and Computing Technologies [ICCPCT], IEEE (2016)
6. Venkatalakshmi B et al., "Heart Disease Diagnosis using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology [IJIRSET], ISSN (Online): 2319-8753, Volume: 3, Issue: 3, Pages :18731877 (2014)
7. Hlaudi Daniel Masethe et al., "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engg and Computer Science [WCECS], Volume: II, ISBN: 978-988-19253-7-4 ISSN: 2078-0958 (Print), ISSN: 2078-0966 (Online) (2014)
8. Tingyan Wang1, 3,Jason L. Qiu2, Robin G. Qiu3*, Ming Yu1," Early Detection Models for Persons with Probable Alzheimer's Disease with Deep Learning" Dept of Industrial Engg, Tsinghua University, Beijing 100084, China 978-1-5386-1803-5/18/$31.00 IEEE (2018)
9. Anshul Bhagtani,Tanupriya Choudhury., An Efficient Survey to Detect Alzheimer Disease Using Data Mining Techniques",978-1-5386-1144-9$31.00 IEEE (2017)
10. Veeramuthu, A., S. Meenakshi, et al. (2014). A New Approach for Alzheimer's Disease Diagnosis by using Association Rule over PET Images. International Journal of Computer Applications 91(9), 9-14.
11. Aditi Gavhane, GouthamiKokkula, "Prediction of Heart Disease Using Machine Learning ",Proceedings of the 2nd International conference on Electronics, Communication and Aerospace (ICECA 2018) IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1.
12. Kalaiselvi C, "Diagnosing of heart diseases using Average KNearest Neighbor Algorithm of Data Mining", 2016 InternationalConference on Computing for Sustainable Global Development (INDIACom), IEEE, Pages: 3099-3103 (2016).
13. Sunil Ray, "Learn Naïve Bayes algorithm" and "Decision tree- Simplified", URL: www.analyticsvidhya.com, Retrieved on- 10.01.2018.