Heart Disease Prediction using Machine Learning

Rohit Saroha[1] Vineet Kumar[2] Mohd Anwar Hussain[3] Manish Kumar[4] Prof. Geeta Patil[5]
[1,2,3,4,5]Department of Information & Technology
[1,2,3,4,5]Army Institute of Technology Pune, India

Abstract— The majority of the passings over the globe are brought about by heart maladies and it is a standout amongst the most fatal ailment turning into the purpose behind a large portion of the deaths. Many calculations are being connected to identify the heart illnesses before they achieve a phase at which it can't be restored. Building up a restorative conclusion framework dependent on AI for forecast of coronary illness gives more precise analysis than customary way. This paper exhibits a review of backpropagation exactness and backpropagation alongside the hereditary calculations precision and break down their performance. In this paper, a coronary illness expectation framework which utilizes counterfeit neural system back propagation calculation is proposed. Thirteen clinical parameter[2] were utilized as contribution for the neural system and afterward the neural system was prepared with backpropagation calculation alongside the hereditary calculations to foresee non-appearance or nearness of coronary illness with precision of 92.

Key words: Heart Disease, Artificial Neural Network, Multilayer Perceptron, Genetic Algorithm, Back-Propagation Algorithm

I. INTRODUCTION

Due to the tremendous growth in the volume of healthcare data, data analytics in a healthcare information system play crucial role. [1] Recently, healthcare organizations have been moving towards digitization of the massive volume of healthcare data to leverage data analytics in healthcare to realize extensive benets. The potential benefits which include detection of diseases at earlier stages, cost reduction, personalized care. The healthcare datasets available in the EMRs are generated from different medical sources such as diagnosis, procedures, medications,and lab results [4] and are maintained under different attributes or features.

Clinical Features	Description
Age	Age of person
Cp	Chest pain of person
Chol(mg/dl)	Serum cholesterol level of person
Ca	Number of major vessels(0-3) colored by flourosopy
Exang	Exercise induced angina
Num	Diagnosis of heart disease
Fbs	Fasting blood sugar
Oldpeak	ST depression induced by the exercise relative to rest
Restecg	Resting electrocardiographic results
Sex	Gender
Thal	3=normal ; 6 = fixed defect; 7= reversible defect
Thalach	Maximum heart rate
Trestbps(mmHg)	Resting Blood Pressure of person

Fig. 1: Feature Table [2]

Feature selection or attribute selection has become an important focus in many research applications, for which datasets with tens or hundreds or thousands of variables are available. Feature selection [2] can greatly improve the accuracy of the resulting classier model. Furthermore, it is important in ending the relevant subset of predictive features. For example, a physician might take a decision on the critically of a particular disease based on a classication carried out using the selected features. The accuracy of the prediction is improved by optimizing the feature selection. Optimization is the process of obtaining the best possible values of decision variables based on the selected objective function [3]. In this paper, feature selection is carried out by selecting the critical features that form the root cause for the objective function of the problem under consideration, which is the prediction of heart disease. The heart disease dataset taken from UCI data repository which is used for experimentation.

II. LITERATURE SURVEY

Distinctive sorts of studies have been done to concentrate on forecast of coronary illness and different information digging procedures are utilized for analysis which help to accomplish diverse exactness level for various strategies.[7] Initially, R.W.Jones, M.Clarke, Z.Shen and T. Alberti have proposed an investigation applying neural system to self-connected poll (SAQ) information to build up a coronary illness forecast framework.[1] The examination not just clears up regular hazard components of the sickness yet in addition different information gathered in SAQ.[3] The approval of the work was given by checking against the consequence of the neural system with "Dundee Rank Factor Score" which is identified with measurably 3 chance variables (circulatory strain, smoking and blood cholesterol) together with sex and age to decide danger of having coronary illness. In the examination, they utilized multi-layered feedforward neural system which was prepared with Backpropagation Algorithm. There were three layers in the neural system they utilized: input, covered up and yield layers.Heart infection causes a greater number of passings yearly than some other reason. It is assessed that 17.7 million individuals passed on because of coronary illness in 2015 [5]. The execution was improved to Relative Operating Characteristic (ROC) zone of 98% by expanding input quantities of the neural network.[1]

The quantity of frameworks for expectation of various maladies are proposed and actualized by utilizing distinctive procedures and strategies. The grouping and forecast was prepared by means of learning Vector Quantization Algorithm which is one of Artificial Neural Network learning technique.[3] There were three stages in their procedure. The first was to choose of 13 clinical highlights which are imperative contrasted with others, i.e., age, cholesterol, chest torment type, practice actuated angina, max pulse, fasting glucose, number of vessels hued, old pinnacle, resting ecg, sex, incline, thal and trestbps.[2] Second one was utilizing Artificial Neural Network calculation for order.]. The danger of creating coronary

illness over a lifetime is one of every five [6]. The hazard dimension of coronary illness forecast through cross breed calculation has been proposed by Shovon K.Pramanik.etl [8]. Finally, the coronary illness expectation framework was created. The exactness of forecast rate which was gotten from the investigation is close to 80%.[4]

III. BACKGROUND STUDY

A. Artificial Neural Network

Counterfeit Neural Networks are organize comprised of essential preparing unit which is called as perceptron. Perceptron utilizes one layer for taking care of direct detachable issues. The issues which are not directly distinct is fathomed utilizing Multilayer Perceptron Neural Network (MLP). MLP has numerous layers, which comprise at least three layers: input, covered up and yield layer. A neural system is a kind of neural system in which the stream of data is from the information hubs to the concealed layers and from that point to yield node,which gives us our outcome. ANN utilize three layers:

Information Layer utilize thirteen neurons. Number of neurons utilized ought to be equivalent to the quantity of traits utilized in the informational index.

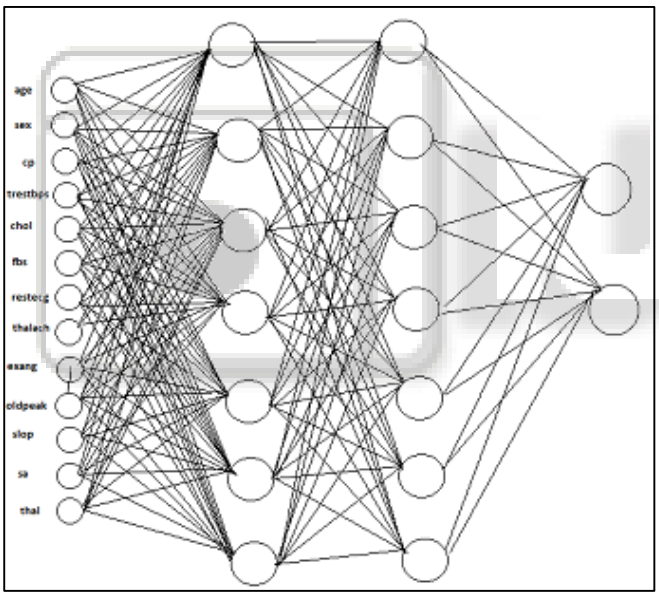

Fig. 2: Backpropagation architecture.

Concealed Layer was intended to contain any number of neurons however for our expectation we utilize three neurons in it. This number use as a beginning stage. The number will build one by one till we came to the quantity of neurons of the info layer by looking at execution of them and furthermore center around intricacy of structure by including more neurons in it and afterward choosing the best one. Yield Layer utilize two neuron which gives result whether an individual have coronary illness or not in parallel configuration where one demonstrate nearness and zero show nonattendance.

B. Backpropagation

Backpropagation is utilized for preparing of neural system and it is additionally used to prepare multilayer perceptron and numerous other neural networks.In backpropagation the yield obtained is contrasted and expected yield and blunder is computed.This registered mistake is again offered back to neural system and loads are changed in accordance with decline the error,so coming about yield will have less error.This process is rehashed number of times to such an extent that at every emphasis the mistake will get diminished and yield get all the more closer to anticipated yield.

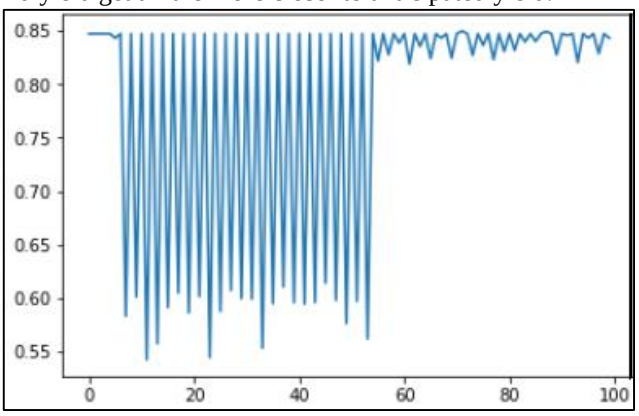

Fig. 3: Backpropagation Algorithm (Accuracy vs Epochs).

C. Feature Selection using Genetic Algorithm

The way toward disposing of the excess highlights and choosing just the imperative highlights from datasets is known as highlight choice. This is a one of a kind procedure that is utilized in AI and furthermore utilized for streamlining problem.The term hereditary calculation is identified with the Gens of the human being.The genes are exchanged starting with one age then onto the next generation.At each progression, hereditary calculation select irregular populace and utilize these arbitrary populace to create the populace for the following generation.Over and over progressive ages, the populate ”develops” towards an ideal arrangement. We can apply the genetic algorithm in many optimization problems.

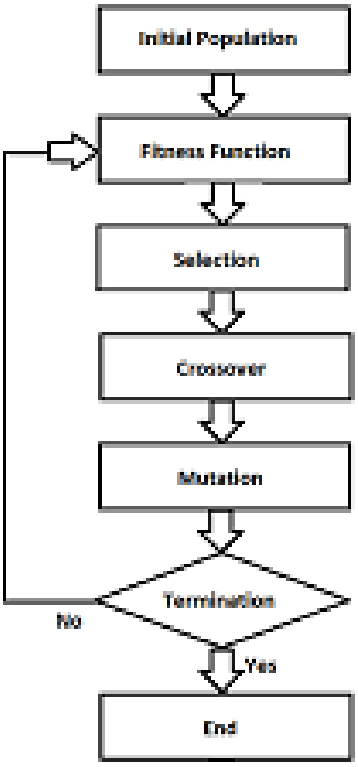

Fig. 4: Genetic Algorithm architecture.

The genetic algorithm has total of five phases.

1) Stage 1-Initial Phase: It is the main period of the hereditary calculation . In this we haphazardly select some populace and give them opportunity to go to the people to come.
2) Stage 2-Fitness Function: It decides how fit an individual is (the capacity of a person to rival different people to find the opportunity to move to the people to come). It gives a wellness score to every person.
3) Stage 3-Mutation: It is the pursuit system utilized in the creation of a freak vector.
4) Stage 4-Selection: DE utilizes choice to coordinate the hunt towards the planned area.
5) Stage 5-Crossover: It is a component of probabilistic and valuable exchange of data among answers for find better arrangements.

IV. OUR APPROACH

Our approach was to try different algorithms and try and compare which one of will works better for the given dataset of heart disease. The algorithms that we have compared are BackPropagation, and Genetic Algorithm on the given vector representation of Cleveland dataset [4]. For our experimental purposes, we have divided our dataset into 5 parts and then we do the crossover validation. Cleveland dataset has 303 instances and 14 attributes. Our first step is to apply dimensionality reduction and for that purpose we have used the genetic algorithm. In genetic algorithm first we generate some random set of attributes and then on these sets of attributes we find the fitness value with the help of backpropagation algorithm.Then we set the selection criteria for the fitness value and move into the next generation is upto 85%.Because we can easily get the accuracy of 85% with the help of backpropagation algorithm. Then we perform the mutation and crossover to get the child generation from the selected parent. we repeat the process again and again until we reach the termination criteria or get the maximum accuracy upto 95%.

A. Problem

We applied backpropagation on heart disease Cleveland dataset having 13 clinical attributes where we get an accuracy of 82% after that when we iterate the algorithm 100 times and take average of it then we get an accuracy of 84%. But that accuracy wasn't sufficient for us. So we use genetic algorithm with backpropagation as a fitness function which makes hybrid model to optimize the clinical features at run time which reduce the complexity and increase accuracy for heart disease prediction.

V. MATHEMATICAL MODEL

A. Calculation for ideal element determination.

Input: health dataset.
Output: Give the accuracy of the model that check for the heart disease.
Steps:

1) Convert scope of qualities somewhere in the range of 0 and 1 by utilizing the fundamental min - max standardization work.
2) Produce the current - age populace vector

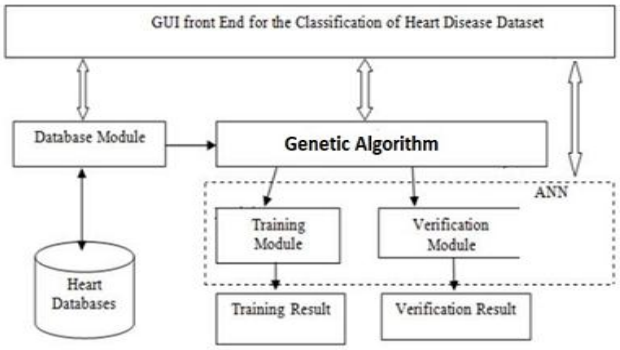

Fig. 5: Architecture of System

Xi,G =x1 i,G,x2 i,G,...,xn i,G ;i = 1,2,...NP.

3) While (next - age vectors are unaltered contrasted this vector and the current - age vectors) .
4) Select the objective of conclusive vector, Xi,G
5) Create the freak vector Mi,G by utilizing change with the last vector Xi,G.
   Mi,G =m1 i,G,m2 i,G,...,mn i,G ;i = 1,2,...NP.
6) Produce the preliminary vector utilizing the hybrid and change technique.
7) On the off chance that the preliminary vector has lesser target work an incentive than the comparing last vector,then the preliminary vector esteem will supplant the last vector esteem.
8) The primary vector of the following - age will be the yield of stage 8 .
9) Rehash this procedure to produce different vectors that structure the following - age vector.
10) End While loop
11) Stop

VI. RESULT

In this paper we have implemented the backpropagation algorithm and genetic algorithm using backpropagation on the cleveland dataset having 13 attributes then we got the accuracy of 85 percents by using backpropagation and we got the accuracy of 93 percents by using the hybrid model of genetic algorithm and the backpropagation.

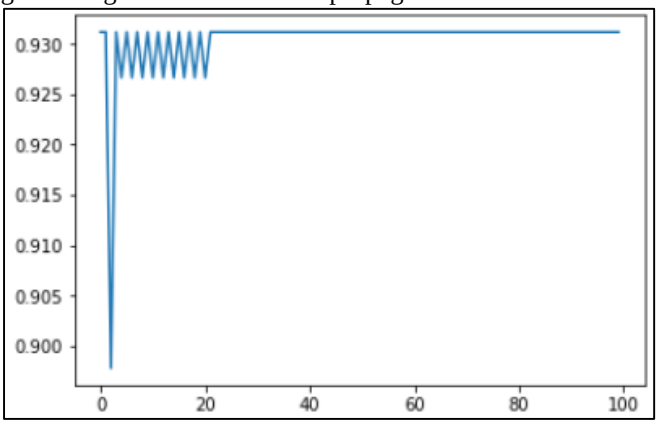

Fig. 6: Genetic Algorithm with Backpropagation.(Accuracy vs Epochs)

VII. CONCLUSION

The proposed heart disease prediction system has been designed as a Multilayer Perceptron Neural Network. For the system dataset we use UCI repository which contain dataset

from four countries Cleveland, Hungarian, Switzerland (Zurich), Switzerland (Basel). The neural network in the system used 13 clinical data as input. It was trained with Backpropagation Algorithm in order to predict whether heart disease present or not in the patient.Also Genetic algorithm is used to reduce the number of features and increase the accuracy. There are a lot of studies on prediction of heart disease. Results of these studies vary up to almost accuracy of 95%.The proposed system gives 92% accuracy rate which means a very good rate according to related studies on this field. As a further study, the proposed methodology can be enhanced as a hybrid model with other classification algorithms in order to obtain more accurate diagnosis for heart disease.

REFERENCES

[1] Z. Shen, M. Clarke, R. W. Jones, T. Alberti, "Detecting the risk factors of coronary heart disease by use of neural networks", Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 277-278, 1993.
[2] Index of/ ml/ machine-learning-databases Index of/ ml/ machine-learning databases, [online] Available: http:// archive.ics.uci.edu/ml/machinelearning-databases/heart-disease/heart-disease.names.
[3] Rami N. Khushaba, Ahmed AL-Ani, Akram Al Sukker, Adel Al-Jumaily, A combined ant colony and differential evolution feature selection algorithm, in: International Conference on Ant Colony Optimization and Swarm Intelligence, 2008, pp. 2936.
[4] I. Kadi, A. Idri, J.L. Fernandez-Aleman, Knowledge discovery in cardiology: a systematic literature review, Int. J. Med. Inf. 97 (2017) 1232.
[5] Key Facts about Heart disease, World Health Organization, Headquarters (Geneva), June 2017. http://www.who.int/mediacentre/factsheets/fs317/en/.
[6] [48] A.L. Bui, T.B. Horwich, G.C. Fonarow, Epidemiology and risk profile of heart failure, Nat. Rev. Cardiol.
[7] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications (0975 888), Volume 47 No.10, pp.44-48, June 2012.
[8] Beant Kaur h, Williamjeet Singh, Review on Heart Disease Prediction System using Data Mining Techniques, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, pp.3003-08,October 2014.

