International Conference on Intelligent Technologies & Science - 2022
(ICITS-2022)

Heart Disease Classification using Decision Tree and ANN

Shristhi Gupta[1], Ram Kailash Gupta[2]
Computer science and Engineering Department,
Bansal Institute of Engineering and Technology, Lucknow
shristhigupta97@gmail.com

Abstract: By using AI techniques, numerous investigations into the sequence of cardiac sickness have been completed. The informational collection is available on the uci AI store for creating the AI model. The Cleveland dataset is the most widely used informative collection. The Cleveland informational index has 303 records and 14 specific attributes, including age, sex, chest pain type (4 qualities), resting pulse, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0, 1, 2), most extreme pulse achieved, practice-induced angina, old pinnacle = ST despondency triggered by practise compared with rest, the slant of the pinnacle practise ST portion, and number of significant vessels (0–3) (the anticipated characteristic). These 14 credits are often used by the analysts out of the initial 72 traits in the informational index.

Keywords: Cardiovascular Disease, Decision Support System, Data Mining, Hybrid Intelligent System

1. Introduction:

Twenty million people die each year from coronary sickness, which is the leading cause of death worldwide (WHO)[26]. For people with cardiac sickness, an accurate and prompt diagnosis could be the difference between life and death. Nevertheless, according to an English Clinical Announcement [1], doctors incorrectly identify roughly one-third of patients as not having coronary disease, causing them to forgo potentially curative care. Given that the American Heart Association predicts a 46% increase in cardiovascular breakdown cases by 2030, this is a growing cause for concern. Any doctor will find it challenging to diagnose coronary illness because, while fatigue and chest pain are common side effects of atherosclerosis, up to 50% of people don't experience any symptoms of coronary illness until they experience their first respiratory failure (Community for Infectious prevention) [3]. Finding biomarkers (for coronary illness)—quantitative indicators of the severity or presence of a disease—is encouraged [4], but in most cases there are no reliable biomarkers, necessitating a battery of testing and joint investigation. Most doctors use the American Heart Association's (AHA) recommended guidelines, which evaluate eight commonly recognised risk factors such high blood pressure, cholesterol, smoking, and diabetes [5]. However, this hazard evaluation model is flawed because it assumes a direct relationship between each risk factor and the outcome of coronary sickness, even if the links are convoluted and involve indirect relationships [7]. Oversimplification could cause physicians to misjudge their expectations or ignore important factors that could determine whether a patient receives treatment. Additionally, experts should be able to interpret the analytical results of clinical studies, which vary between patients and call for extensive knowledge. Utilizing AI (ML) information research techniques can increase prediction accuracy while reducing the need for human judgement and the risk of human error[8]. In light of the information dataset's learned relationships between various parameters, ML algorithms apply adaptable forecast models. This can prevent fixed finding models like the AHA rules from being oversimplified. In fact, it has been shown that a brain network calculation with 76% accuracy can accurately predict 7.6% more occurrences than the AHA technique [9].

2. Related Work:

Artificial intelligence (AI) computations using relapse tree, backward vector machine (SVM), and fake brain organisations (ANN) are studied for assessing and forecasting the severity of cardiovascular breakdown. Last but not least, the inventors claimed that SVM outperforms the following two calculations. SVM, which provides a precision of 94.60% [8] and is also more precise and makes less errors in sickness prediction [9] and [10], is the ideal methodology for coronary illness demonstrative critical initiatives produced by inventors.

Seven AI calculations, including Gullible Bayes, Choice trees, K-Closest Neighbor, Multi-facet perceptrons, Spiral premise capabilities, single conjunctive students, and SVM, were used in HD Forecast to assess and explore 302 events. Creator investigated a variety of possibilities in relation to the events and discovered that the SVM technique worked effectively [11]. In the HD conclusion, SVM algorithms are also used for diabetes patients [12].

Uncommonly designed for mobile phones and used for checking coronary sickness, versatile AI model aids in seeing HD. 200 patients' clinical records were examined from various angles, and it was successful in obtaining a precision of 90.5%. [13]. To predict HD, research is conducted on the execution analysis of various ML calculations[14]. A patient's HD risk level is predicted using ML computations, and a unified Framework has also been created so that both patients and clinicians can access the cloud-based e-wellness data[15].

A decision tree analysis is used to find a predictive model for detecting cardiac disease. Informational indicators were

collected from 1159 healthy people and 1187 people who had had coronary angiography. finally claimed that the risk factors for coronary sickness had specificity, awareness, and accuracy of 87%, 96%, and 94%, respectively. to improve the HD presentation setup and forecast Techniques based on trees are used [16]. Compared to AI and information mining strategies, conventional calculated relapse had the option to predict more precise outcomes among the patients in terms of cardiovascular breakdown [17].

3. Methodology:

The most hotly debated topic in the world of AI right now is profound learning, which is receiving a tonne of media attention. Deep learning can be seen as a collection of calculations that were developed to best prepare with numerous layers. The multi-faceted fake brain networks that we will look at in this part have false neurons that address their structural building components. The fundamental notion behind fake brain networks was founded on hypotheses and models of how the human mind tries to handle challenging problem-solving tasks. Even though fake brain networks have been quite popular recently, the first studies of brain networks date back to the 1940s, when Warren McCulloch and Walter Pitt first described how neurons could behave. However, in the many years that followed the initial application of the McCulloch-Pitt neuron model, Rosenblatt's perceptron in the 1950s, many scientists and AI specialists gradually began to lose interest in brain networks because no one had a good solution for creating a brain network with multiple layers. In 1986, D.E. Rumelhart, G.E. Hinton, and R.J. Williams finally rekindled interest in brain networks by (re)discovering and promoting the backpropagation computation to better prepare brain networks. In the previous ten years, much more substantial advances led to the development of what we now refer to as profound learning calculations, which may be used to incorporate markers from unlabeled data to pre-train profound brain organisations, or brain networks with several layers. Brain networks are a contentious topic in academic study as well as in major technology companies like Facebook, Microsoft, and others that have invested much in research on fake brain networks and in-depth learning. In terms of complicated critical thinking, such as picture and voice recognition, complex brain networks controlled by sophisticated learning computations are now regarded as cutting edge. Google's image search and Google Decipher, a mobile app that can intuitively discern message in photographs for continuous interpretation into 20 dialects, are well-known examples of the products in our daily lives that are fuelled by profound learning.

Let's briefly revisit the early AI starting points before delving deeper into the perceptron and related calculations. Warren McCullock and Walter Pitts proposed the basic idea of a worked synapse, the purportedly interconnected nerve cells in the cerebrum that are involved with the handling and sending of chemical and electrical signals, in an effort to understand how the natural mind attempts to plan man-made consciousness. This idea is shown in figure 1.

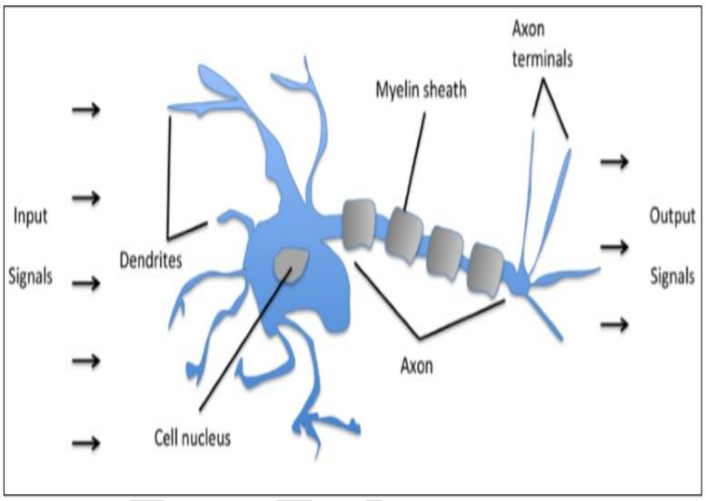


Fig 1: Biological brain cell (neuron)

McCullock and Pitts described such a nerve cell as a simple rational entryway with two results: first, different signs arrive at the dendrites, are then incorporated into the phone body, and, if the total number of signs exceeds a certain threshold, a result signal is generated and transmitted by the axon. Candid Rosenblatt published the primary idea of the perceptron learning rule in light of the MCP neuron model within a few years after it was first proposed (F. Rosenblatt, The Perceptron, a Seeing and Perceiving Robot. Cornell Aeronautical Research facility, 1957). Rosenblatt's perceptron rule is a calculation that would logically become familiar with the optimum weight coefficients, which are then replicated with the information highlights to decide whether or not a neuron fires. Such a formula might therefore be used to predict whether an example belonged with one class or the other in terms of controlled learning and grouping. To describe this problem more formally, we may frame it as a parallel characterisation task and refer to our two classes as 1 (positive class) and - 1 (negative class) for simplicity. Then, where z is the alleged net info, we can define an enactment capability (z) that accepts a direct blend of particular data values x and a comparing weight vector w. The perceptron gathers the contributions of an example x and joins them with the loads w to analyse the network information, as shown in figure 2. The initiation capability (in this case, the unit step capability), after receiving the net information, generates a double result of 1 or +1, which corresponds to the example's expected class grade. This outcome is used to correct the expectation error and update the weights throughout the learning phase. A quantizer, which is similar to the unit step capability we have previously seen, can then be used to predict the class names while the direct initiation capacity is being used to learn the loads.

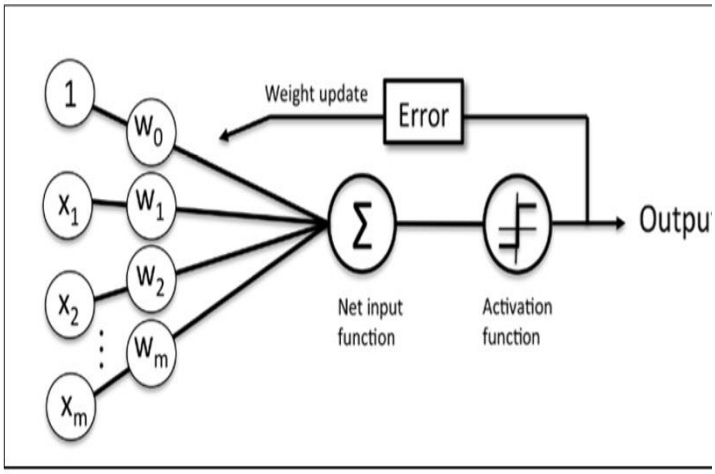
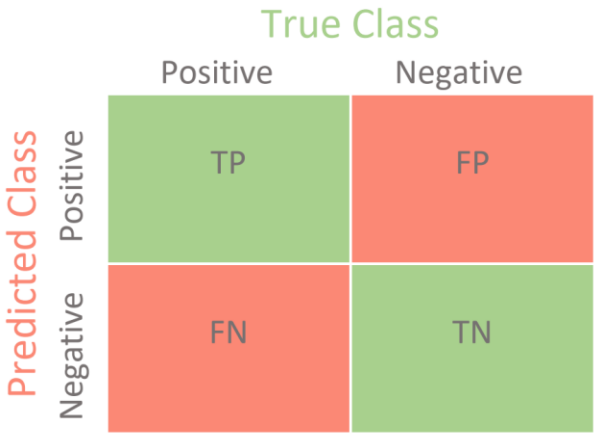


Fig 2. Simple perceptron's in neural network

Proposed Work:

In machine learning problems, Confusion Matrix is generally used for statistical analysis of data. It is also known as the error matrix. A confusion matrix [15] basically denotes the relationship between the actual and the predicted results of a classifier. Classifier performance can be viewed based on the values obtained in the confusion matrix.As the name suggests it denotes how much the model is confused when it makes predictions on the test set. If we have a binary classifier the confusion matrix will be a 2X2 matrix which can be represented as follows. Assume that there are two classes positive and negative.



Where TP represents True Positive It depicts the number of tuples of positive that have been correctly predicted.FN represents False Negative: It depicts the number of tuples of Class 1 that have been incorrectly predicted to be in negative.FP represents False Positive: It depicts the number of tuples of negative that have been predicted to be in positive.TN represents True Negative: It depicts the number of tuples of positive that have been correctly predicted to be in negative.

4. Result and Discussion:

The exploratory data analysis is performed on the data set to find the variance of the data set .the exploratry data analysis is the most important phase of data analysis by using which the data is interprited and gives a better understanding for the type of model to be used with the data . the vaious features of the feature set is plotted by using count plots and histograms so that we can check the kind of data and data distribution availabe.the below image shows the analysis of each columns of the data set.

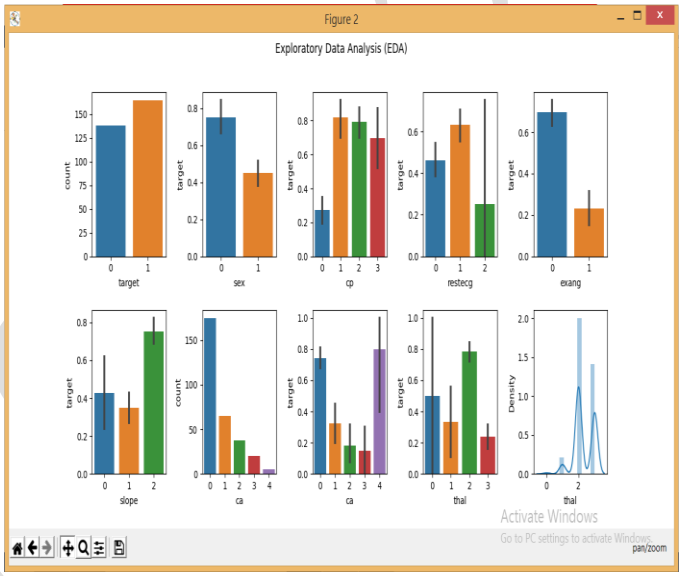


Fig 3: exploratory data analysis of the data set

The data is skewd for the removal of skewness of data we can perform either over sampling or under sampling but in our case we have only 310 rows in our data hence we will prefer the over sampling by using the ADASYN over sampler. The data is preproccesd using the min max scalar, the the data is divided into 80:20 ratio as the training set and testing set. The train data set is firstly treated for the data skewness and then the data set is used to traing a decision tree so that the important features of the data set can be obtained by building the decision tree. The clolumns used as the nodes of the tree is used as the important features . for the optimization of the model we have used the random search cv as the hyper parameter tuning model the ANN model is optimized for the activation layers and the loss model. The ANN model used contains three different layers the first layer is take the input of 13 columns and it is transformed to 10 by uising the relu activation function,the second layer converts the 10 columns into 8 columns by using the relu activation function then finally the 8 input is usedto genrate one output by using the activation function of sigmoid. The values predicted is between 0 to 1 then by round function we have converted the values back to either 0 or 1. Then by using the different metrics system we have calculated the precision, accuracy, recall, f1-score. The same data is used to train the various other machine learning models which were previously to build the models the result of metrics of each model is shown in the

below figure. The machine learning models are also hyper tuned for different parameters in the different models like k neaighbors is tuned for the n neighbours concept, the optimal value for k neighbors is found to be seven. The decsion tree and the random forest are hyper tuned for the concept of random_state , the optimal value for random state in decision tree is 11 and in case of random forest it is 323. The logistic regression model, suppportvector machine algoritms are tuned for the parameters of the learning rate. For the implementation of the ANN model wehave used the tensorflow and the keras modules and for other models the sklearn module is used. The result of each model is shown in the below figure. The results clearly shows that the ANN model has the highest accuracy as compared to other models. and the k neighbours has the least accuracy. The Ann model resulted a accuracy of 92.16% accuracy, 94.12 % precision, 88.89% recall, 91.43% f1-score.

heart disease prediction using ANN

ALGORITHMS	ACCURACY	PRECISION	RECALL	F1-SCORE
Logistic Regression	85.25	88.24	85.71	86.96
Naive Bayes	85.25	91.18	83.78	87.32
Support Vector Machine	81.97	88.24	81.08	84.51
K-Nearest Neighbors	67.21	67.65	71.88	69.7
Decision Tree	81.97	82.35	84.85	83.58
Random Forest	88.24	87.41	85.85	90.0
ANN	92.16	94.12	88.89	91.43

Fig 4: results of the ANN model and other models.

The results obtained are graphicaly shown in the below figure for the better result analysyis for implementing the graphical representation of the results we have used the matplotlib and the seaborn library.

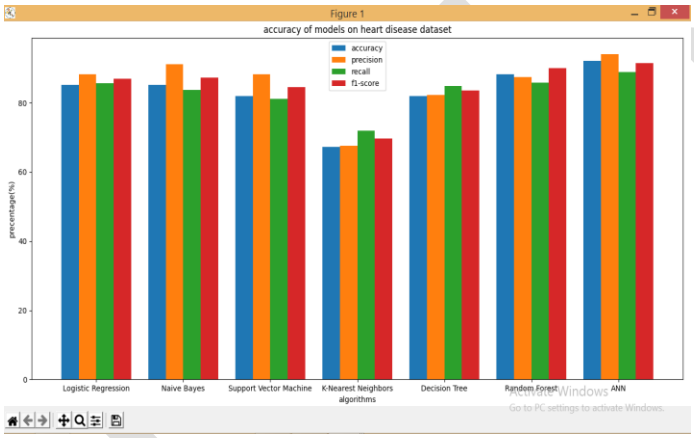


Fig 5: results analysis

5. Conclusion:

In our work we have proposed an ANN-DT (Artificial neural network decision tree) based model for predicting the heart disease. Previously various Machine learning, Datamining and Neural Network models were employed for the same. Machine learning and the Datamining did not have a good method for the selection of the important features. We have used the Decision tree as the model for the feature selection. And the selected features are used to train the ANN model. The ANN model was built by using the relu action function, sigmoid function, Adams solver and the binary entropy loss function. The ANN approach has outperform the existing techniques for the heart disease prediction. We have done the proper pre-processing of data by using the scalers and the skewness is removed by using the ADASYN over sampler.

References:

[1] Nidhi Bhatla, and Kiran Jyoti, Oct. 2012,"An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), Vol. 1, Issue 8, ISSN: 2278-0181, pp. 1-4.

[2] Sumitra Sangwan, and Tazeem Ahmad Khan, Mar. 2015,"Review Paper Automatic Console for Disease Prediction using Integrated Module of A-priori and k-mean through ECG Signal",International Journal For Technological Research In Engineering, Vol. 2, Issue 7, ISSN(Online): 2347-4718, pp. 1368-1372.

[3] Rishi Dubey, and Santosh Chandrakar, Aug. 2015,"Review on Hybrid Data Mining Techniques for The Diagnosis of Heart Diseases in Medical Ground" ,Vol. 5, Issue 8, ISSN: 2249-555X, pp. 715-718.

[4] Ashish Chhabbi,Lakhan Ahuja,Sahil Ahir, and Y. K. Sharma,19 March 2016,"Heart Disease Prediction Using Data Mining Techniques", International Journal of Research in Advent Technology,E-ISSN:2321-9637,Special Issue National Conference "NCPC-2016", pp. 104-106.

[5] Shaikh Abdul Hannan, A.V. Mane, R. R. Manza,and R. J. Ramteke, Dec 2010,"Prediction of Heart Disease Medical Prescription using Radial Basis Function",IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2010.5705900 ,28-29 .

[6] AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "HDPS: Heart Disease Prediction System",Computing in Cardiology, ISSN: 0276-6574, pp.557-560. [7] Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010,"Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network",International Conference on Computer and Communication Technology (ICCCT),DOI:10.1109/ICCCT.2010.5640377, 17-19.

[8] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map",IEEE International Conference on Fuzzy Systems (FUZZ),pp. 1377-1382.

[9] Carlos Ordonez,2006, "Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine (TITB), pp. 334-343, vol. 10, no. 2.

[10] Prajakta Ghadge,Vrushali Girme, Kajal Kokane, and Prajakta Deshmukh, 2016,"Intelligent Heart Attack Prediction System Using Big Data", International Journal of Recent Research in Mathematics Computer Science and Information Technology,Vol. 2, Issue 2, pp.73-77, October 2015–March.

[11] Asha Rajkumar, and Mrs G. Sophia Reena, 2010, "Diagnosis of Heart Disease using Data Mining Algorithms",Global Journal of Computer Science and Technology,Vol. 10,Issue 10, pp.38-43, September.
[12] K. S. Kavitha, K. V. Ramakrishnan, and Manoj Kumar Singh, September 2010,"Modelling and Design of Evolutionary Neural Network for Heart Disease Detection" ,International Journal of Computer Science Issues (IJCSI), Vol. 7, Issue 5, pp. 272-283.
[13] K.Sudhakar, and Dr. M. Manimekalai, January 2014, "Study of Heart Disease Prediction using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 1,pp. 1157-1160.
[14] Shantakumar B. Patil,and Dr. Y. S. Kumaraswamy, February 2009, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", IJCSNS International Journal of Computer Science and Network Security, Vol. 9 , No. 2,pp. 228-235.
[15] Sairabi H. Mujawar, and P. R. Devale, October 2015,"Prediction of Heart Disease using Modified k-means and by using Naive Bayes", International Journal of Innovative Research in Computer and Communication Engineering(An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, pp. 10265-10273.
[16] S. Suganya, and P.Tamije Selvy, January 2016, "A Proficient Heart Disease Prediction Method using Fuzzy-Cart Algorithm", International Journal of Scientific Engineering and Applied Science (IJSEAS), Vol. 2, Issue1,ISSN: 2395-3470.
[17] Ashwini Shetty A, and Chandra Naik, May 2016,"Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative Research in Science,Engineering and Technology(An ISO 3297: 2007 Certified Organization), Vol. 5, Special Issue 9, pp. 277-281.
[18] K Cinetha, and Dr. P. Uma Maheswari, Mar.-Apr. 2014,"Decision Support System for Precluding Coronary Heart Disease using Fuzzy Logic.", International Journal of Computer Science Trends and Technology (IJCST), Vol. 2, Issue 2, pp. 102-107.
[19] Indira S. Fal Dessai,2013,"Intelligent Heart Disease Prediction System Using Probabilistic Neural Network", International Journal on Advanced Computer Theory and Engineering (IJACTE),Vol. 2, Issue 3, pp. 38-44.
[20] Mai Shouman, Tim Turner, and Rob Stocker, June 2012,"Applying k-Nearest Neighbors in Diagnosing HeartDiseasePatients",International Journal of Information and Education Technology, Vol. 2, No. 3,pp. 220-223.
[21] Serdar AYDIN, Meysam Ahanpanjeh,and Sogol Mohabbatiyan,February 2016, "Comparison And Evaluation of Data Mining Techniques in the Diagnosis of Heart Disease",International Journal on Computational Science & Applications (IJCSA), Vol. 6,No.1, pp. 1-15.