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

Heart Disease Classification using Decision Tree and ANN

Shailendra[1], Rahul Gupta[2]
Computer Science and Engineering,
S. R. Institute of Management and Technology, Lucknow, India
786unique.shailendra@gmail.com

Abstract: The keys issues with the previous research works were improper pre-processing of the data due to which the data was skewed and contained larger numbers. Hence in this paper the model would be biased due to the skewness of the data and it would take more time to process the data due to not scaling the data and it makes the model more complex to learn the linear and non-linear relation between the features. The data were directly used to train the models without even hyper tuning the models, without hyper tuning the model would be build the models on the default parameters hence makes it more difficult to respond according to the data used for training purpose. Feature extraction is most important portion of any learning process.

Keyword: ANN, AHA, Deep Learning, Heart issue.

1. Introduction:

Heart disease is the leading cause of death worldwide, killing twenty million people per year (WHO)[26]. An accurate and early diagnosis could be the difference between life and death for people with heart disease. However, doctors misdiagnose nearly 1/3 of patients as not having heart disease (British Medical Bulletin) [1], causing these patients to miss out on potentially lifesaving treatment. This is a problem of increasing concern, as the number of Americans with heart failure is expected to increase by 46 percent by 2030 (American Heart Association)[2]. What makes diagnosing heart disease a challenging endeavor for any physician is that while chest pain and fatigue are common symptoms of atherosclerosis, as many as 50 percent of people lack any symptoms of heart disease until their first heart attack (Center for Disease Control) [3]. Discovering biomarkers (for heart disease) – measurable indicators of the severity or presence of some disease – is preferred [4], but in many cases there are no clear biomarkers, and multiple tests may berequired and analyzed together. Most doctors use the guidelines recommended by the American Heart Association (AHA) [5], which test eight widely recognized risk factors such as hypertension, cholesterol, smoking, and diabetes [6]. However, this risk assessment model is flawed since it rests on an assumed linear relationship between each risk factor and heart disease outcome, while the relationships are complex and with non-linear interactions [7]. Oversimplification may cause doctors to make errors in their predictions, or overlook important factors that could determine whether or not a patient receives treatment. Doctors must also know how to interpret the diagnostic implications of medical tests which vary between patients and require tremendous expertise. The use of Machine Learning (ML) data analysis techniques can alleviate the need for human expertise and the possibility of human error while increasing prediction accuracy[8]. ML algorithms apply flexible prediction models based on learned relationships between variables in the input dataset. This can prevent the oversimplification of fixed diagnosis models such as the AHA guidelines. In fact, a neural network algorithm with 76 percent accuracy has been proven to correctly predict 7.6% more events than the AHA method [9].

2. Related Work:

Many research have been done for the classification of Coronary heart disease by using the machine learning techniques. For training the machine learning model the data set is available on the uci machine learning repository. Mostly used data set is the Cleveland dataset. The Cleveland data set contains 303 records with 14 attributes namely age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, old peak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) coloured by fluoroscopy , thal: 3 = normal; 6 = fixed defect; 7 = reversible defect, num (the predicted attribute). Originally the data set contains 72 attributes out of which these 14 attributes are commonly used by the researchers.

Machine learning algorithms are investigated for assessing and predicting the severity of heart failure by artificial neural networks (ANN), Support vector machine (SVM), classification and regression tree. Finally, out of these algorithms the authors claimed SVM performs better than the other two algorithms [56]. In the heart disease diagnostic significant attempts are made by authors, best approach is SVM which provides an accuracy of 94.60% [8] and also SVM approach is more accurate and less errors in disease prediction [9] and [10].

In HD Prediction 302 instances were compared and investigated using seven machine learning algorithms such as Naïve Bayes, Decision tree, K-Nearest Neighbor, Multilayer perceptron, Radial basis function, single conjunctive learner and SVM. Author has done experiment with the instances and resulted SVM method performed well [11]. SVM methods also used for diabetic patients in HD diagnosis [12].

Mobile Machine Learning Model is designed for Monitoring Heart Disease, specially designed for mobile devices which helps to monitor HD. Experimented with clinical datasets of 200 patients and was successful in obtaining an accuracy of 90.5%[13].Research is carried out on performance analysis of various ML algorithms to predict HD[14].HD risk level of a patient is predicted using ML algorithms and also created a centralized System to view the e-health data on cloud for both patients and doctors[15].

3. Methodology:

Deep learning is getting a lot of press and is without any doubt the hottest topic in the machine learning field. Deep learning can be understood as a set of algorithms that were developed to train with many layers most efficiently. Artificial neurons represent the building blocks of the multi-layer artificial neural networks that we are going to discuss in this chapter. The basic concept behind artificial neural networks was built upon hypotheses and models of how the human brain works to solve complex problem tasks. Although artificial neural networks have gained a lot of popularity in recent years, early studies of neural networks go back to the 1940s when Warren McCulloch and Walter Pitt first described how neurons could work. However, in the decades that followed the first implementation of the McCulloch-Pitt neuron model, Rosenblatt's perceptron in the 1950s, many researchers and machine learning practitioners slowly began to lose interest in neural networks since no one had a good solution for training a neural network with multiple layers. Eventually, interest in neural networks was rekindled in 1986 when D.E. Rumelhart, G.E. Hinton, and R.J. Williams were involved in the (re)discovery and popularization of the backpropagation algorithm to train neural networks more efficiently. During the previous decade, many more major breakthroughs resulted in what we now call deep learning algorithms, which can be used to create feature detectors from unlabelled data to pre-train deep neural networks—neural networks that are composed of many layers. Neural networks are a hot topic not only in academic research, but also in big technology companies such as Facebook, Microsoft, and Google who invest heavily in artificial neural networks and deep learning research. As of today, complex neural networks powered by deep learning algorithms are considered as state-of-the-art when it comes to complex problem solving such as image and voice recognition. Popular examples of the products in our everyday life that are powered by deep learning are Google's image search and Google Translate, an application for smartphones that can automatically recognize text in images for real-time translation into 20 languages.

Before we discuss the perceptron and related algorithms in more detail, let us take a brief tour through the early beginnings of machine learning. Trying to understand how the biological brain works to design artificial intelligence, Warren McCullock and Walter Pitts published the first concept of a simplified brain cell, the so-called interconnected nerve cells in the brain that are involved in the processing and transmitting of chemical and electrical signals, which is illustrated in the figure 1.

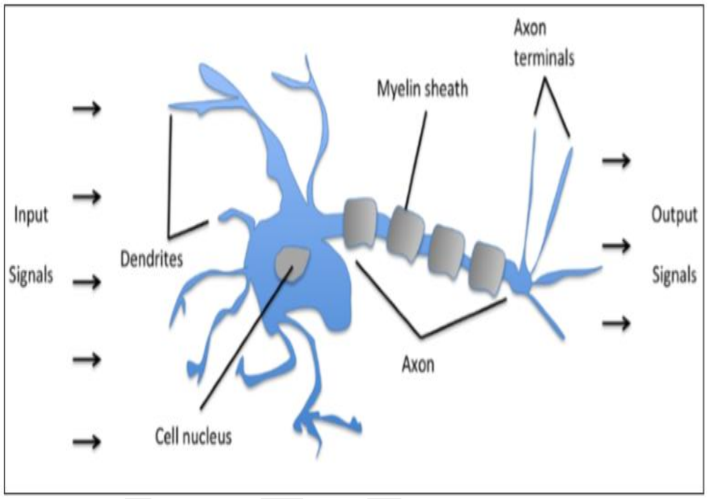


Fig 1: biological brain cell (neuron)

The proposed model includes 2 stages in which the first stage is to generate an optimized model for prediction of the presence of heart disease in the patient. The block diagram is as shown in the figure 4.For building the optimized model we use the dataset discussed above. The dataset is treated for the missing values by using the NB (naïve Bayes) technique.Naïve Bayes is a probabilistic model based on the Bayes' theorem [13].

4. Result and Discussion:

The proposed methodology is implemented on python 3.6. Python is a interpreted high level programming language which is been used for development of websites by using the Django framework of python, in field of AI (artificial intelligence) ,and also used in the field of IOT (internet of things) and in the field of research and analytics. For implementation of our model we have used the various tools of python like Tensorflow , Keras, sklearn, pandas, imblearn, matplotlib, seaborn, jupyter notebook. For the frontend is developed by using the tkinter tool available in python. The dataset is obtained from the uci machine learning repository website.

For implementation of model we have designed a user interface. The user interface is divided in 3 parts the first part is for the exploratory data analysis, the second part is for the training the models on the heart disease dataset and to acquire the result , the last part is for the result analysis which includes the generation of accuracy, precision, recall and f1-score and the visualization of the results. The user interface is shown in the below figure for the implementation of 3 different phases of the model we have used different buttons through which we can navigate through the different portion.

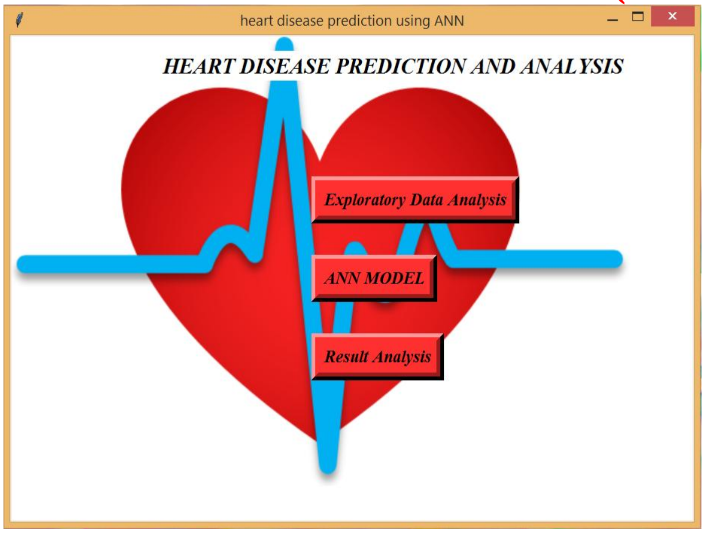
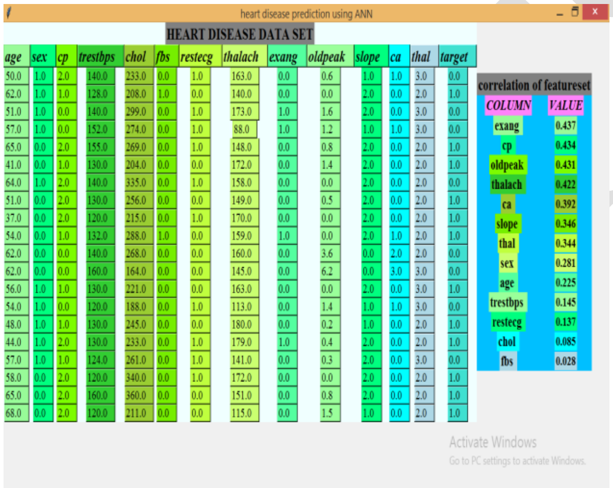

heart disease prediction using ANN
HEART DISEASE PREDICTION AND ANALYSIS
Exploratory Data Analysis
ANN MODEL
Result Analysis

Fig. 2: The default GUI page

For initialling the first phase of exploratory data analysis we need to click on the first button. Once the button is clicked the data set of heart diseases is loaded by using the pandas library. The data is treated for the missing values and the Pearson correlation is calculated among the feature set and the target label. The dataset is and the correlation is as in the below figure. The correlation is used for the feature selection only the columns which have some correlation with the target label is used for training the models.


heart disease prediction using ANN
Activate Windows
Go to PC settings to activate Windows.

HEART DISEASE DATA SET

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
50.0	1.0	2.0	140.0	233.0	0.0	1.0	163.0	0.0	0.6	1.0	1.0	3.0	0.0
62.0	1.0	1.0	128.0	208.0	1.0	0.0	140.0	0.0	0.0	2.0	0.0	2.0	1.0
51.0	1.0	0.0	140.0	299.0	0.0	1.0	173.0	1.0	1.6	2.0	0.0	3.0	0.0
57.0	1.0	0.0	152.0	274.0	0.0	1.0	88.0	1.0	1.2	1.0	1.0	3.0	0.0
65.0	0.0	2.0	155.0	269.0	0.0	1.0	148.0	0.0	0.8	2.0	0.0	2.0	1.0
41.0	0.0	1.0	130.0	204.0	0.0	0.0	172.0	0.0	1.4	2.0	0.0	2.0	1.0
64.0	1.0	2.0	140.0	335.0	0.0	1.0	158.0	0.0	0.0	2.0	0.0	2.0	0.0
51.0	0.0	2.0	130.0	256.0	0.0	0.0	149.0	0.0	0.5	2.0	0.0	2.0	1.0
37.0	0.0	2.0	120.0	215.0	0.0	1.0	170.0	0.0	0.0	2.0	0.0	2.0	1.0
54.0	0.0	1.0	132.0	288.0	1.0	0.0	159.0	1.0	0.0	2.0	1.0	2.0	1.0
62.0	0.0	0.0	140.0	268.0	0.0	0.0	160.0	0.0	3.6	0.0	2.0	2.0	0.0
62.0	0.0	0.0	160.0	164.0	0.0	0.0	145.0	0.0	6.2	0.0	3.0	3.0	0.0
56.0	1.0	1.0	130.0	221.0	0.0	0.0	163.0	0.0	0.0	2.0	0.0	3.0	1.0
54.0	1.0	0.0	120.0	188.0	0.0	1.0	113.0	0.0	1.4	1.0	1.0	3.0	0.0
48.0	1.0	1.0	130.0	245.0	0.0	0.0	180.0	0.0	0.2	1.0	0.0	2.0	1.0
44.0	1.0	2.0	130.0	233.0	0.0	1.0	179.0	1.0	0.4	2.0	0.0	2.0	1.0
57.0	1.0	1.0	124.0	261.0	0.0	1.0	141.0	0.0	0.3	2.0	0.0	3.0	0.0
58.0	0.0	2.0	120.0	340.0	0.0	1.0	172.0	0.0	0.0	2.0	0.0	2.0	1.0
65.0	0.0	2.0	160.0	360.0	0.0	0.0	151.0	0.0	0.8	2.0	0.0	2.0	1.0
68.0	0.0	2.0	120.0	211.0	0.0	0.0	115.0	0.0	1.5	1.0	0.0	2.0	1.0

correlation of featureset

COLUMN	VALUE
exang	0.437
cp	0.434
oldpeak	0.431
thalach	0.422
ca	0.392
slope	0.346
thal	0.344
sex	0.281
age	0.225
trestbps	0.145
restecg	0.137
chol	0.085
fbs	0.028

Fig 3: First 21 rows of the heart disease dataset

The calculated correlation is visualized by using the heat map , the heat map uses various color codes to represent the correlation among the columns. For calculation of the correlation the pandas library is used and the heat map is plotted by using the seaborn liubrary . the heat map is shown in the figure below.

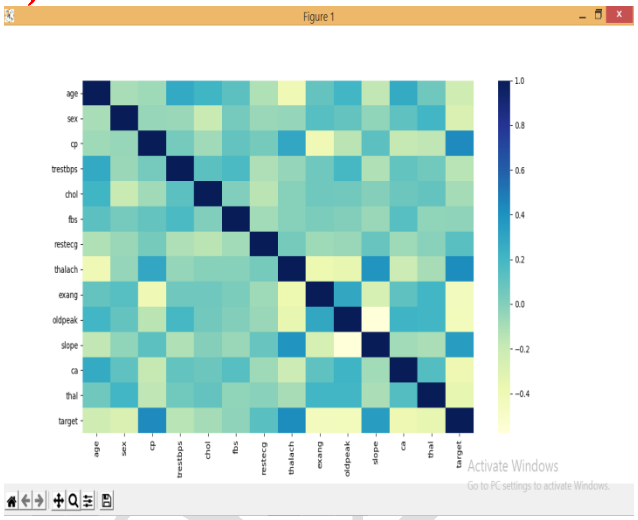


Fig 4: Heat map of data correlation

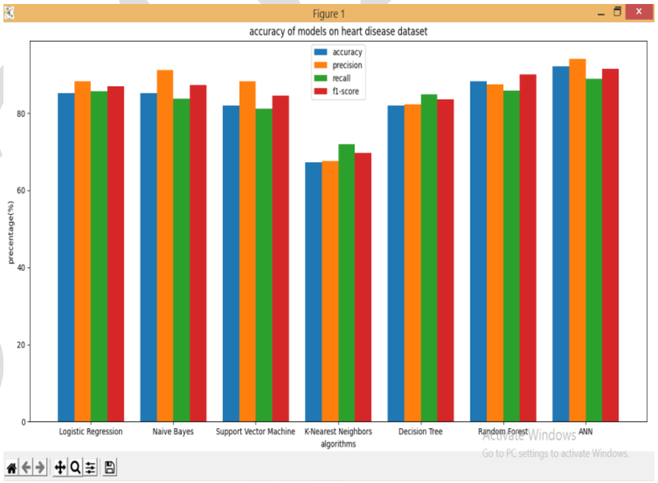


Fig 5: Results analysis

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]. Cheryl Ann Alexander and Lidong Wang(2017)‘ Big Data Analytics in Heart Attack Prediction‘, Journal of Nursing and Care , Volume 6 ,Issue 2 , ISSN:2167-1168 .

[2]. J.Archenaa and E.A. MaryAnita (2015)'A Survey of Big Data Analytics in Healthcare and Government', Elsevier Procedia Computer Science, Volume 50, 2015, Pages 408-413.

[3]. Roheet Bhatnagar (2018) _Machine Learning and Big Data Processing: A Technological Perspective and Review' , The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA20), pp.468-47818.

[4]. Apoorva Sharma, Pallavi Rawat, Kajal Pandey, Ravi Shankar Rai (2017), _Health Analytics Using Machine Learning: A Survey', International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 4, PAGES 6650-6657.

[5]. Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy (1996) _From Data Mining to Knowledge Discovery: An Overview', Advances in Knowledge Discovery and DataMining,pp.1-34.

[6]. N. K. Salma Banu , Suma Swamy . (2016) _ Prediction of Heart Disease at Early Stage Using Data Mining and Big Data Analytics: A Survey', IEEE international conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), INSPEC Accession Number: 16981012,DOI: 10.1109/ICEECCOT.2016.7955226.

[7]. Goodfellow, Y. Bengio, and A. Courville.(2016) _Deep learning', MIT Press. Sharma Purushottam, Dr. Kanak Saxena, Richa Sharma, "Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree", Springer, Computational Intelligence in Data Mining,vol.2, 2015, pp.285-294

[8]. Meherwar Fatima,Maruf Pasha.(2017) _Survey of Machine Learning Algorithms for Disease Diagnostic', Journal of intelligent learning systems and applications, Vol. 09. pages 10.4236/jilsa.2017.91001.

[9]. Akhila C S,Vidya M.(2017) A Survey on Machine Learning Approaches for Disease Predicting System',Journal of Emerging Technologies and Innovative Research,Volume 4,Issue 6,pages 203-205.

[10]. Gagan Kumar,Rohit Kalra.(2016)_A Survey on Machine Learning Techniques in Health Care Industry', International Journal of Recent Research Aspects, ISSN: 2349-7688 Vol. 3 Issue ,pp. 128-132.

[11]. Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez. (2017) _A Comprehensive Investigation and Comparison of Machine Learning Techniques in The Domain of Heart Disease', Published in: Computers and Communications (ISCC), 2017 IEEE Symposium on 3-6 July 2017, DOI: 10.1109/ISCC.2017.8024530.

[12]. G.Parthiban,S.K.Srivatsa.(2012) _Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients', International Journal of Applied Information Systems (IJAIS), ISSN : 2249-0868 , Volume 3– No.7.

[13]. Omar Boursaliea, Reza Samavia, Thomas E. Doylea. (2015) _M4CVD: Mobile Machine Learning Model for Monitoring Cardiovascular Disease', The 5th International Conference on Current and Future Trends of Information andCommunication Technologies in Healthcare- Elsevier Computer Science, Pages:384 – 391.

[14]. M. Chandralekha and N. Shenbagavadivu .(2018) _Performance Analysis of Various Machine Learning Techniques to Predict Cardiovascular Disease: An Empirical Study', Applied mathematics and information sciences, Appl. Math. Inf. Sci. 12, No. 1, 217-226.