# An Approach to Predict Heart Diseases Using Data Mining Techniques

**[1]Javed Akhtar, [2]D. L. Gupta**

[1]M. Tech Scholar, [2]Associate Professor
Department of Computer Science and Engineering
Kamala Nehru Institute of Technology, Lucknow, India

***Abstract:*** **Heart disease is a well-known cause of death and as we know many on the human in the world having breathing difficulties due to change in human lifestyle and change in the environment. Many of these strokes occur when not being flagged positively as a heart patient so there is a need for more accurate and precise methods for heart disease prediction. So in this research, I am solving these difficulties using emerging technologies like Deep Learning. There is a chance we get good results in terms of speed and Accuracy of well build model architecture.**

***Index Terms*****: Deep neural networks, binary cross entropy, regularization, hyper parameter, accuracy.**

## I. Introduction:

According to the World Health Organization (WHO), in 2015, cardiovascular diseases caused 31% of all global deaths in one year, heart disease is a well-known cause of deaths wold wide. By observing the public health statistics we came to know an increase of patients with some form of cardiovascular disease in countries with low or middle gross national income is increasing in countries like India. Although very serious and often life-threatening, cardiovascular disease in individuals can be managed in clinics as a chronic condition, and treated with medicines available in the market, diet, regular monitoring of specific health indicators and working on body fitness like ancient techniques yoga. Risk factors are fairly well defined and lifestyle changes can mitigate some risks of flagging as a heart patient. The motivation to prevent and manage heart disease has started the development of numerous Health applications for consumer use, some of which have been scientifically approved and tested for efficiency. In this paper, we provide deep learning techniques in these systems. And then we solve issues and give solutions associated with heart disease using deep learning algorithms for medical applications with higher accuracy. A conclusive and early diagnosis of heart disease could be the difference between life and death for some which will be possible due to deep learning techniques. In daily practice, one out of 3 patients is miss diagnosed results in miss of early treatment to cure for heart diseases. With the number of heart disease patients expected to rise soon, it is vital to find a solution. The use of Artificial Intelligence (AI, Conversational AI) and more specifically Machine Learning (ML)[10] algorithms and Deep Learning (Deep Neural Networks, DNN )[9] can mitigate the possibility of human error while increasing prediction accuracy rates for future by using previous data. The expected outcome of the use of these data analysis techniques is a higher accuracy prediction rate of more than or minimum 75%. And, it is expected for the deep neural network to have a higher accuracy rate tahn machine learning model because of its ability to back-propagate and adjust weights, and as theory states. Means, the model, whether it is the machine learning algorithm or deep neural networks, with the best accuracy will be used to create an application that reads required data inputs for patients to determine an accurate heart disease diagnosis. This tool can be a great contribution to the cardiology field as it can be used by medical care professionals to assist them in more accurate diagnoses.

## II. DATA-SET DESCRIPTION:

In this study, we have collected a heart disease dataset known as Cleveland heart disease database from an online machine learning and data mining repository of the University of California, Irvine (UCI). The data-set was gathered by Dr. Robert Detrano and was obtained from V.A. Medical Centre, Long Beach and Cleveland Clinic Foundation. This data-set consists of total raw 76 attributes from which we have considered 13 of them. But, as per previous machine learning researchers we have used 13 of them for greater accuracy. Hence, in this research paper we have considered 13 attribute from overall data-set. For more information about 13

attributes we have described and tabulated all features in Table 1. We have excluded the column of target. In the below table C1, C2, C3,.....,C13 represents the 13 columns containing data of 1025 patients.

**TABLE 1.Features Description of the HF Dataset.**

| Feature Number | Feature column number | Feature Description | Feature abbreviation |
|---|---|---|---|
| 1 | C1 | Age (Years) | age |
| 2 | C2 | Sex | sex |
| 3 | C3 | Chest Pain Type | cp |
| 4 | C4 | Resting Blood Pressure | trestbps |
| 5 | C5 | Serum Cholesterol | chol |
| 6 | C6 | Fasting Blood Sugar | fbs |
| 7 | C7 | Resting Electrocardiographic Results | restecg |
| 8 | C8 | Maximum Heart Rate Achieved | thalach |
| 9 | C9 | Exercise Induced Angina | exang |
| 10 | C10 | Old Peak | oldpeak |
| 11 | C11 | Peak Exercise Slope | slope |
| 12 | C12 | Number of Major Vessels Colored by flourosopy | ca |
| 13 | C13 | Thalium Scan | thal |

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 | 1.0 | 2 | 2 | 3 | 0 |
| **1** | 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 1 | 3.1 | 0 | 0 | 3 | 0 |
| **2** | 70 | 1 | 0 | 145 | 174 | 0 | 1 | 125 | 1 | 2.6 | 0 | 0 | 3 | 0 |
| **3** | 61 | 1 | 0 | 148 | 203 | 0 | 1 | 161 | 0 | 0.0 | 2 | 1 | 3 | 0 |
| **4** | 62 | 0 | 0 | 138 | 294 | 1 | 1 | 106 | 0 | 1.9 | 1 | 3 | 2 | 0 |

## III. Problem :

In India, more than 17 Lakh humans die per annum due to heart diseases and by 2030, the number of humans will die is expected to increase with 2.3 crore deaths per annum. Around 26 percent of total deaths in India occur due to non-communicable diseases (NCD) which are not spread by communication, which are largely referred to the heart diseases and diabetes diseases. High blood pressure usually has no symptoms, so it can't be detected without being measured that is why blood pressure (BP) is one of the most important screenings for human body. Quantity of cholesterol and triglycerides in the body, increase in body weight, blood glucose level. Smoking any kind of drugs and tobacco, physical activity, diet, age, diabetes, chest pain are the main reasons why heart disease is occurring worldwide. By regular observations on the public health data, we came across one thing that increases of patients with some form of cardiovascular disease in countries with low or middle gross national income. A country like India comes under it. Although there are best practices in the area of health care to prevent cardiovascular diseases which are quite life-threatening very often and very serious diseases can be mitigated if we know it at the starting level of diseases. Very powerful clinical practices, medicines available in the market, regular diagnosis, and diet can help to mitigate the diseases. Well defined diagnosis and change in way of living can mitigate some risks of flagging as a heart patient. The motivation to prevent and manage heart disease has spurred the development of numerous Health applications for consumer use, some of which have been scientifically assessed for efficacy. In this paper, we provide deep learning techniques in these systems. And then we solve issues and give solutions associated with heart diseases using deep learning algorithms for medical applications with higher accuracy. A conclusive and early diagnosis of heart disease could be the difference between life and death for some which will be possible due to deep learning techniques. In daily practice, one out of 3 patients is miss diagnosed results in miss of early treatment to cure for heart diseases. With the number of heart disease patients expected to rise shortly, it is vital to find a solution.

## IV. Idea :

Here, we significantly improve on these already very promising ML results by designing and tuning deep neural network (DNN) architectures of increasing depth for detecting heart disease based on routine clinical data. We show that a flexible design of algorithm and the subsequent tuning many of the hyper parameters of a deep neural networks can yield up to 99% accuracy. The results were evaluated and validated using matheves correlation coefficient and confusion metrics measure the quality of the classifications. The model accuracy lies in between for each training 99% , which basically outperforms many of the currently published research in the area oh heart disease predictions. We also show that the hyper parameter optimization technique is that Bayesian optimization for hyper parameter optimization and tuning helps to increase the inference speed and robustness of model.

## V. Idea Details:

***Model architecture***

The steps for Forwarding propagation are given by range from 0 to L, which represents a non-linear hypothesis/prediction function as shown H for X as they give inputs and the fixed weights W,b. After this, we have to modify weights so that the predictions H

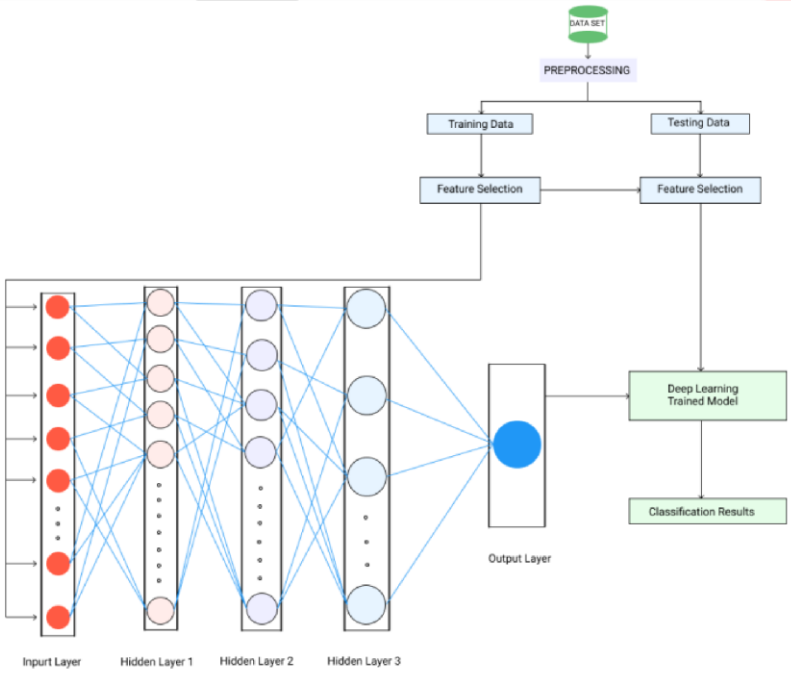
become close to given known outcomes stored in Y. This is known as a binary classification problem where we have 2 classes and is a case of so called supervised learning because we have dependent and independent both the variables. The weights modification for the model is defined as a minimization problem on a cost function or loss function , e.g. Binary cross entropy is also called as sigmoid cross entropy . It is a sigmoid activation plus cross entropy loss. Unlike softmax loss function it is independent ofor each vector of class (component) , means that for each different component class in dataset the loss that is computed by proposed model is not affected by the other classes values and decisions. Because of the characteristic that the values of certain class are not influenced by the decisions of the other class this is beneficial to used for multilable classification. And also it is called as binary classification because it sets up a binary classification condition such that for every class C there is C' = 2 classes as explained above . Loss Function is given as Binary cross entropy-

$$CE = -\sum_{i=1}^{C'=2} t_i log(f(s_i)) = -t_1 log(f(s_1)) - (1 - t_1) log(1 - f(s_1))$$

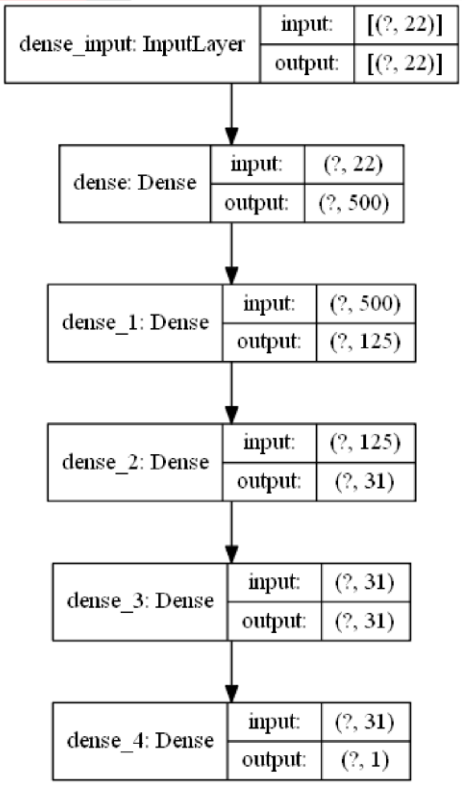
This difficulty made it easier by using a stochastic gradient descent method – which is an iterative algorithm using n training examples at a time. The derivatives of H concerning the weights (W and b) are derived over the layers using the chain rule for differentiating compositions of functions known as chain rule in differential calculus. They are computed then by the back prop means backward propagation steps and used to modify their respective weights and biases, during the iterative training process for each layer where higher-level parameters is a known as hyper parameter referred to as learning rate. The activation functions (possibly different) for each different layer of the same neural network, and represents their derivatives. We have used and coded activation function choices for ReLU, sigmoid, tanh, and leaky ReLU which are well-known activation functions in Deep Learning.

***Weight normalization*:**

Algorithmic Optimization: Regularization is a standard technique that we have to use to prevent overfitting by penalizing large weight values. We can apply regularization to various levels in our model. DNNs tend to assign higher weight values for certain training data points, which corresponds to high variance. Regularization can improve accuracy on test data by helping the model to address the problem of high variance on training data. Regularization is usually done by adding a penalty term of the form to our loss function or cost function J, where ||W,b|| is some norm of the weights, e.g., L1 or L2 [8]. The regularization parameters can impose a penalty on larger weights, thereby ensuring that we do not over fit training data means model tries to memories our training data. Another advantage of regularization is that it can prevent an algorithm from learning from data outliers, which can result in a more robust model. This is helpful for smaller datasets such as heart disease prediction used in our research. Regularization don't remove the outliers but remains them in the dataset but reduces the algorithm's probability of learning from these outlier values. That is why we use the regularization technique to make it default use of an increase in accuracy and by reducing the overfitting and thus automatically decreasing the impact of any outliers in the dataset.




Deep Neural Networks Architecture (5 Layers Neural Network ) For Heart diseases Dataser




***Validation Schemes And Evaluation Matrices*:**

VALIDATION SCHEME

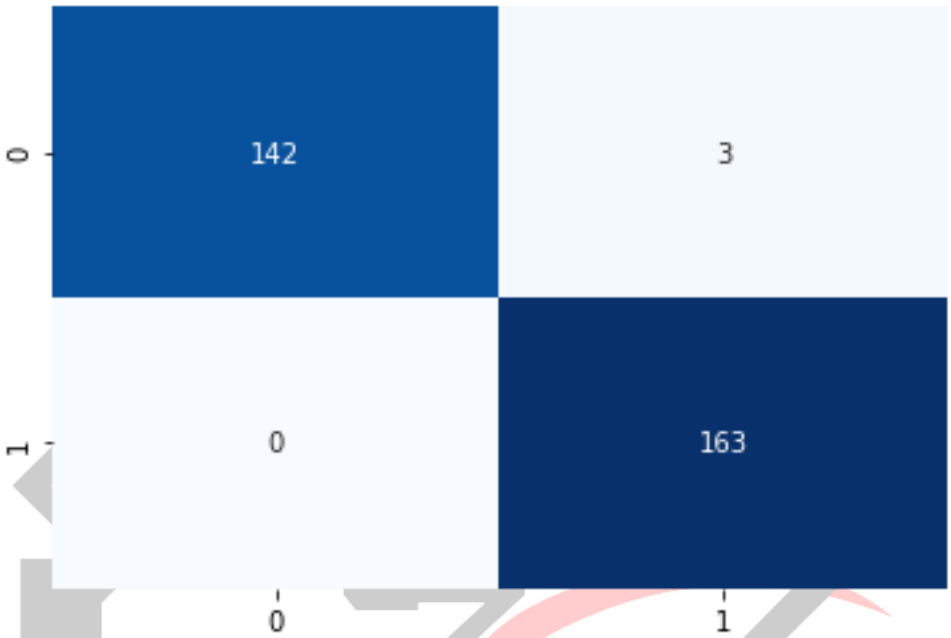
In the previous years, researchers have studied validation schemes for evaluating the performance of the developed diagnostic system architecture. The percentage of train-test-split for data partitioning was different for different studies. Most of these studies like Z. Jin in [1],K. Vembandasamy in [2] and K.R. Lakshmi in [3] have used validation techniques with 70-30 split and 80-20 split. That means 70% to 80% data from dataset is used for training the proposed model on the other hand 20% to 30% data from the dataset is used for testing purpose. We have used 70% of the data from dataset as the training and 30% of the dataset for testing purpose, we have used the approach with the amount of percentage is as same as used in above mentioned papers as train-test-split partitioning.

EVALUATION METRICS

To evaluate the effectiveness and efficiency of the our proposed methodologies and system, different evaluation metrics including accuracy, precision, recall, specificity, f1 score calculated using confusion matrix metrics have been used. Accuracy for a model is defined as the percentage(%) of correctly classified subjects by proposed model.

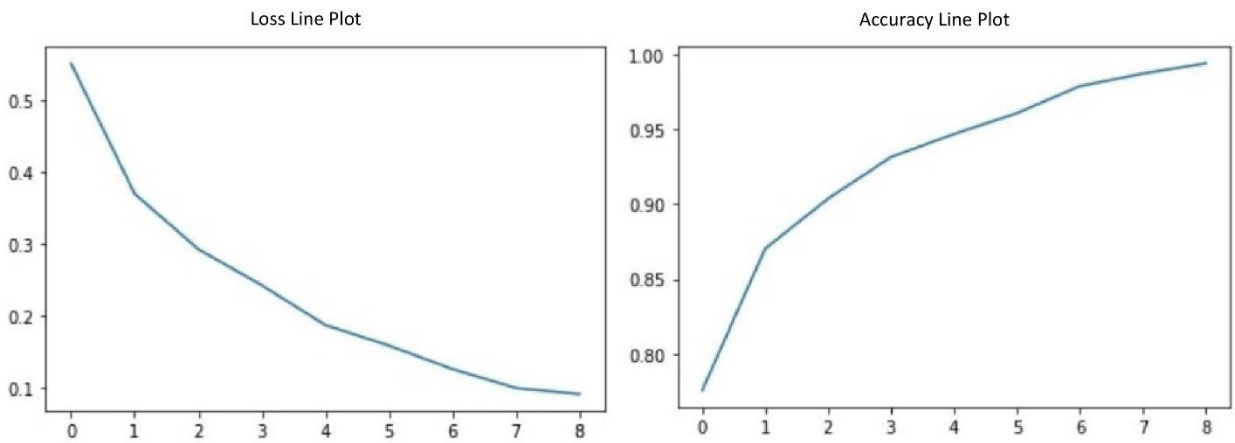
Accuracy = (TP+TN) / (TP+TN+FP+FN)

where- TP denotes number of true positives, TN denotes number of true negatives, FP denotes number of false positives, and FN denotes number of false negatives.




EXPERIMENT RESULTS AND COMPARISONS:

In this experiment, we have used three types of models to get better results with different layers and architectures. In the beginning we have tried 3 layers neural networks architecture which gave us 87% training accuracy and 84% testing accuracy. In the experiment two we have used different types of layers and defined a neural network of 7 layers, which gave us 92% accuracy of training and 86% testing accuracy. And the final approach was a architecture of deep neural networks with 5 layered structure, which gave us the best results of 99.17% training accuracy and 98.57 % testing accuracy.




***Comparisons with other researches*** :

Table 2 : Comaparision Table for Accuracy for Different Researchs

| Study (Year) | Method | Accuracy(%) |
|---|---|---|
| R. Sharmila, 2018 [4] | Support Vector Machine(SVM) in parallel fashion | 85 |
| Aniruddha Dutta, 2019 [5] | CNN model | 83.17 |
| Martin Abadi et al., 2016 [7] | Differential privacy | 98 |
| Our model | Regularized Deep neural networks | 99 |

## VI. Conclusion And Future directions:

We used deep learning methods and algorithms and get way better performance in terms of accuracy. And the use of deep neural networks results in improvements such as robustness of proposed model . This work investigated and showed the potential of using DNN-based data analysis for detecting heart disease based on routine clinical data. DNN data analysis techniques can yield very high accuracy (99% accuracy) results of our model shows that, which significantly outperforms currently published research in the area of machine learning, Future directions include extending this analysis to construct a more thorough model that includes heart visualizations and CT image data. And planning to include User Interface like chat bots(Conversational AI) for taking input data and providing output visualization and predictions to humans.

***References***:

[1] Z. Jin, W. Chaorong, H. Chengguang, W. Feng, "Parameter optimization algorithm of SVM for fault classification in traction converter", The 26th Chinese Control and Decision Conference (2014 CCDC), Changsha, 2014, pp. 3786-3791, doi: 10.1109/CCDC.2014.6852839.

[2] K. Vembandasamy, R. Sasipriya, E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", IJISET-International Journal of Innovative Science, Engineering & Technology, Vol.2, pp.441-444, 2015

[3] K.R. Lakshmi, M.Veera Krishna, and S.Prem Kumar, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability", International Journal of Scientific and Research Publications, Vol.3, Issue 6, pp.1-10, June 2013.

[4] R. Sharmila, S. Chellammal, "A conceptual method to enhance the prediction of heart diseases using the data techniques", International Journal of Computer Science and Engineering, May 2018.

[5] Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton,"An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction",arXiv preprint arXiv:1909.00489, September 2019.

[6] Heart Disease Prediction using Neural Networks, Kaggle: https://www.kaggle.com/bulentsiyah/heart-disease-prediction-using-neural-networks, March 2020.

[7] Martin Abadi et al., "Deep Learning with Differential Privacy", arXiv preprint arXiv:1607.00133, October 2016

[8] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," in IEEE Access, vol. 7, pp. 54007-54014, 2019, doi: 10.1109/ACCESS.2019.2909969.

[9] L. Ali et al., "An Automated Diagnostic System for Heart Disease Prediction Based on ${\chi^{2}}$ Statistical Model and Optimally Configured Deep Neural Network", in IEEE Access, vol. 7, pp. 34938-34945, 2019, doi: 10.1109/ACCESS.2019.2904800.

[10] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.