CARDIO VASCULAR DISEASE PREDICTION USING DEEP LEARNING

Fall 2018
Deep and Reinforcement Learning

Apurva Katti (AGK17C)
Tejas Hasarali (TDH17)

Introduction

Heart attack or the cardio vascular disease is the leading cause of death both in male and female throughout the world. Heart being a vital organ, pumps the pure oxygenated blood and supplies it to other organs through blood vessels. Any small malfunction can severely affect other organs and in the extreme case it is proved to be fatal. Most of the times, the diagnosis is found to result in error because of the unspecialized doctors or misdiagnosis. However, heart attack shows a substantial number of symptoms which could easily be recorded and studied. These symptoms are available from any patient's record and it contains all the details of the patient (like his blood pressure, blood sugar levels, results of several tests etc.,) from the time of his admission into a hospital to the time at which he was discharged from the hospital. By doing so we could establish a direct relationship between the symptoms and the event of a cardiac arrest. This intrinsically will establish the new technique for early prediction or the prevention of cardiac disease. Therefore, in our work we have aimed at establishing a technique which takes in the selected symptoms as input and based on the evaluations, it predicts whether the patient has cardio vascular disease.
It is implemented by constructing a Fully connected neural network model over heart disease dataset from UCI machine learning repository. We implemented the fully connected model by generating training and testing samples and have also included several optimization strategies to improve the accuracy of the model. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems operate such as the brain [1]. It learns by processing the information which are in the form of symptoms in this project. Finally, it helps in the prediction of the disease. It is composed of a large number of highly interconnected processing elements called the neurons working in unison to solve specific problems. The accuracy achieved by using ensemble methods in our work is about 81.11%.

Literature Survey

The existing works have used several other techniques for prediction of heart disease. In [2] MLP architecture is used for the diagnosis. Over 40 attributes were considered, and back propagation algorithm is used. The results are validated by using cross validation, holdout and bootstrapping. However, this work did not consider the UCI dataset. An extension is provided in [3], we considered the momentum parameter, learning rate and conjugate mechanics to improve the back-propagation algorithm. Paper [4] proposed the method of using Supervised Multilayer feed forward networks by updating weights in each epoch. It was analyzed using Weight Linear Analysis technique. This work used only two of the datasets from UCI database. In paper [5] the Artificial Neural Network (ANN) algorithm was used for classifying the Heart Disease based on output. It used the method of Learning Vector Quantization (LVQ) .It is a prototype based Supervised Classification Algorithm. Paper [6] uses heart sound as the

parameter. The abnormality is detected by carefully studying the patterns of the heart sound signals. It was implemented using the support vector method.
However, we have come up with a novel technique for prediction of the cardio vascular disease. Our approach includes considering all four datasets of UCI machine learning repository and building a fully connected neural network model by relevant pre-processing on the data. All the existing works considered a maximum of 14 attributes whereas our work considered 24 parameters. We were also able to generalize the model better by using more data, more parameters and by better generalization techniques which could be seen in the results section. The description of the data, pre-processing, methodology, results are provided in successive sections.

Dataset

We have used the heart disease dataset of the UCI machine learning repository. It consists of the data in the form of patients record from four distinguished hospitals in different countries. They are the Hungarian Institute of Cardiology, Budapest; University Hospital, Zurich, Switzerland; University Hospital, Basel, Switzerland. Medical Center, Long Beach and Cleveland Clinic Foundation. Each dataset contains about 76 attributes which have been recorded by the hospitals about the patients. The Cleveland dataset is widely used for research purposes. We extended the research by considering all four datasets and combing them into one big data. The existing work considered only 14 features out of 76 for building a prediction model. We have selected 24 features of which the 23 features are the important information regarding a patient's health and the last feature represents the goal field or the output which indicates the presence of heart disease in the patient. The description of the 24 selected features is tabulated in table 2. The class distribution is provided in table 1.

Database	No Disease	Disease	Total
Cleveland	164	139	303
Hungarian	188	106	294
Switzerland	8	115	123
Long Beach VA	51	149	200

Table 1. Class Distribution

Keywords:

- *Angina:* a condition of severe pain in the chest usually detected during a cardiac attack episode. It is often found to spread to the shoulders, arms, and neck. It is usually caused by an inadequate blood supply to the heart.
- *Fluoroscopy:* An x-ray procedure which helps to supervise the internal organs in motion. It uses x-ray to produce real-time video images by passing x-rays through the patient and then capturing the statistics by using a device called the image intensifier. This is further

converted to light and the light is then captured by a TV camera and displayed on a video monitor.

- *StressTest:* Thallium stress test, sometimes called a treadmill test or exercise test, helps the doctor to find out the working condition of the heart with workload. The body is made to exercise harder during the test. For the body to cope up with the intense work, it requires more fuel and the heart has to pump more blood. The test can indicate if there's a lack of blood supply through the arteries that go into the heart.
- *Electrocardiogram (ECG):* It is a medical test that detects heart problems by measuring the electrical activity generated by the heart as it contracts. **ST Depression or elevation** measures the interval between ventricular depolarization and repolarization. It represents abnormality due to myocardial infarction.

NAME OF THE ATTRIBUTE	DESCRIPTION	VALUES
Age	Describes the age of the patient	Contains a real number in the range of 28 to 77
Sex	Describes if the patient is male or female	Contains 0 for female and 1 for male
Pain Location	Describes if the patient has substernal pain (chest pain)	Contains 1 for existing pain and 0 otherwise
Resting BP	Describes the value of Resting Blood Pressure. Performed upon admission to hospital	Contains the value in the range of 80 to 200
Hypertension	Describes if the patient has Hypertension or not --Value 0: BP value is <120 mm in Hg --Value 1: BP value is >120 mm in Hg	Contains the value 1 for existing hypertension 0 otherwise
StressTest Duration	Describes the duration of exercise test in minutes	Contains the value in the range of 1 to 250
StressTest ST Time	Describes the time when ST measure depression was noted	Contains the value in the range of 0 to 20 and -1 to missing values
StressTest Max HR	Describes the maximum heart rate achieved	Contains the value in the range of 60 to 202
StressTestResting HR	Describes the resting heart rate	Contains the value in the range of 40 to 139
StressTestMaxFirstBPS	Describes the peak exercise blood pressure (part 1)	Contains the value in the range of 84 to 240
StressTestMaxSecondBPS	Describes the peak exercise blood pressure (part 2)	Contains the value in the range of 40 to 134
StressTestResting BP	Describes the stress test resting Blood Pressure	Contains the value in the range of 50 to 120
ExerciseAngina	Describes if the angina was induced due to the exercise	Contains the value 1 if it exists otherwise 0
STDepressionExercise	Describes the ST depression induced by exercise relative to rest	Contains the value in the range of 1 to 6.2

ChestPainType	Describes the type of chest pain. There are 4 types. -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic	Contains the value in the range of 1 to 4
Cholesterol	Describes serum cholesterol in mg/dl --Value 0: cholesterol value<=200 (normal) --Value 1: cholesterol value <=239(borderline) --Value 2: cholesterol value >240(high cholesterol levels)	Contains the value 0, 1, 2
Smoker	Describes if the person is smoker or not	1 indicates smoker 0 otherwise
BloodSugar	Describes the fasting blood sugar levels in mm --Value 0: value<120 mg/dl --Value 1:value>120 mg/dl	Contains 0 ,1
HistoryofHA	Describes if the patient has the history of coronary heart condition	Contains 1 if existing 0 otherwise
RestingECG	Describes the resting electrocardiographic results --Value 0: normal --Value 1: abnormal (either having ST wave abnormality or showing probable or definite left ventricular hypertrophy	Contains 0,1
STDepressionSlope	Describes the slope of the peak exercise ST segment --Value 0: abnormal(upsloping) --Value 1: normal(flat) --Value 2: abnormal(downsloping) --Value -1: Missing values	Contains values -1,0,1,2
ColoredVesselsFluroscopy	Describes the number of major blood vessels colored --Value (0-3): Number of blood vessels colored --Value -1: Missing values	Contains -1,0,1,2,3
HeartWall Damage	Describes the defect observed during the stress test --Value 0 (thal<=3): normal --Value 1 (thal=7): reversible defect (when the heart muscle compromised can be reversed (medication, surgery) --Value 2(thal<=6): fixed defect (when the heart muscle compromised cannot be reversed) --Value -1: Missing values	Contains -1,0,1,2
Output	Describes if the person has cardio vascular heart disease or not (0-4) --Value 0: value 0 --Value 1: values (1-4)	Contains 1 if exists 0 otherwise

Table 2: Provides the description of the attributes considered.

Reasons for Considering the Attributes:

Each attribute describes a direct connection to the output, yet it is impossible to use all the 76 attributes. It is usually due to the lack of data (missing data) or due to some of the parameters being trivial or irrelevant to deep learning paradigm. The 14 parameters of age, sex, Chest Pain Type, Resting BP, Blood Sugar, Cholesterol, Resting ECG, Stress Test Max HR, Exercise Angina, ST Depression Exercise, ST Depression Slope, Colored Vessels Fluoroscopy, Heart Wall Damage, Output are commonly used in the research experiments since these fields contain relevant data and also describes the direct relationship with the output. We have considered the other attributes which also describes the relationship with the output.

- *Pain Location:* Cardiac arrest is most prominently diagnosed with the help of pain location. We have tried to establish the relationship with 2 parameters. The first is the patients who have chest pain and the patients who have diabetes because the heart attack can happen either silently or with chest pain.
- *Hypertension:* This parameter simple represent if the person has high blood pressure or not. Just the values of blood pressure seem trivial as it is important to establish a metric to define the normality and abnormality.
- *StressTestDuration:* This parameter represents the duration of the stress test. It is important because the healthy patients would easily be able to complete the entire test where as people with cardiac infarction would not be able to complete the duration due to decreased breathing or physical fatigue caused by myocardial infarction.
- *StressTest ST Time:* This parameter represents the time at which the ST depression was measured. It could also indicate the depth of the problem. If the value measure is high, we can infer that the percentage of blockage is high (indicating heart attack) and the smaller value can indicate lesser value which can be treated with medication or minor procedures**.**
- *StressTestResting HR:* This parameter is usually compared along with the maximum heart rate reached. If the maximum heart rate value is high, we can infer that the patient can be affected with heart attack because the heart takes more cycles to pump the blood.
- *StressTestMaxFirstBPS:*
- *StressTestMaxSecondBPS:* These three parameters are usually used together to
- *StressTestResting BP:* compare. Just like the heart rate, the blood pressure variation is keenly observed to check for abnormality caused due to heart attacks.
- *Smoker:* One of the leading causes of cardiac infarction is caused due to smoking.
- *HistoryofHA:* History of cardiac disease in the family is an important criterion since the cardiac disease can also be inherited which indicates the inherited genetic risk factors. Hence, we have considered this attribute in our study.

Design

Data Processing:

- *Reading:* The heart disease raw data of patients from four hospitals was obtained from UCI database. The obtained data was reformatted into CSV to read it as data frame. The dataset from four hospitals was combined into one dataset and it was shuffled to improve the generalization.
- *Filling Missing Values:* Some of missing values was filled by finding the relationship between different features like:
    - A person was assumed as smoker if the data is present for no of cigarettes smoked in a day or the number of years since he has been smoking.
    - The pain location was assumed to be substernal if the patient has a heart disease and the pain location was assumed to be not substernal when the patient is diabetic even if the patient has a heart disease.
    - The patient with resting blood pressure more than 120 was assumed to have hypertension condition.
    - The exercise angina was assumed to be present if the patient has pain in substernal area.

  The remaining missing values of continues data was filled with mean and categorical data with -1.
- *Pre-processing of Data:* Some of the features was preprocessed to make it suitable for training like:
    - The continues cholesterol data was converted into categorical by assigning 0 if the patient has less than or equal to 200 mg/dL (desirable), 1 if it is between 200 mg/dL and 239mg/dL (borderline high) and 2 if it is above 239mg/dL (high)
    - The resting ECG was simplified from 0, 1, 2 (normal, ST abnormality, left ventricular hypertrophy) to 0, 1 (normal, abnormal).
    - The slope produced by ST depression was changed 1, 2, 3 representing 1 as high slope, 2 has flat or normal and 3 as down slope to 1, 0, 2 respectively.
    - The heart wall damage values 1 to 7 was converted into 0, 1, 2. The values less than 3 being no damages was converted into 0, the values between 3 and 6 being irreparable damage to 2 and the value 7 being reparable damage to 1.
    - The output of some of the hospitals had values from 0 to 4 representing the severity of the disease, which was simplified into 0, 1 (no disease, disease) to make it consistent.

  All the categorical values except the output was converted into one hot encoder.
- *Preparing data for training:* The preprocessed data was normalized into values between 0 and 1 using Min Max Scaler to bring all the data into same scale. Then the data was split into training and testing data, 20% of the was used for testing the accuracy of the model. Of the 80% data selected for training, 10% was used for validation of the model.

Training of the Model:

- *Building the model:* Sequential model using fully connected layers was used for prediction of heart disease. The fully connected layer consists of fully connected

neurons, meaning each neuron is dependent on all the output of the previous layer. The choice of fully connected neural network was obvious as the patient data is independent of one another and it is not an image dataset. We trained five models with different weight initialization, number of layers, activation function and optimizer as shown below in the tables. All the models were trained for 1000 epochs with batch size of 16.

Model - 1:

Layer	Activation	Neurons	Weight Initializer	Bias Initializer	Regularization
Dense	Relu	46	He Normal	Zero	None
Dropout - 0.2		46			
Dense	Relu	23	He Normal	Zero	L1 0.01
Dropout - 0.2		23			
Dense	Relu	11	He Normal	Zero	None
Dropout - 0.2					
Dense	Sigmoid	1	He Normal	Zero	None

Loss Function – Mean Squared Error
Optimizer – Adam
Learning rate – 0.0001

Model - 2:

Layer	Activation	Neurons	Weight Initializer	Bias Initializer	Regularization
Dense	Relu	46	He Normal	Zero	None
Dropout - 0.2		46			
Dense	Relu	11	He Normal	Zero	L1 0.01
Dropout - 0.2		11			
Dense	Sigmoid	1	He Normal	Zero	None

Loss Function – Mean Squared Error
Optimizer – Stochastic Gradient Descent

Learning rate – 0.005

Model - 3:

Layer	Activation	Neurons	Weight Initializer	Bias Initializer	Regularization
Dense	Relu	46	He Normal	Zero	None
Dropout - 0.2		46			
Dense	Relu	23	He Normal	Zero	L1 0.01
Dropout - 0.4		23			
Dense	Sigmoid	1	He Normal	Zero	None

Loss Function – Mean Squared Error
Optimizer – Adam
Learning rate – 0.0001

Model - 4:

Layer	Activation	Neurons	Weight Initializer	Bias Initializer	Regularization
Dense	Relu	46	Glorot Uniform	Zero	None
Dropout - 0.2		46			
Dense	tanh	23	Glorot Uniform	Zero	L1 0.01
Dropout - 0.4		23			
Dense	Sigmoid	1	Glorot Uniform	Zero	None

Loss Function – Mean Squared Error
Optimizer – Adam
Learning rate – 0.0001

Model - 5:

Layer	Activation	Neurons	Weight Initializer	Bias Initializer	Regularization
Dense	tanh	46	Glorot Uniform	Zero	None
Dropout - 0.2		46			
Dense	relu	23	Glorot Uniform	Zero	L1 0.01
Dropout - 0.4		23			
Dense	Sigmoid	1	Glorot Uniform	Zero	None

Loss Function – Mean Squared Error
Optimizer – RMSprop
Learning rate – 0.0002

- *Tuning of the model:* Initially due to small amount of data the model was overfitting, so we used generalization techniques to improve the model. We used L1 regularizers and dropout layers to increase the generalization of the model. The L1 regularizers or lasso regularizers works by adding a squared magnitude penalty term to the loss function, which helps to avoid the overfitting of the model and it is resistant to outliers. The dropout layers help in improving the generalization by dropping a fixed percentage of neurons in every epoch, which creates an effect of training models with different architecture. The number of neurons was also reduced to reduce the capacity of the model, which in turn increases the generalization of the model. Even after extensive parameter tuning the maximum accuracy achieved was around 75%.
- *Ensemble Method:* As the accuracy of the trained model was less, we used ensemble method to improve the accuracy. Here we trained 5 different models with different parameters as shown in the tables above and we performed voting on the predicted output. The prediction from all the five trained models on the same 20% testing data was obtained, then the prediction was compared with one another to find the output which appeared maximum number of times. The output which appeared maximum number of times was chosen as the correct output. The accuracy of the model improved from around 75% to 81.11%. The 5 different models built not only has different parameters but may also has different data in different order, this is due to shuffling of the data and random test, training data split before training the model.

Results

The accuracy of all the five models along with the ensemble method could be seen in the table below. The maximum accuracy obtained by a single model was 76.67%, which was improved to 81.11% using ensemble method. The accuracy of the model could not be improved more than 81.11% due to large number of missing values which creates a randomness effect on the trained model.

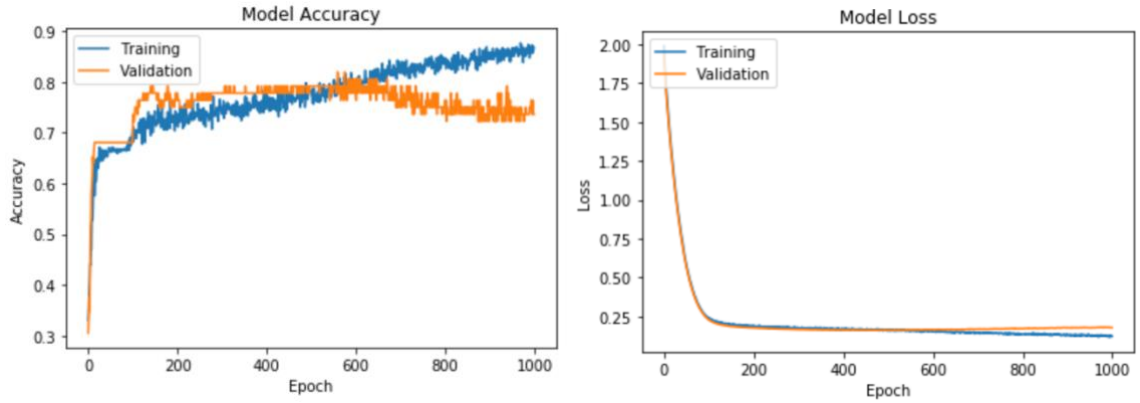
	Accuracy (%)
Model 1	76.11
Model 2	73.33
Model 3	76.67
Model 4	73.33
Model 5	75.00
Ensemble Method	81.11

From the confusion matrix shown below we can see that model was better at prediction when the patient had disease rather than when the patient did not. This could be due to less number patients with the heart disease in the used dataset.

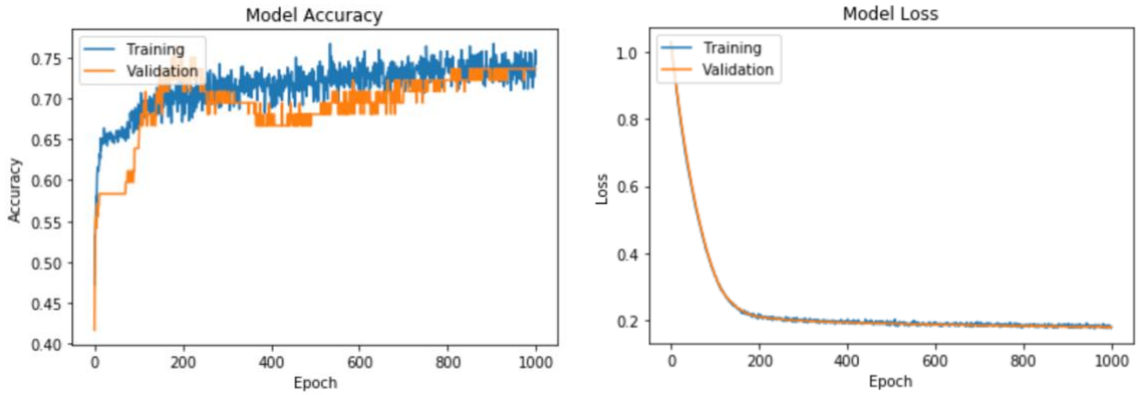
	Pred: yes	Pred: no
True:yes	96	19
True: no	14	51

Confusion matrix

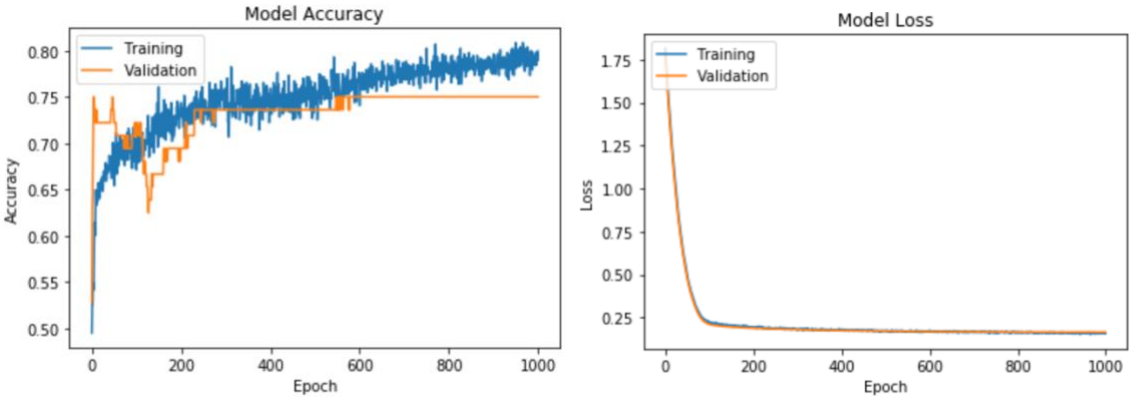
The graphs below show the accuracy and loss of all the models, it could be seen that even though the model loss remains stable throughout the training period the accuracy of the model fluctuate a lot. This is could due to large number of missing values in the data, which are filled with mean values and -1. This creates randomness in the data, which can never be learnt. But the variation in the model loss is minimal due to good generalization of the model.



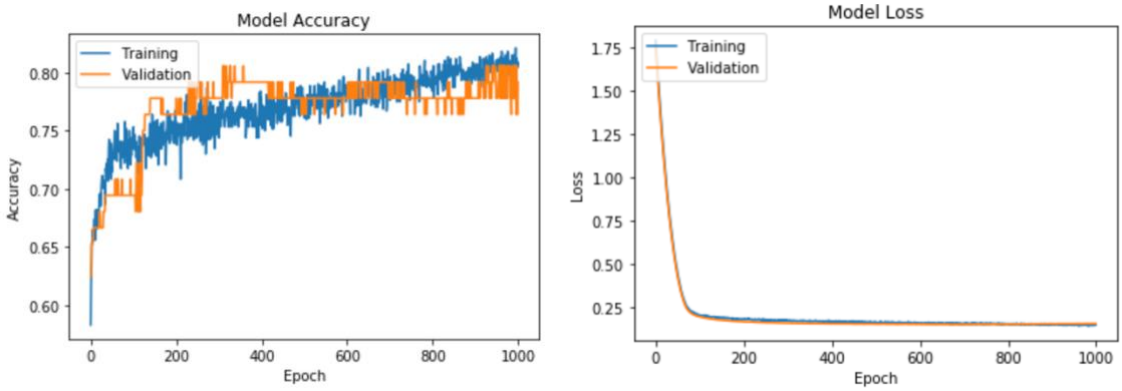
Model 1



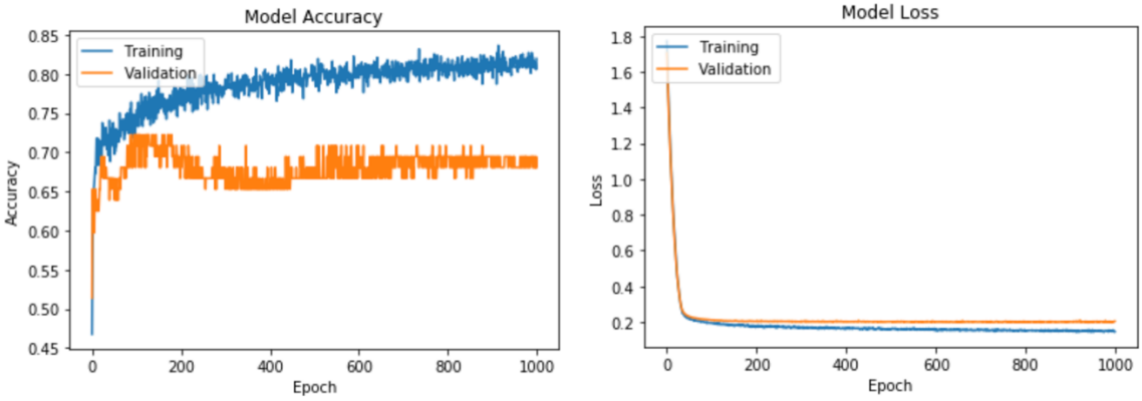
Model 2



Model 3



Model 4



Model 5

Conclusion

This work proposed a novel technique of building a fully connected neural network for cardio vascular disease prediction using the UCI datasets. We extensively studied the relationship between different parameters (attributes of a patient's data) and are successful in prediction of cardiovascular disease. All other existing works considered just the 14 significant features of UCI dataset and our work considered 24 features which is an improvement. A lot of

generalization and regularization strategies were used to improve the accuracy. We also used the ensemble method of bagging by building 5 similar models and thereby decided the overall accuracy based on the voting scheme. The experiments displayed good results performance results. A novel approach by using the fully connected layered neural network proved to be successful in the prediction step with an overall accuracy of 81%.

References

[1]. https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html
[2]. Hongmei yan, Yingtao Jiang, Jun Zheng, Chenglin Peng, Qinghui Li,". A Multilayer Perceptron-Based Decision Support System for heart Disease Diagnosis", 30(2006) 272-281
[3]. Miss. Chaitrali S. Dangarw, Dr. Mrs. Sulabha S. Apte". A Data Mining Approach For Prediction of Heart Disease Using Neural Networks" volume 3, Issue 3, October December (2012), pp.30-40.
[4]. YAN Hongmeil, PENG Chenglin1, DING Xiaojun2, XIAO Shouzhong, "Improving the accuracy of heart disease diagnosis with an augment back propagation algorithm", June 2003.
[5]. Durairaj M, Sivagowry S and Persia A," An Empircial Study on Applying Data Mining Techniques for the Analysis and Prediction of Heart Disease",2013.
[6]. Durairaj M, Revathi V," Soft Computing Methodology to Measure Heart Sound-A Survey",2015.