Heart Disease Prediction Using ANN

1st Monica Gattupalli
Masters in computer science
Blekinge institute of technology
Karlskrona, Sweden
moga20@student.bth.se

Abstract - One of the most common diseases which is seen in women and men is a heart disease which has cases throughout the world. The most common type of heart disease is heart arrhythmias, Cardiomyopathy, valvular heart disease and heart failure. Machine learning techniques are performing a vital role in detecting heart disease by analyzing the features like chest pain, resting blood pressure, maximum heart rate, number of major vessels that can predict whether the person has heart disease or not. The designed method in this paper is to build three different models where one model is trained and tested on the whole dataset and remaining two models are trained and tested on male and female patient's data and the accuracy of the three models are compared.

Keywords: Artificial Neutral Network, heart disease, machine learning, Z-scores.

I.Introduction

In recent years, due to the lifestyle of an individual people were affected by different types of heart diseases. Coronary Artery disease is the most common type of infection which ultimately leads to heart attacks. In the year 2012, 7.4 million people died due to coronary artery disease [1]. In recent years technology reached its height where all the types of heart diseases can be detected by advanced medical

techniques (Developing through Artificial neural network). These techniques take the details like heart rate, blood pressure, sugar level etc., and predict the type of heart disease for the patient.

Machine learning algorithms are playing an important role in the field of medicine, especially Neural networks stood in the first place in detecting the types of disease and stage of disease as they are capable of dealing with images, they even can detect the disease more accurately[2]. These advanced algorithms can also deal with time-series data. More the data more accurate the prediction is experienced in the Deep learning algorithms.

Artificial Neural Network is one of the most commonly used algorithms for diseases predictions. These algorithms are trained upon a large amount of data where the data is passed to the levels of architecture and trained upon the different number of units with an activation function. The major advantage of the ANN is easy to train and can predict more accurately with large inputs [3].

In this paper, three different models are built upon an Artificial neural network and trained upon the selected dataset. The accuracies of the three models are compared after performing the data pre-processing on the datasets. The Artificial Neutral Network achieved an accuracy of 86% on the whole dataset. 73% and 82% on male and female patient's data respectively

II.Method

a) Dataset:

The dataset used for the experiment is taken from Kaggle that is "heart.csv"[4]. The dataset contains 303 rows with 14 attributes. The "target" attribute contains two binary values which indicates the binary classification.

Attribute information

1. Age: Age of the person
2. Sex: Gender of the person (1= male, 0=female)
3. Chest pain type(cp): level of chest pain (4 values (0,1,2,3))
4. Resting blood pressure(trestbps): numerical values
5. Serum cholesterol(chol): numerical values in mg/dl
6. Fasting blood sugar(fbs): fbs rate ,120 mg/dl(1=True,0=False)
7. Resting electrocardiographic result(restecg): values are 1 or 0.
8. Maximum heart rate received(thalach): numerical values.
9. Exercise includes angina(exang): only 2 values (1 =yes,0=no)
10. Oldpeak: ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment(slope): Values various between (0,1,2)
12. Number of major vessels(ca): colored y floursopy (0-3)
13. Thal: values various between (3,6,7; 3= normal, 6= fixed defect, 7=reversable defect)
14. Target: values are 1 or 0.

The below Figure-1 gives the first 5 rows in the dataset by using "head ()" command.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63.0	1.0	3.0	145.0	233.0	1.0	0.0	150.0	0.0	2.3	0.0	0.0	1.0	1.0
1	37.0	1.0	2.0	130.0	250.0	0.0	1.0	187.0	0.0	3.5	0.0	0.0	2.0	1.0
2	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
3	56.0	1.0	1.0	120.0	236.0	0.0	1.0	178.0	0.0	0.8	2.0	0.0	2.0	1.0
4	57.0	0.0	0.0	120.0	354.0	0.0	1.0	163.0	1.0	0.6	2.0	0.0	2.0	1.0

Figure-1: First 5 rows of dataset.

a) Data Selection

Using the panda library, the dataset "heart.csv" is read and some of the preprocessing techniques are applied to the dataset. Using "train_test _split" method the entire dataset is divided into 70:30 ratio and for training another 2 model's whole dataset is divided into 2 sub-datasets where female and male patient's data are separated into 2 datasets.

b) Data Visualization

Histogram plot is used to observe the distribution of all 14 attributes. The correlation map or heat map is generated for all the 3 datasets to check the dependencies of each attribute and which is depicted on both x-axis and y-axis.

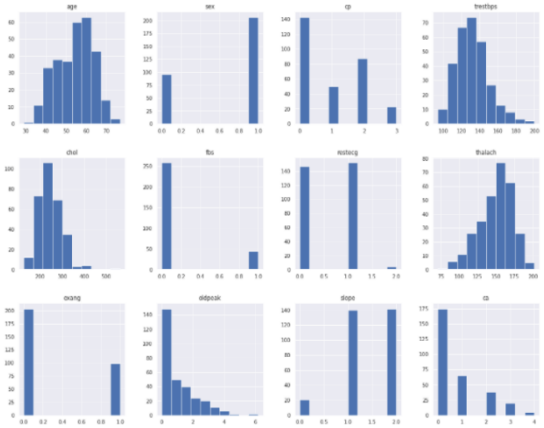

Figure-2: Histogram plot

c) Data Pre-processing

Data Pre-processing contains 2 steps they are 1. Data cleaning 2. Data transformation, performing these techniques helps to increase the performance and accuracy of the model by polishing the dataset.

1. Data cleaning: Data Cleaning steps include

Removing or replacing the missing or null values: The selected dataset doesn't have any missing or null values in the dataset.

Identification and removing the outliers: Z-scores[5] method is used for identifying and removing the outliers. In total 16 outliers are identified in the dataset when the threshold value is set to 3 and those outliers are removed from the dataset then the size of the dataset is (287,14). The same technique is applied for the remaining 2 datasets.

2. Data Transformation:

Standard scaler techniques are used to standardize the features by removing the mean and variance [6]. This technique is used and all the 3 datasets are transformed and standardized. These datasets are divided and sent to the model.

d) Proposed ANN Model:

The proposed ANN has 4 neural network layers, the first 3 layers are followed by the dropout layer. Rectified Linear Unit (ReLU) activation function is used in all 4 layers. The number of units of the layers is 20,25,30, 35 respectively. And the last layer is called a final layer which has the sigmoid activation function with 1 unit. The same layers are designed for all the 3 Models and they are trained and tested against the 3 datasets respectively.
At the stage of compilation “adam” optimizer is used, “binary_crossentropy” is used as loss function and “accuracy” is the measuring metrics. This is followed by all the 3 models. All the designed 3 Models are executed to 90 epochs with a batch size of 32.
The below figure explains the summary of the designed models

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 layer1 (Dense)              (None, 20)                280

 dropout (Dropout)           (None, 20)                0

 layer2 (Dense)              (None, 25)                525

 dropout_1 (Dropout)         (None, 25)                0

 layer3 (Dense)              (None, 30)                780

 dropout_2 (Dropout)         (None, 30)                0

 layer4 (Dense)              (None, 10)                310

 f-layer (Dense)             (None, 1)                 11

=================================================================
Total params: 1,906
Trainable params: 1,906
Non-trainable params: 0
_________________________________________________________________
```

Figure-3: Model Summary

e) Performance

Accuracy: The accuracies of the training and testing data is calculated for all the 3 models and accuracies are compared the highest accuracy model among the testing data is considered as the best model.

III.Result and Analysis

Model-1 is trained and tested upon whole dataset with the proposed ANN architecture -1 the accuracy of the training and testing data is 88% an 86% respectively.

```
7/7 - 0s - loss: 0.3009 - accuracy: 0.8800 - 18ms/epoch - 3ms/step
Epoch 86/90
7/7 - 0s - loss: 0.2799 - accuracy: 0.8700 - 15ms/epoch - 2ms/step
Epoch 87/90
7/7 - 0s - loss: 0.2849 - accuracy: 0.8750 - 18ms/epoch - 3ms/step
Epoch 88/90
7/7 - 0s - loss: 0.3254 - accuracy: 0.8750 - 19ms/epoch - 3ms/step
Epoch 89/90
7/7 - 0s - loss: 0.3273 - accuracy: 0.9000 - 21ms/epoch - 3ms/step
Epoch 90/90
7/7 - 0s - loss: 0.2882 - accuracy: 0.8850 - 19ms/epoch - 3ms/step
```

Figure-4: Accuracy of the training data

```
3/3 [==============================] - 0s 6ms/step - loss: 0.4048 - accuracy: 0.8621
Accuracy: 86.21%
```

Figure-5: Accuracy of the testing data

Model-2 is trained and tested upon female patients in the dataset with the proposed ANN architecture and the accuracy of the training and testing data is 92% and 82% respectively.

```
3/3 - 0s - loss: 0.1628 - accuracy: 0.9104 - 9ms/epoch - 3ms/step
Epoch 86/90
3/3 - 0s - loss: 0.1934 - accuracy: 0.8955 - 13ms/epoch - 4ms/step
Epoch 87/90
3/3 - 0s - loss: 0.1745 - accuracy: 0.9254 - 21ms/epoch - 7ms/step
Epoch 88/90
3/3 - 0s - loss: 0.2405 - accuracy: 0.8806 - 16ms/epoch - 5ms/step
Epoch 89/90
3/3 - 0s - loss: 0.1446 - accuracy: 0.9254 - 17ms/epoch - 6ms/step
Epoch 90/90
```

Figure-6: Accuracy of the training data

```
1/1 [==============================] - 0s 176ms/step - loss: 0.7893 - accuracy: 0.8276
Accuracy: 82.76%
```

Figure-7: Accuracy of the testing data

Model-3 is trained and tested upon male patients in the dataset with the proposed ANN architecture and the accuracy of the training and testing data is 86% and 73% respectively.

```
6/6 - 0s - loss: 0.5220 - accuracy: 0.6970 - 18ms/epoch - 3ms/step
Epoch 24/90
6/6 - 0s - loss: 0.5026 - accuracy: 0.7758 - 12ms/epoch - 2ms/step
Epoch 25/90
6/6 - 0s - loss: 0.5121 - accuracy: 0.7515 - 18ms/epoch - 3ms/step
Epoch 26/90
6/6 - 0s - loss: 0.5148 - accuracy: 0.7212 - 20ms/epoch - 3ms/step
Epoch 27/90
6/6 - 0s - loss: 0.4960 - accuracy: 0.7212 - 21ms/epoch - 3ms/step
Epoch 28/90
6/6 - 0s - loss: 0.4823 - accuracy: 0.7697 - 19ms/epoch - 3ms/step
Epoch 29/90
6/6 - 0s - loss: 0.4941 - accuracy: 0.7758 - 28ms/epoch - 5ms/step
```

Figure-8: Accuracy of training data

```
2/2 [==============================] - 0s 7ms/step - loss: 0.7892 - accuracy: 0.7302
Accuracy: 73.02%
```

Figure-9: Accuracy of testing data

IV.Conclusion

On comparing the obtained accuracies of the 3 designed models, model which is trained and tested with whole dataset performed well with the highest accuracy of 86 % over remaining 2 models.

References

[1] S. Safdar, S. Zafar, N. Zafar, and N. F. Khan, "Machine learning based decision support systems (DSS) for heart disease diagnosis: a review," *Artif. Intell. Rev.*, vol. 50, no. 4, pp. 597–623, Dec. 2018, doi: 10.1007/s10462-017-9552-8.

[2] S. I. Ayon, Md. M. Islam, and Md. R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," *IETE J. Res.*, vol. 0, no. 0, pp. 1–20, Jan. 2020, doi: 10.1080/03772063.2020.1713916.

[3] "State-of-the-art in artificial neural network applications: A survey | Elsevier Enhanced Reader." https://reader.elsevier.com/reader/sd/pii/S2405844018332067?token=4E91433035BAE6D62FFD03FE0FCE73FEDA095EBD112B2DD1F95849FA25CBB659D6EB7CB82CDED04739ECECD31BDF10C1&originRegion=eu-west-1&originCreation=20220108113145 (accessed Jan. 08, 2022).

[4] "Heart Disease UCI." https://kaggle.com/ronitf/heart-disease-uci (accessed Jan. 09, 2022).

[5] J. J. Carey and M. F. Delaney, "T-Scores and Z-Scores," *Clin. Rev. Bone Miner. Metab.*, vol. 8, no. 3, pp. 113–121, Sep. 2010, doi: 10.1007/s12018-009-9064-4.

[6] "sklearn.preprocessing.StandardScaler," *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (accessed Jan. 08, 2022).