Heart Disease Prediction Using Neural Network

SHRADDHA PATEL
Computer Science & Engineering, ITM (SLS) Baroda University, Paldi, Vadodara, India

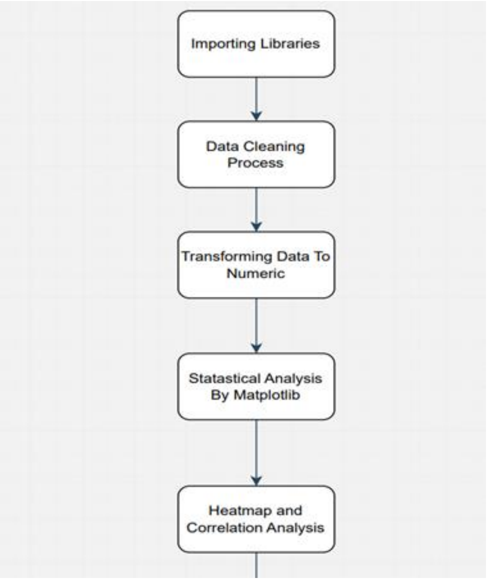
Abstract- Heart disease is a critical health concern worldwide, contributing significantly to morbidity and mortality. Timely prediction and accurate identification of individuals at risk are essential for effective preventive measures and personalized healthcare. In this study, we propose a novel approach for heart disease prediction utilizing a neural network-based model. The neural network is designed to analyze a comprehensive set of input features, including attributes such as blood pressure, cholesterol levels, heart rate, and other characteristic attributes, patients will be categorized based on the different stages of coronary artery disease. The neural network employs advanced deep learning techniques, leveraging its ability to automatically learn intricate patterns and representations from complex datasets. To assess the model's efficacy, extensive experiments are conducted, employing various performance metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve. In this study, we will be using common Python libraries, such as pandas, matplotlib, sklearn and keras for visualization and implementing deep learning algorithm and also softmax classification function.

Indexed Terms- Neural Network, Deep Learning Techniques, Sklearn, Softmax Classification Function

I. INTRODUCTION

Heart disease continue to be a major global health challenge, necessitating innovative approaches for timely detection and prevention. In recent years, the application of neural networks, a subset of deep learning techniques, has gained prominence in the domain of heart disease prediction. This shift towards neural networks is driven by the limitations inherent in traditional methods, which often struggle to capture the intricate patterns and nonlinear relationships present in complex patience dataset. Additionally, neural networks have demonstrated superior performance in handling unstructured data, such as medical images and signals, providing a holistic view of an individual's health status. This capability is crucial in capturing nuanced patterns that may elude traditional methods, offering a more comprehensive understanding of the complex factors contributing to heart disease.

This research aims to explore the efficacy of neural networks in heart disease prediction, addressing the limitations of conventional methods and emphasizing the need for a more nuanced and data-driven approach. Through rigorous experimentation and validation, we seek to demonstrate the superiority of neural networks in capturing the subtleties of cardiovascular risk, ultimately contributing to more accurate predictions and improved patient outcomes. In an era where precision medicine is gaining prominence, the integration of neural networks stands as a pivotal step towards enhancing our ability to identify individuals at risk of heart disease and tailor interventions accordingly.


Importing Libraries
Data Cleaning Process
Transforming Data To Numeric
Statastical Analysis By Matplotlib
Heatmap and Correlation Analysis

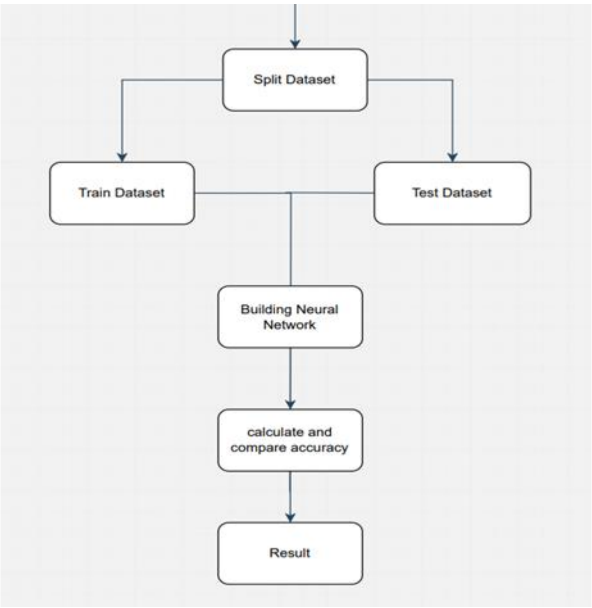


Fig.1: Flowchart of research work

II. LITERATURE SURVEY

This work has been prompted by a significant amount of work linked to the diagnosis of heart disease using machine learning techniques. There is a brief review of the literature in this publication. A heart disease forecast that is accurate has been generated utilizing a variety of algorithms, some of which are logistic

KNN, Random Forest Classifier, Regression, and so forth. The results show that every algorithm has a different capacity to record the specified goals.

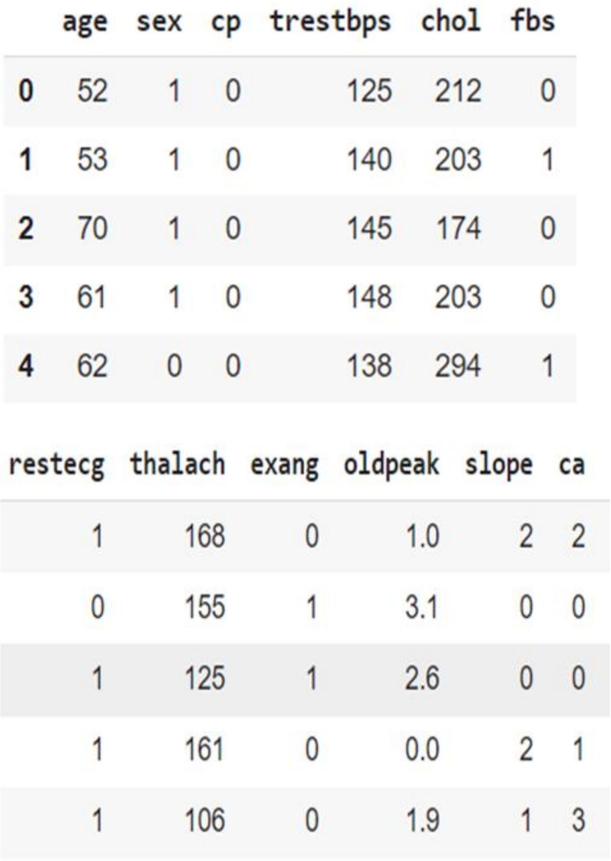
The goal of the research presented by Narain et al. is to improve the accuracy of the commonly used Framingham risk score (FRS) by developing a novel machine-learning-based cardiovascular disease (CVD) prediction system. This proposed system employs a quantum neural network to learn and recognise patterns of CVD. It was experimentally validated and compared with the Framingham Research System (FRS) using data from 689 individuals with symptoms of CVD and a validation dataset from the Framingham research. The accuracy of the proposed approach was shown to be 98.57% in predicting the risk of CVD, far higher than the accuracy of the FRS (19.22%) and other methods already in use. The study's conclusions indicate that the recommended strategy may help physicians predict their patients' risk for CVD, develop more effective treatment regimens, and enable early diagnosis.

Using machine learning (ML) approaches, Drod et al. (2022) sought to determine the most important risk factors for cardiovascular disease (CVD) in individuals with metabolic-associated fatty liver disease (MAFLD). On 191 MAFLD patients, blood biochemical investigation and subclinical atherosclerosis assessment were carried out. ML techniques, including as principle component analysis (PCA), univariate feature ranking, and multiple logistic regression classifier, were used to build a model to identify those with the highest risk of CVD. The most important clinical features, per the study, were duration of diabetes, plaque scores, and hypercholesterolemia. With an AUC of 0.87, the machine learning method successfully identified 40/47 (85.11%) high-risk patients and 114/144 (79.17%) low-risk patients. Based on basic patient criteria, an ML technique is helpful in identifying MAFLD patients with extensive CVD, according to the study's findings.

The authors of a study by Shah et al. (2020) [18] set out to create a machine learning-based model for cardiovascular disease prediction. The UCI machine learning repository provided the 303 instances and 17 attributes that made up the Cleveland heart disease dataset, which provided the data used for this study. The authors used a range of supervised classification techniques, such as random forest, k-nearest neighbour (KKN), decision trees, and naive Bayes. With an accuracy rate of 90.8%, the study's findings showed that the KKN model had the highest level of performance. The work underscores the potential value of machine learning methods in the prediction of cardiovascular illness and stresses the significance of model and technique selection for best outcomes.

III. DATASET

Heart disease prediction is necessary in this growing world. The dataset is downloaded from Kaggle, which was used in the building model and had many different fields in it, from which 14 fields were used which are age, sex, chest-pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, max heart rate achieved, exercise included angina, ST depression included by exercise relative to rest, peak exercise ST segment, Number of major vessels, Thal, Diagnosis of heart disease. The following data is stored in "CSV" format. In the system, we have trained the model using different features where 80% of the data is used for training while the rest 20% is for testing purposes. Pandas, numpy, and matplotlib are the libraries used in the model, where pandas are used in analyzing and manipulating, numpy is used for working with arrays, and matplotlib is for interactive visualization. Last, we check accuracy of both model categorical model and binary model.



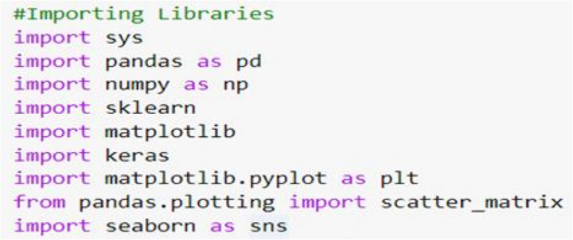
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
0	52	1	0	125	212	0	1	168	0	1.0	2	2
1	53	1	0	140	203	1	0	155	1	3.1	0	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1
4	62	0	0	138	294	1	1	106	0	1.9	1	3

Fig.2: Dataset

IV. METHODOLOGY

1) Importing libraries

The first crucial step is to import all the python related libraries. By using this libraries we can use the functions to model and build neural network. After building neural network we can predict heart disease also finds the best accuracy result



```
#Importing Libraries
import sys
import pandas as pd
import numpy as np
import sklearn
import matplotlib
import keras
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import seaborn as sns
```

Fig.3: Libraries

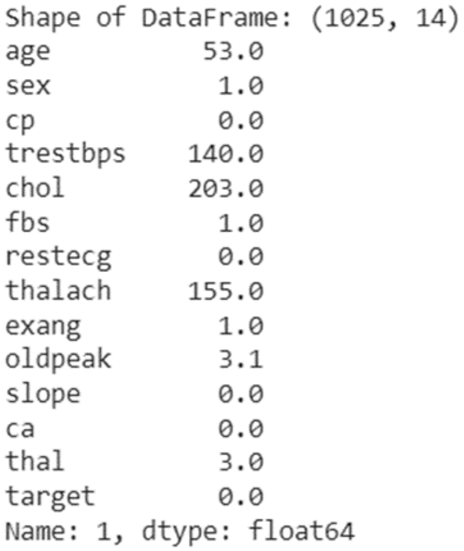
2) Read CSV file

Here, we use read_csv() function to read dataset. This is inbuilt function of panda library. Now, we can print first five rows of heart csv dataset.

data = pd.read_csv('heart.csv')
data.head()

Now, we will see dimension of each attribute, so for that we find shape of the dataframe.

print('Shape of DataFrame: {}'.format(data.shape))
print (data.loc[1])



```
Shape of DataFrame: (1025, 14)
age          53.0
sex           1.0
cp            0.0
trestbps    140.0
chol        203.0
fbs           1.0
restecg       0.0
thalach     155.0
exang         1.0
oldpeak       3.1
slope         0.0
ca            0.0
thal          3.0
target        0.0
Name: 1, dtype: float64
```

Fig.4: Shape of attribute

3) Data cleaning process

A vital stage in data preparation that guarantees the accuracy and dependability of analytical results is the removal of missing data. Missing values can be handled using a variety of strategies, including interpolation, imputation, and deletion. Frequently, missing data is removed from rows or columns; however, caution must be exercised in this process to prevent major data loss. Whereas interpolation uses nearby data points to forecast missing values, imputation uses the information that is now available to estimate the missing values. To preserve the integrity of the dataset and reduce its influence on further analysis, the technique of choice must be carefully considered.

```
# drop all rows wich contains NaN values
data = data.dropna(axis=0)
data.loc[280:]
```

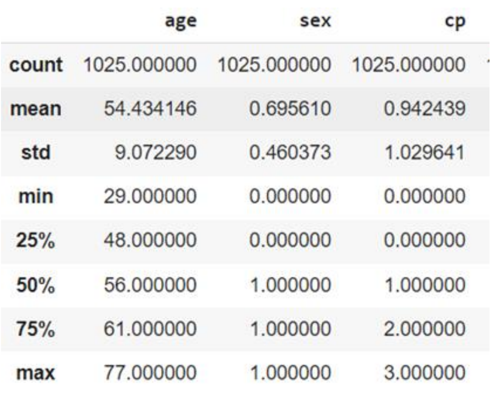
4) Transformation of data to numeric

We need to convert all the data into numeric to process the data and for further analysis. We use to_numeric() function to transform attributes value.

```
# transform data to numeric for further analysis
data = data.apply(pd.to_numeric)
data.dtypes
```

5) Print data characteristics for further analysis

A thorough description of a dataset's key trends and dispersion metrics is given by this function. It provides a brief summary of the distribution and properties of the data by incorporating statistics such as mean, median, standard deviation, minimum, and maximum values. In exploratory data analysis, the "describe" function is particularly helpful since it helps researchers, analysts, and data scientists understand the underlying patterns and characteristics of the data they are working with.



	age	sex	cp
count	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439
std	9.072290	0.460373	1.029641
min	29.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000
50%	56.000000	1.000000	1.000000
75%	61.000000	1.000000	2.000000
max	77.000000	1.000000	3.000000

Fig.5: Data characteristics

6) Statistical analysis by using matplotlib

Histograms provide a clear representation of the frequency or count of values within specific intervals, offering insights into the overall shape and characteristics of each variable. By plotting histograms, analysts can quickly identify patterns, outliers, and potential skewness in the data.

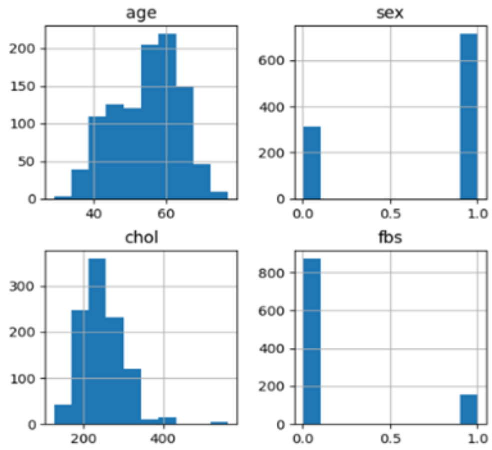


Fig.6: Histogram

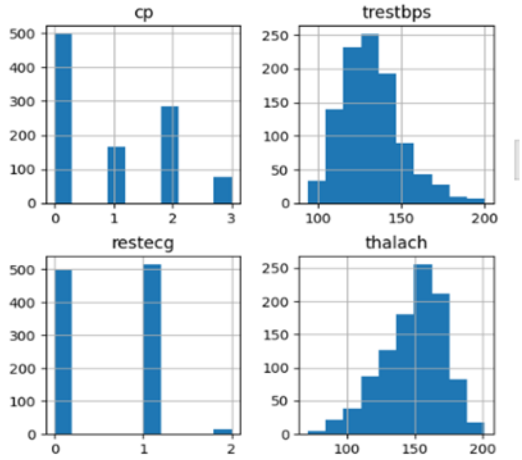


Fig.7: Histogram

7) Plotting heatmap and correlation

A heatmap is a data visualization where colors are used to represent the values in a matrix. Visualizing the strength of the associations between two categorical variables is a popular application for it. The primary goal of correlation plots is to show the correlation coefficients between numerical variables. These charts frequently take the shape of matrices, where the correlation between two variables is displayed in each cell.

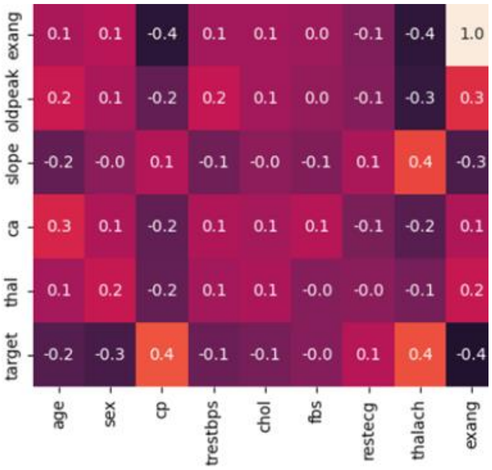


Fig.8: Heatmap

8) Splitting training and testing dataset

Usually, a random split is created, with the bulk of the data going to the training set for training the model and the rest going to the testing set for validation. This section makes that the model is evaluated using untested data, giving a trustworthy indication of how well it generalizes. When the training-testing split is properly implemented, it can assist avoid overfitting, which is the phenomenon where a model works well on training data but is unable to generalize to new, unobserved cases. Here, we split our dataset into two parts, first part contains 80% of the data which is called training dataset and second one is contains 20% of the data which is known as testing dataset. We will use one library Sklearn's train_test_split() function to split the dataset. Further, we convert the data into categorical labels using to_categorical() function.

9) Building and training neural network

Having finished processing our data and dividing it into training and testing datasets, we are now ready to start developing a neural network to tackle this classification issue. We will define a basic neural network with one hidden layer using Keras. Given that this is a categorical classification problem, we will train using a categorical_crossentropy loss and apply a softmax activation function in the network's last layer.

10) Results

The following graph shows the result of the model. The graphs are of model's accuracy and model loss.

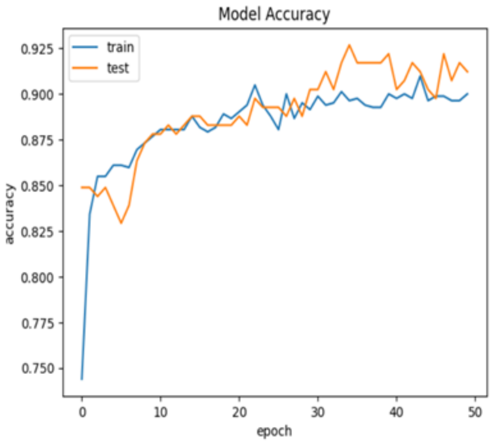


Fig.9: Model accuracy

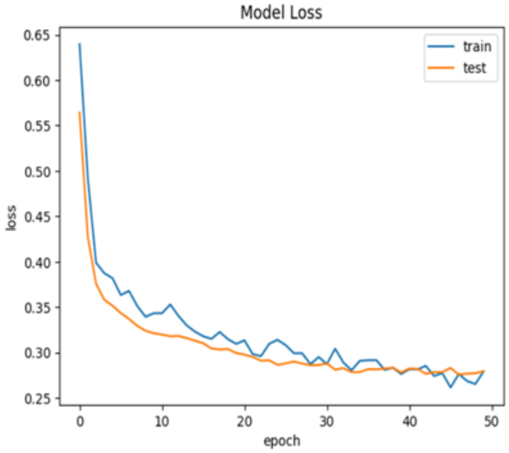


Fig.10: Model loss

11) Improving result

Even with our encouraging results, we still have a sizable error. This may be due to the fact that it is highly challenging to differentiate between heart disease's four severity grades. By transforming the data into a binary classification problem—heart disease or no heart disease

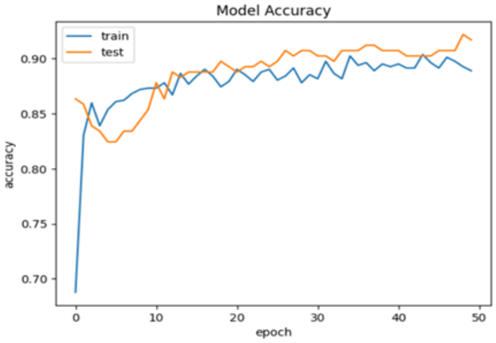


Fig.11: Model accuracy

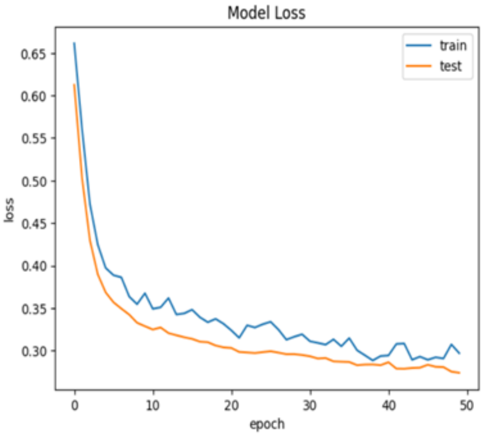
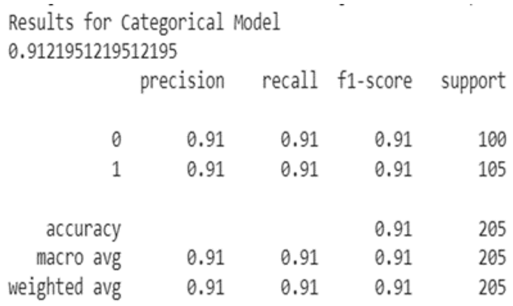


Fig.12: Model loss

V. RESULTS AND METRICS

Now, test the performance of both model categorical and also binary. To do this, we need to make predictions on training and testing dataset.

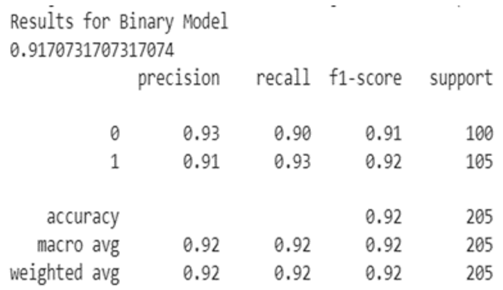


```
Results for Categorical Model
0.9121951219512195
              precision    recall  f1-score   support

           0       0.91      0.91      0.91       100
           1       0.91      0.91      0.91       105

    accuracy                           0.91       205
   macro avg       0.91      0.91      0.91       205
weighted avg       0.91      0.91      0.91       205
```

Fig.13: Categorical model accuracy



```
Results for Binary Model
0.9170731707317074
              precision    recall  f1-score   support

           0       0.93      0.90      0.91       100
           1       0.91      0.93      0.92       105

    accuracy                           0.92       205
   macro avg       0.92      0.92      0.92       205
weighted avg       0.92      0.92      0.92       205
```

Fig.14: Binary model accuracy

CONCLUSION

The goal of this research paper is to predict the heart disease using neural network. This research proposes a neural network architecture in which we have used two classification model. One is binary and the other one is categorical model. In this study we have improved accuracy of the model by converting the data category into 0 and 1. Also used matplot library to analyses the attributes and improve the efficiency of the model.

REFERENCES

[1] Narin, A.; Isler, Y.; Ozer, M. - “Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability.” 2016, Medical Technologies National Congress

[2] Shah, D.; Patel, S.; Bharti, S.K. - “Heart Disease Prediction using Machine Learning Techniques.” *SN Computer. Sci.* 2020.

[3] Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; - “Heart disease and stroke statistics—2019”.

[4] Shorewala, V.- “Early detection of coronary heart disease using ensemble techniques.” 2021.

[5] Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. – “ Heart disease and stroke statistics”—2015

[6] Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P. “Work stress and cardiovascular disease: A life course perspective.” 2016.

[7] Mohan, S.; Thirumalai, C.; Srivastava, G. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” *IEEE*, 2019.

[8] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. “Predicting the risk of heart disease using advanced machine learning approach.” 2020.

[9] Breiman, L. Random forests. *Mach. Learn.* 2001.