Heart Failure Detection Using Artificial Intelligence

A Project By:

Name	Section	BN
Mostafa Ibraheem Basheer	2	31
Ahmed Osama Ismail	1	2
Ahmed Gamal Abbas	1	3

The Source Code is provided via a hyperlink in the references page.

Index

Cover………………………………………………………………………… 1

Index………………………………………………………………………… 2

Introduction…………………………………………………………………3

The data……………………………………………………………………4

Data Acquisition……………………………………………………………4

Data preparation……………………………………………………………5

Data visualization…………………………………………………………6

Method……………………………………………………………………8

Candidate models…………………………………………………………8

Model Results………………………………………………………………8

Model Selection……………………………………………………………8

Tuning………………………………………………………………………9

Summary…………………………………………………………………10

References………………………………………………………………11

Introduction

We would like to shed some light on a very
vital matter that endanger many lives if not diagnosed early; Heart failure is a very common disease especially in our country which what led us to believe that this is the problem that we would like to solve, so in our case we will tackle how we can make a model that **predicts heart failure**.

Cardiovascular diseases cause over 18.9 million
deaths globally each year, They are responsible for approximately 31% of all health-related deaths worldwide, one in five people will develop hearth failure and half of that will potentially decease[1].

it's Critical to try and prevent HF diseases by accurate prediction, Machine learning models such as decision trees, logistic regression, and SVM were used in similar prediction and classification tasks[2], but there are some challenges that should be taken into consideration like the need of algorithms that must be developed to allow the full integration of the widely diverse data available in the EHR, ranging from textual medical reports, a wide variety of imaging data formats[1].

Heart failure prediction is critical to accurate application of many available therapeutic options, such as pharmacologic, highly invasive mechanical ventricular assistance and cardiac transplantation[3].

Available literature between January 2017 and April 2022 by a search of the PubMed library database for relevant papers in our topic we found more than 685 publications with a significant increase in the number of papers published annually.

The Data

A. Data Acquisition

Now let us talk about what really matters;the Data.
We found a couple of datasets that will assist us in our goal:

a) Open Data Commons' Heart Failure Prediction Dataset (containing 11 different features).

b) Data files © Stroke Prediction Dataset (contains 11 features with 5000+ records).

c) Several Data sets was also found in the reviewed literature, However; They had many issues that would've made the preprocessing phase alot longer than the previous datasets.

The table below shows a sample of the used dataset.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Upon visualizing the data we see that it's almost evenly distributed between healthy and diseased records.

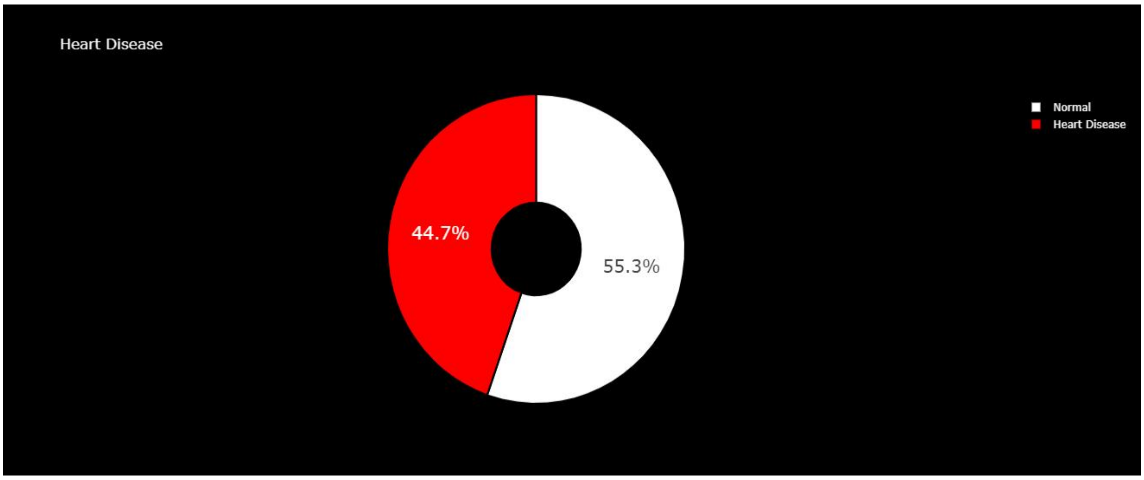

Fig.(1) data distribution

B. Data Preparation

The cholesterol level feature has some values that are equal to zero which is not a record that can be used[4](most likely the CL was not measured for this patient): we can fix that by removing all the records that are related to any patient with zero cholesterol level.

We also preformed several preprocessing methods such as binary encoding for the some features and one-hot encoding for categorical features to have a total of 19 features.

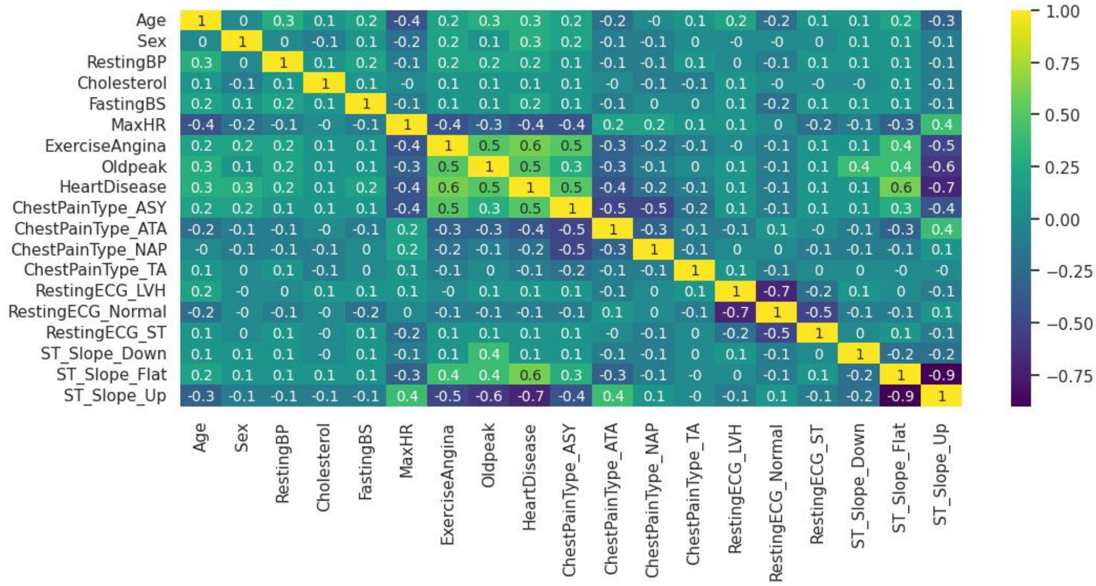

Fig.(2) Correlation map between the features.

using a python feature selection library based on corelation, 14 features were selected.

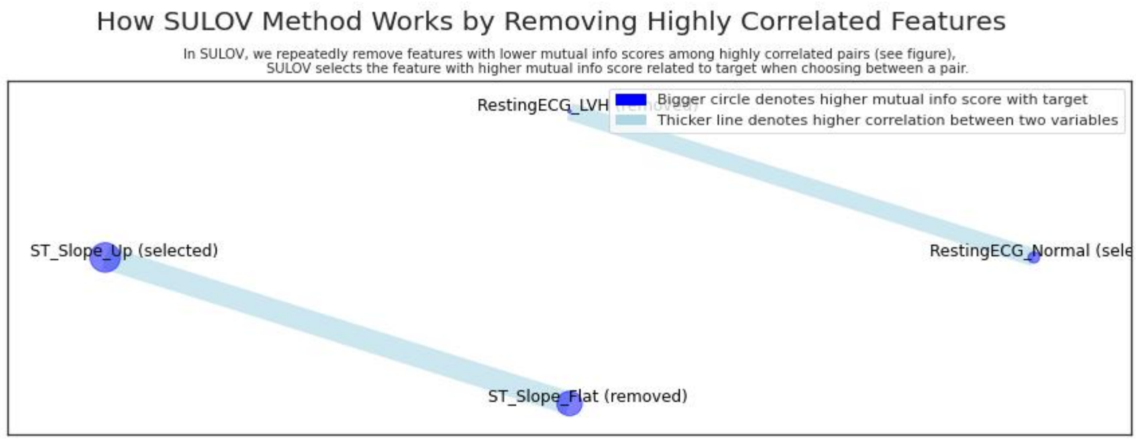

Fig.(3) SULOV feature selection method.

12 out of the 19 features were selected.
The feature selection algorithm output was as follows:

```
Selected 12 important features:['ST_Slope_Up', 'ChestPainType_ASY', 'ExerciseAngina', 'Sex',
'RestingECG_ST', 'ChestPainType_TA', 'FastingBS', 'RestingECG_LVH', 'ST_Slope_Down',
'Oldpeak', 'ChestPainType_ATA', 'ChestPainType_NAP']
```

Fig.(4) output snippet of SULOV algorithm

C. A closer look at the different features

Now we take a look at some of the features and some of the observations we found out.

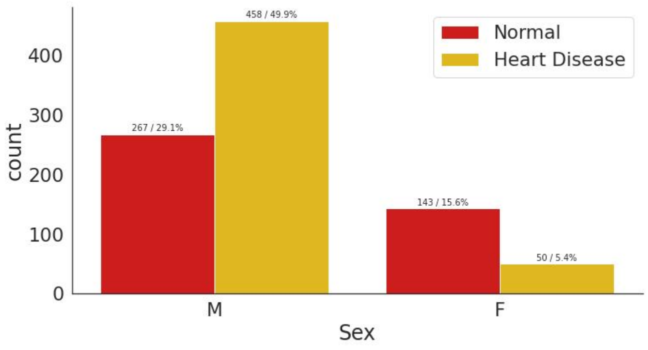

Fig.(5) histogram for diseased and healthy records for males (M) and females(F)

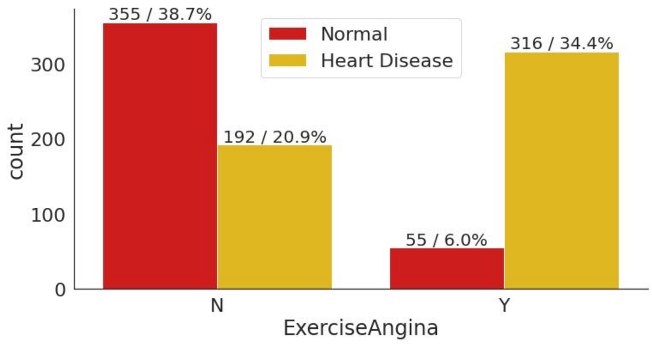

Fig.(6) histogram for diseased and healthy records with exercise angina (Y) and without (N)

Figure (5) shows that males are more likely to have heart disease more than females.

Figure (6) shows that Heart disease is often diagnosed when exercise-induced angina is present.

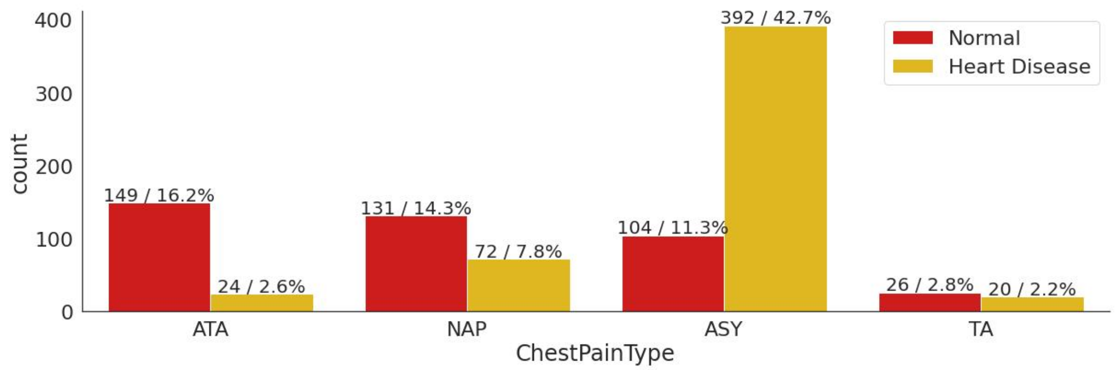

Fig.(7) histogram for diseased and healthy records for the different chest pain types:TA: typical angina, ATA: atypical angina, NAP: non-anginal pain, ASY: asymptomatic

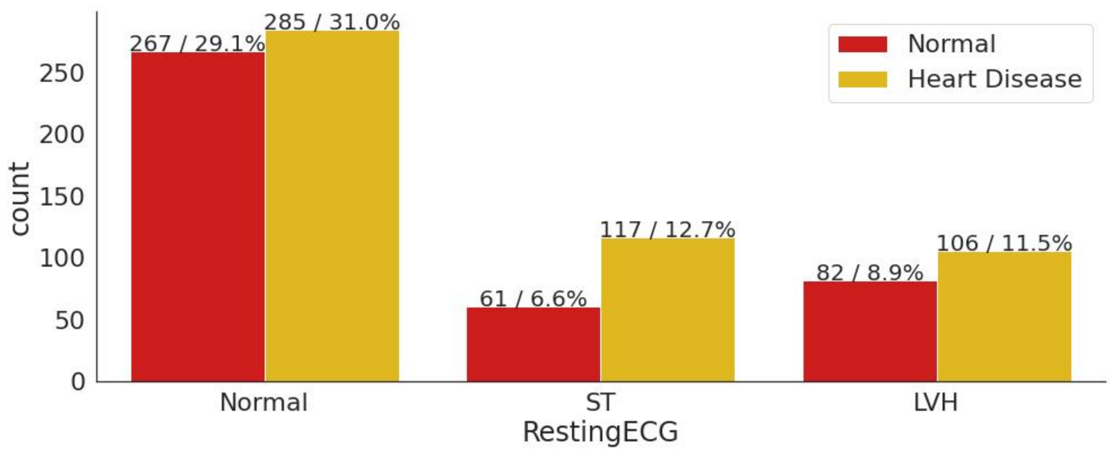

Fig. (8) histogram for diseased and healthy records with different ECG resting potential:
Normal , ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria

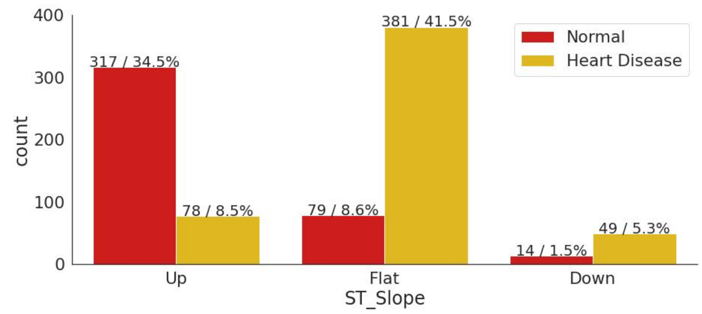

Fig.(9) histogram for healthy and diseased records with the different Slopes of the ST-T wave

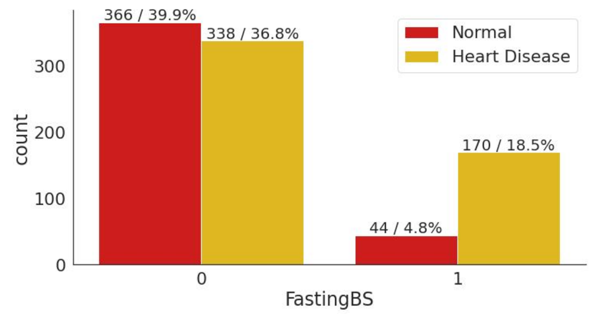

Fig.(10) histogram for healthy and diseased records with the different fasting boold sugar fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

Figure (7) shows that Asymptomatic patients are most likely to be diagnosed as having heart disease.

Figure (9) shows that if the slope of the ST wave is downward or flat, patients are more likely to be diagnosed with heart disease.

Figure (10) shows that patients with a fasting blood sugar more than 120 mg/dl is more likely to have heart disease.

From these observations we can also conclude that the all the selected features are indeed important and not redundant or uncorrelated to our model.

Method

B. Candidate Models

In order to find a suitable model for our problem we have to define what type of ML it is.
Our problem is to find whether some one has a heart disease or not Which makes it a **Classification** type in general or a binary classification in particular.

We covered several classification models in our course:

1. Logistic Regression: used in both classification and regression
2. K-Nearest Neighbour : a good candidate as the record size is less than 1k.
3. Decision Tree: also used in classification.
4. Random Forest (Ensemble Learning): Random forest is also viable since it uses decision tree as a weak learner.
5. Support-Vector Machine : exceedingly efficient in binary classification.

Logistic Regression , KNN and the Decision tree was used in previous attempts ,We also added Ensemble Learning and SVM later.

A. Model Results

After testing all the above models we are met with the following testing scores:

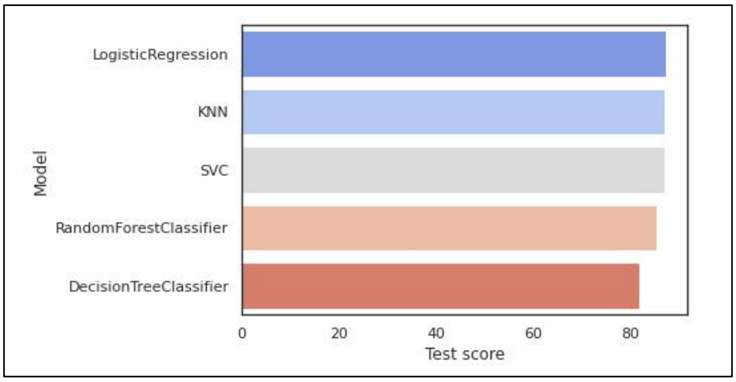

Fig.(11) Bar plot for models score

- Logistic Regression : 87.4%
- K-Nearest neighbour : 87%
- Decision Tree: 81.8%
- Random Forest : 85.4%
- Support-Vector Machine: 87%

Upon closer inspection of the scores of the different models we can say that none of them actually preform badly since all of them are used for classification but naturally some preformed better than others.
The Logistic Regression , KNN and the SVM preformed almost the same with a slightly higher score for Logistic Regression.

C. Model Selection

Previously we decided to only chose the logistic regression model since it yielded significantly more score than the rest but now we have to go a little deeper to determine the winner.

We will go fourth with the Logistic Regression , KNN and SVM
Until We decide the best option and in order to do that we have to tune them as best as possible in order to truly maximize their score, And the only way we can that is by adjusting their hyper parameter.

Hyper parameters Tuning

First we specify which parameters we are going to adjust.

1. For logistic Regression we will adjust is the C parameter which controls the penalty strength[5].
2. For KNN we will adjust the k parameter which accounts for the number of selected neighbors
3. For SVM we will adjust the C which also adjusts the penalty which adjusts the boundary in return[6].

We can determine the best hyper parameter value via Cross validation using grid search.

Results:
1. C for logistic Regression = 10
2. K = 10
3. C for SVM = 1

New Scores:
1. Logistic Regression : 87.4%
2. KNN : 87%
3. SVM : 88.3%

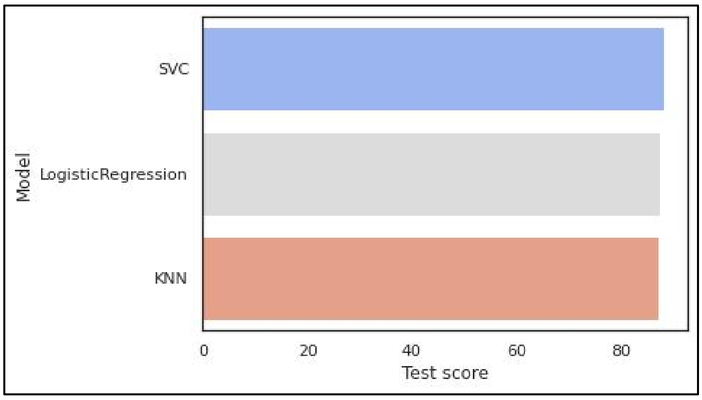

Fig.(12) final test score after tuning the parameters.

Although the scores are still pretty similar but we got more out of the SVM model, Thus we select that model as our algorithm for the problem.

Summary

Heart failure is a matter of life and death especially in the developing countries which has the most amount of diseased patients.
Early detection of HF can be very helpful and can save many lives.
Our goal is to develop a ML model accurate enough to detect that failure which in return can help treat these patients.

Several Datasets were found and we selected a suitable one with 11 different features and a 1K record count, We also made some preprocessing such as one hot encoding and removing some records with faulty or missing data, we also used a feature selection algorithm to the newly created features and 12 out of 19 features were selected and the rest had very high correlation so they were discarded .

Since our problem can be a classification type we can turn to the models that will help us in that specific type, several models were used like KNN , SVM and Logistic Regression , after tuning them up using grid search cross validation; the SVM yielded the highest score of 88.3% which is a very satisfying accuracy for us.

References

1. Heart Failure Diagnosis, Readmission, and Mortality Prediction Using Machine Learning and Artificial Intelligence Models.

2. Clinical applications of machine learning in the diagnosis, classification, And prediction of heart failure.

3. A Machine Learning Approach for Chronic Heart Failure Diagnosis.

4. Cholesterol levels and age.

5. Tune Hyperparameters for Classification Machine Learning Algorithms.

6. Hyperparameter Tuning for Support Vector Machines — C and Gamma Parameters.

7. Source Code for the Project.

All References are clickable.