Determining which physical parameters are significant for heart disease

Zhuoning Li\textsuperscript{1}, Dingxin Tao\textsuperscript{2*,}, Jiaao Zheng\textsuperscript{3}, Chenhao Zhu\textsuperscript{4}

\textsuperscript{1}Shandong Experimental High School
\textsuperscript{2}Beijing Mingcheng Academy
\textsuperscript{3}Seton Hall Preparatory School
\textsuperscript{4}Shanghai hongrun Boyuan school
\*Corresponding author’s e-mail: dingxin.tao@go.shoreline.edu

Abstract. Heart disease currently, as one of the most important causes of death in the world which can take one’s life immediately, should be taken seriously. In 2019, there were almost 9 millions people died from this disease, and the mortality was still increasing. This study aims to warn people be more focus on their heart condition, and provide a further suggestion about whether they can speculate their heart disease at home. In order to test the hypothesis that the cardiac parameters examined at home can also demonstrates the correct heart condition, many different machine learning algorithms were used to predict the final heart disease accuracy by different data sets (from home and from hospital). Also, there would be spot diagram and Pierson correlation diagram to show the correlation between features and condition, and the correlation among the features. In the end, chi-square method was used to determine the influential degree of those factors which can cause the heart disease. The results of these illustrated that even though several important factors which could only be examined in the hospital rather than at home played the leading role of predicting heart disease, people could still try to predict or observe the heart condition at home, because accuracy difference of the predicting results were only 2% to 4%. Therefore, the result from home could be the reference of whether people should go to the hospital to do more advanced examinations.

1. Introduction
As our society improves, people prefer to stay up late at night and having an unhealthy diet which causes the incidence of heart diseases to increase. Many heart diseases have minor reactions at first, which means that heart diseases become more pervasive. Therefore, people would be better to have the ability to predict the probability of catching a heart problem at home. This is where this research heading and different from others. Our assumption is that people’s self-diagnosis is accurate enough to predict diseases. To test that, both the data which can be accessed by sensory perception or household instruments and detection with professional instruments in the hospital is required to compare the accuracy.

Twelve algorithms are applied in this research, which will be demonstrated in Result part, with comparison to find out the most accurate one. Same as the research described in Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques(Mohan, Thirumalai, and Srivastava), the more accurate algorithm in this research is respectively MLP classifier and Random Forest Classifier.
2. Background
This paper bases on the data set Heart Disease UCI, published by Ronit and re-processed by Chernges, on Kaggle[1]. Several algorithms such as regression, and KNN are available to be used to process the data. There are many discussions and codes under the data set which have the core idea with our research. The similarity between this research and others is that the basic process is to determine the best factors to predict heart disease. While the goal for this project, which makes it distinguished from others, is to offer a public reference to self-diagnose at home. The information that can be detected or measured by people themselves is considered separately. Comparing the accuracy analyzed at home with that analyzed in hospital, if the difference is not so significant, people can just self-diagnose at home and combine with the most significant factors examined in the hospital to further determine whether or not they have heart disease. If the difference is significant, which factors cause the large difference should be examined, and people would better go to examine in the hospital.

3. Approach
This study first compares the correlation between parameters to find out the factors that have an impact on heart condition through the chart. Then machine learning algorithms will be used to figure out the accuracy of heart disease prediction. With the accuracy of the two data sets (at home and in hospital), it is able to compare the accuracy of two data sets to determine whether people can ultimately predict heart disease at home.

The dataset in this study is from Kaggle which is called Heart Disease Cleveland UCI[1], and it is mainly used to do the prediction research. However, there are already a bunch of machine learning prediction projects have done by others. The most goal of those projects is to find the fittest algorithm to predict the heart condition based on this dataset. But what will be there is to find whether the results examined at home are similar to the results tested in the hospital.

![Flow Chart of the study](image)

Figure 1. Flow Chart of the study

The dataset includes 297 patients’ information and contains 13 parameters and one condition which is the symbol of whether the patient suffers from heart disease.
Table 1. Parameters’ definition

| Parameter | Definition | Meaning of values |
|-----------|------------|-------------------|
| age       | Age of the patient | age in years |
| sex       | Gender of the patient | 1: male 0: female |
| cp        | Chest pain type | 0: asymptomatic 1: atypical angina 2: non-anginal plain 3: typical angina |
| trestbps  | resting blood pressure | (in mm Hg on admission to the hospital) |
| chol      | serum cholesterol | (in mg/dl) |
| fbs       | fasting blood sugar > 120mg/dl | 1: true 0: false |
| restecg   | resting electrocardiographic results | 0: normal |
| thalach   | maximum heart rate achieved | unit: bpm |
| exang     | Exercise-induced angina | 1: yes 0: no |
| oldpeak   | ST depression induced by exercise | |
| slope     | the slope of the peak exercise ST | 0: upsloping 1: flat 2: downsloping |
| ca        | number of major vessels(0-3) | |
| thal      | state of heart illness | 0: normal 1: fixed defect 2: reversible |
| condition | whether the patient has disease | 0: no disease 1: disease |

All these definition and measurement of parameters are from official website and Kaggle[3]

3.1 Data Pre-Processing
The final goal is to compare the prediction accuracies both from home and from the hospital, so that the extremely first step is to divide the dataset into two parts - the home part covering 7 parameters which can be simply examined at home, and the hospital part incorporating the whole 13 parameters. The method there is dividing them directly into two data files.
In home data set, it contains ‘age’, ‘sex’, ‘cp’, ‘trestbps’, ‘fbs’, ‘thalach’, ‘exang’.
In preparation, all parameters in both the home and the hospital dataset except the condition are collected together as the features. Hence, in later portions, those features can be used straightway to do the rest analysis.

3.2 Finding Correlation
In the finding correlation process, the python library used there is Seaborn which can be used to plot the graph to discover certain patterns. In the graph of this portion, there will be 169 different graphs to show the relationship between any two features. The features are on X and Y axis, and the condition including 0 and 1, which means respectively no disease (0) and disease (1), will be the content in the graph. An example is following.
Figure 2. Examples of spot graph

In these two example graphs, orange indicates diseased, and blue indicates non-diseased. It doesn’t hard to get from above that when thal (state of heart illness) becomes greater, the possibility of being ill raises up. The purpose of using this graphing method is to allow researchers to better observe the impact of the data on heart disease directly. More details will show in the result section.

Besides, the correlation between every two factors can be analyzed by Pearson Correlation Matrix. The greater value represents a stronger relationship between the two factors. As shown by the graph, when ignoring how “condition” the exercise-induced angina and chest pain type to relate to each most strongly, indicating by Pearson Correlation coefficient 0.38. But since this smallest value is lesser than 0.6, it does not affect the result.

3.3. Analyzing data

The predicting module used is sklearn in python containing an amount of machine learning predict algorithms that have already integrated so that they can be utilized directly. This takes the same approach that Baris Cal used on Kaggle[2]. In his code, he puts together a number of predictive algorithms to find the one with the highest accuracy. The reason for using these algorithms is to compare the accuracy of them and find the most one. Don’t use the result from others’ projects is because the accuracy will change when there are different datasets.

First of all, these machine learning algorithms which will be used in the research are imported from the python library into code, including GaussianNB, MultinomialNB, and BernoulliNB, KNN, Logistic Regression, Decision Tree Classifier, Support Vector Classifier, Random Forest Classifier, Gradient Boosting Classifier, Stochastic Gradient Descent, Neural Networks, and XGB Classifier. The reason why using such many algorithms is to make the result as accurate as possible. The more algorithms are used, the more opportunity to find the higher accuracy of certain predicting methods there will be.

In the next portion, the features and condition will be named as X_h and y_h. Then the whole data from home is divided into three parts, which respectively are the training part, testing part, and validating part. Then, they will be mixed within their own parts to ensure the randomness of the prediction. There are 297 groups of samples, of which 198 groups are used as the training data, 49 groups are used as validation data, and 50 groups are used as testing data.

```python
X_train, X_test, y_train, y_test = train_test_split(X_h, y_h, test_size=0.33, random_state=101)
X_valid, X_test, y_valid, y_test = train_test_split(X_test, y_test, test_size = 0.5, random_state=42)
```
Then, gathering and unifying all of these algorithms into a dictionary which is called models. In the models, the style and some basic information of those algorithms are set for the following predicting. Meanwhile, the name of those algorithms are also put into a new list. With this list, it is possible to use loop to repeatedly predict and test the data accuracy.

However, before predicting those data, initializing the result containers, which are trainScores, validationScores, and testScores is indispensable. The rest of the prediction is finished by a loop that fit the X value(elements used to predict) and y value(predicting result) of training data values into those models and conserves the consequence into the result containers above one by one.

The most important portion in the prediction is to calculate the Accuracy, Precision, Recall, F1 score, and Specificity. All these consequences are gotten from the result of predicting. The way to calculate those values are based on the confusion matrix, and the graph under this is the code about how to calculate those values.

For the sake of understanding the values better, there will be some charts as to the following to explain the meanings with details. In each cell of the Confusion Matrix, they have diverse significance, which is TN, FP, FN, and TP, which of the explanations beneath this. Meanwhile, the values during the analysis are also calculated by these things. Accuracy is the sum of TP and TN divided by the total number. Precision is TP divided by the addition of TP and FP. The Recall is TP divided by the sum of TN and FP. Finally, the F1 score is twice precision times recall divided by the sum of the precision and the recall values. In one word, the information we need accuracy derives from the formula (TP + TN)/ (TP+FP+TN+FN).

### Table 2. Meanings of cells of Confusion Matrix

| Predicted Class | Actual Class  |
|-----------------|--------------|
|                 | Negative Class | Positive Class |
| Negative Class  | TN(True Negative) | FP(False Positive) |
| Positive Class  | FN(False Negative) | TP(True Positive) |

#### 3.4. Comparison

With the information and result above, the accuracy can be calculated and began to compare the home and the hospital results. If the accuracy at home is similar or a little less than at hospital, it can be defined that people are also able to predict whether they have heart disease at home, but if not, it shows that people should go to hospital when they want to judge the heart condition.

The next step is to find the importance or the influence degree of a feature posting on the condition. The method using in there is Chi-square test which is from the SelectKBest and chi2 from sklearn library. Firstly, the best feature is defined by chi-square test method, and it is fitted to 'fit'. To display it read easily, it should be changed into DataFrame by pandas to show it in a list. Higher score of certain feature demonstrates it is more significant to condition. Doing so can help us further know which factor should be put in the first place to detect and focus.

#### 4. Results

### 4.1. The correlation of factors

To ensure the accuracy of prediction, the precondition have to be finding the correlation between independent variables. When two or more of the independent variables are correlated, the condition is called multicollinearity. The purpose is to avoid this condition occurring by adjusting the variables. If the correlation coefficient is too high (above 0.6), it will affect the consequence of prediction so that the unimportant variables should be removed. Pearson Correlation matrix directly presents the correlation between each two independent variables with coefficient and color. The darker the green is,
the stronger correlated the two factors are. In this graph, both the factors tested at home and in the hospital are presented. Among the factors tested at home, the highest one is between exang (Exercise-induced angina) and cp (chain pain type), which is 0.38. Among the factors tested in the hospital, the highest one is between oldpeak (ST depression induced by exercise relative to rest) and slope (the slope of the peak exercise ST segment), which is 0.58. Since the maximum coefficient is 0.58, less than 0.6, the consequence will not be affected so much.

**Figure 3. Pearson Correlation Matrix in hospital data**

4.2 The accuracy of algorithms:

**The most accurate algorithm at home:**

Since the project employs random sampling in which the samples of training, testing, and validation are made of random data, the accuracy is shown each time has a subtle difference from usual. By examining for several times, MLP Classifier is the most accurate overall, which has the accuracy of 80%. The graph below is one representative of the test of accuracy.
Table 3. Accuracy of each algorithms at home

| Algorithms                  | Accuracy |
|-----------------------------|----------|
| MLP Classifier              | 80%      |
| Gaussian NB                 | 78%      |
| Random Forest Classifier    | 78%      |
| Bernoulli NB                | 76%      |
| Decision Tree Classifier    | 76%      |
| Gradient Boosting Classifier| 76%      |
| Stochastic Gradient Descent | 76%      |
| XGB Classifier              | 76%      |
| Support Vector Machine      | 72%      |
| Logistic Regression         | 64%      |
| K Neighbors Classifier      | 62%      |
| Multinomial NB              | 48%      |

The most accurate algorithm in hospital:
After testing randomly several times, Random Forest Classifier shows to be the most accurate model, which has an accuracy of 82%-84%.

Table 4. Accuracy of each algorithms in hospital

| Algorithms                  | Accuracy |
|-----------------------------|----------|
| Random Forest Classifier    | 84%      |
| Gradient Boosting Classifier| 82%      |
| Bernoulli NB                | 80%      |
| Stochastic Gradient Descent | 80%      |
| Gaussian NB                 | 78%      |
| MLP Classifier              | 78%      |
| XGB Classifier              | 78%      |
| Decision Tree Classifier    | 72%      |
| Support Vector Machine      | 64%      |
| K Neighbors Classifier      | 64%      |
| Logistic Regression         | 57.99%   |
| Multinomial NB              | 48%      |

4.3 Significance of factors
The most significant factors at home:
Chi-square is used to present the significance of each factor. Thalach (maximum heart rate achieved) is the most significant one. Exang (exercise-induced angina), age, and cp (chest pain type) are relatively significant factors that can be detected at home.
Table 5. Significance of factors at home

| feature_name | Score    |
|--------------|----------|
| 5            | thalach  | 187.053104 |
| 6            | exang    | 35.508090  |
| 0            | age      | 22.917697  |
| 2            | cp       | 21.352432  |
| 3            | trestbps | 16.707463  |
| 1            | sex      | 7.444195   |
| 4            | fbs      | 0.002547   |

The most significant factors in hospital:

It can be found out that the most significant factors are thalach (maximum heart rate achieved), thal (thalassemia), ca (number of major vessels colored by fluoroscopy), and oldpeak (ST depression induced by exercise relative to rest). While thalach exceeds other factors remarkably.

Table 6. Significance of factors in hospital

| feature_name | Score    |
|--------------|----------|
| 7            | thalach  | 187.053104 |
| 12           | thal     | 87.903888  |
| 11           | ca       | 82.730613  |
| 9            | oldpeak  | 68.570533  |
| 8            | exang    | 35.508090  |
| 0            | age      | 22.917697  |
| 2            | cp       | 21.352432  |
| 4            | chol     | 20.855084  |
| 10           | slope    | 20.818579  |
| 3            | trestbps | 16.707463  |
| 6            | restecg  | 8.134652   |
| 1            | sex      | 7.444195   |
| 5            | fbs      | 0.002547   |
5. Conclusion
What we learned is that in the hospital, thalach (maximum heart rate achieved, which could also be tested at home), thalassemia, number of major vessels colored by fluoroscopy, and ST depression induced by exercise relative to test are the most important factors to predict heart disease. For the factors tested at home, thalach (maximum heart rate achieved), exang (exercise-induced angina), age are relatively useful in prediction. Because of that, we can understand why testing at home is not as accurate as the test in the hospital since three of the most important factors must be done in the hospital. However, even though a test at home is not as accurate as the test in the hospital, the maximum accuracy of testing at home and in the hospital is 80% and 82-84% respectively. The discrepancy between home test and hospital test is just 2 to 4 percent. Thus, the test at home is still credible to some extent to advise the user about whether they should go to the hospital.[5]

In future work, in order to make the prediction more accurate, many other ways can be applied to increase the accuracy of the forecast like using different algorithms and statistical methods. For instance, there are only 12 different algorithms are adopted, among which the maximized accuracy is about 80%. Other researches are using the same dataset as this research does but have higher accuracy. In that case, trying some different algorithms is able to increase the accuracy. In addition, a data set that contains more factors can be utilized to make the result more accurate. Moreover, instead of just giving the respective significance of factors, a specific index derived through the overall data can be determined to assist people in better diagnose the possibility of having heart disease. By analyzing the data examined at home, an index can be given, exceeding which the person is more likely to have heart disease. Therefore, people can have a reference to decide whether they should get a more comprehensive examination in the hospital.

References
[1] Cherngs. “Heart Disease Cleveland UCI.” Kaggle.com, 29 Mar. 2020, “www.kaggle.com/cherngs/heart-disease-cleveland-uci”. Accessed 13 Mar. 2021.
[2] Cal, Baris. “12 ML Models + Visualization (92% Accuracy).” Kaggle.com, 7 Mar. 2021, “www.kaggle.com/bariscal/12-ml-models-visualization-92-accuracy”. Accessed 15 Mar. 2021.
[3] Ronit. “Heart Disease UCI.” Kaggle.com, 25 June 2018, “www.kaggle.com/ronit/heart-disease-uci/discussion/105877”. Accessed 15 Mar. 2021.
[4] Blog.csdn.net, 4 Sept. 2018, blog.csdn.net/yangyin007/article/details/82385967”. Accessed 7 Apr. 2021.
[5] Melker R J, Euliano N R. Method and Apparatus for Diagnosing Respiratory Disorders and Determining the Degree of Exacerbations[J]. 2008.