Efficient machine learning based detection of heart disease

Ricardo Buettner
Aalen University, Germany
ricardo.buettner@hs-aalen.de

Marc Schunter
Aalen University, Germany
marc.schunter@studmail.htw-aalen.de

Abstract – This paper describes a method to detect possible heart disease using the Random Forests algorithm. Cardiovascular diseases are the number 1 cause of death globally - an estimated 17.9 million people died from it in 2016. This machine learning work contributes to healthcare and can detect heart disease on the basis of clinical data and test data from different patients. The result and contribution of this paper is to identify whether a patient has heart disease or not, based on the information of clinical data and test results and so support doctors in making decisions about patient treatments.

Keywords – Healthcare, Heart Disease, Machine Learning, Random Forests

I. INTRODUCTION

The World Health Organization (WHO) classifies cardiovascular diseases as the number one cause of death globally. In total, 17.9 million people died from cardiovascular diseases in 2016, representing 31% of all global deaths. Cardiovascular diseases are disorders of the heart and blood vessels. Four out of five cardiovascular diseases deaths are due to heart attacks and strokes. Individuals at risk of cardiovascular diseases may demonstrate raised blood pressure, be overweight or obese [1].

Among the adult population, cardiovascular diseases are the main health problem in general. It mainly affects the heart and the arteries of the brain, heart and legs. Therefore, the lack of blood supply not only damages the heart, but also the legs and brain, which can lead to health disorders prompting a risk of heart attacks, thrombosis or rupturing of blood vessels, among others [2].

The main risk factors were defined in the Framingham Heart study published in 1952 and are listed as follows [3]:
- Age
- Gender
- Body Mass Index
- Smoking Condition
- Homocysteine
- Reactive C-Protein
- Fibrinogen
- Previous familiar cases
- Diet
- Cholesterol HDL Triglycerides Lipoprotein
- Sedentary Condition
- Glucose Tolerance and Metabolic System
- High blood pressure

The WHO defines unhealthy diet, physical inactivity, tobacco use and excessive use of alcohol as the most important behavioral risk factors for heart disease. These “intermediate risk factors” can be measured in primary care facilities and indicate an increased risk of developing a heart attack and other complications [1]. Some of this information can be provided immediately, while in the other cases tests need to be done. These can include blood tests or an electrocardiogram. An electrocardiogram is a diagnostic tool that is routinely used to measure and record different electrical potentials of the heart. Willem Einthoven developed the ECG method in the early 1900s, and while it is a relatively simple test to perform, the interpretation of ECG tracing requires a significant amount of training [4-5].

![Figure 1. Elements of the ECG-complex. [4]](image)

The P wave of the ECG looks at the atria. The QRS complex looks at the ventricles and the T wave evaluates the recovery stage of the ventricles while they are refilling with blood. The ST slope and ST depression, induced by exercise, is part of the database which is used for the method in this paper.

Generally, many health care organizations are facing a major challenge to offer high quality provisions, like diagnosing patients correctly and administrating treatment at reasonable costs. Machine learning techniques have been widely used to mine information from medical databases. In Machine Learning, classification (e.g.: is this specific patient sick or healthy) is a supervised form of learning that can be used to design models describing important data classes. Using those machine learning techniques can support researchers or physicians in making medical decisions and they can answer important and related questions concerning health care.
II. DATABASE AND METHOD

Method

For Supervised Machine Learning Algorithms there are multiple techniques. Some examples include: Nearest Neighbor, which classifies a set of test data based on the k Nearest Neighbor algorithm using the training data [6-7]. Naive Bayes, which is the simplest form of Bayesian network calculates a set of probabilities by counting the frequency and combinations of values in a given data set [8-9]. Support Vector Machines output an optimal hyperplane which categorizes new examples between labeled training data [10-11].

The Decision Tree is a tree-based flowchart model, in which each internal node represents a “test” on an attribute. Each branch represents the outcome of the test and the leaves are a class distribution. The different paths from the root to a leaf represent a classification rule [12-13].

The machine learning technique used in this paper is the Random Forests. It is used to classify whether a person has a heart disease or not, based on clinical information and test results about a group of patients.

The Random Forests algorithm is a popular and very efficient algorithm, for both classification and regression problems. The principle of Random Forests is to combine many binary decision trees by using several bootstrap samples coming from the learning sample and choosing randomly at each node a subset of variables [14-15].

Liaw and Wiener described the Random Forests algorithm (for both classification and regression) as follows [17]:
1. Draw \( n_{\text{tree}} \) bootstrap samples from the original data.
2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample \( m_{\text{try}} \) of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of Random Forests obtained when \( m_{\text{try}} = p \), the number of predictors.)
3. Predict new data by aggregating the predictions of the \( n_{\text{tree}} \) trees (i.e., majority votes for classification, average for regression).

Validation

In order to achieve a reliable result from the Random Forests, cross-validation was used. Cross validation divides the data set into a specific number of subsets. Each subset is used by repeating both as a training record and as a test record. The error estimates of all rounds are then summarized and averaged [18].

As used in the method by Rieg et al., a 10 times 10-Cross-Validation was applied. As the result, the algorithm revealed which subjects were correctly classified. For this purpose, a confusion matrix was generated, and the cross validation classified as follows [19]:
- True positive: The subject has heart disease and the algorithm has correctly indicated it.
- False negative: The subject has heart disease, but the model has falsely classified him as being without heart disease.
- False positive: The patient does not have heart disease, but the model has classified him as a person with heart disease.
- True negative: The patient does not have heart disease and the algorithm has not classified him as a person with heart disease either.

![Figure 2. Example of a decision Tree](image)

![Figure 3. Simplified Random Forest](image)

![Figure 4. Summary of the full process](image)
The database which is used for this paper comes from a publicly accessible repository [20]. The data sets itself are provided by V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Database Characteristics and Preprocessing

The database contains information about different patients with or without heart disease. First, there are four different variables concerning clinical information: The patients’ age and gender, the patients' resting systolic blood pressure in mm Hg at the time of admission to the hospital, and four different classifications of chest pain: a) typical anginal, b) atypical anginal, c) nonanginal and d) asymptomatic.

The youngest patient is 29 years old, the oldest is 77. Sixty-eight percent of the patients are male and 32% are female. Fifty-four percent of all patients have heart disease, while 46% do not. Comparing the heart disease rate between female and male, 75% of all female patients in this database have heart disease. The male patients have heart disease in 45% of all cases, while 55% do not have heart disease.

In addition to the clinical information, which can be recorded immediately by a medical staff, e.g. doctor or nurse, the database contains multiple test data: [20].

- Cholesterol Measurement: Serum cholesterol in mg/dl
- Fasting Blood Pressure: Boolean measure indicating if fasting blood sugar is higher than 120 mg/dl
- Resting Electrocardiographic Measurement: Classified in Three different types of values: “normal”, “Having ST-T wave abnormality” and “Showing probable or definite left ventricular hypertrophy by Estes criteria”
- Maximum Heart Rate achieved
- Exercise induced angina: Boolean measure indicating whether exercise induced angina has occurred: 1 = yes, 0 = no
- Oldpeak: ST depression brought about by exercise relative to rest
- Slope: The slope of the ST segment for peak exercise. There are three different types of values: upsloping, flat, down sloping
- Major Vessels: The number of major vessels (from 0-3)
- Thalassemia: The hearts status (normal, fixed defect or reversible defect)
- Heart Disease (class attribute): Shows if the patient has a heart disease or not

In order to improve the variables for the Random Forests algorithm, the values of several attributes needed to be grouped in finite ordered lists of elements, so called tuples.

The first grouping is the age of the patients. Instead of using the detailed age of every person, the numbers have been grouped to Children (younger than 18) and “10-year packages” based on WHO’s Index ages and age groups [21].

The second grouping contains information about the resting blood pressure (Table 2). The patients’ resting blood pressure has been categorized in normal and elevated blood pressure as well as in three different stages of hypertension [22-23].

The cholesterol levels in the database are measured in milligrams (mg) of cholesterol per deciliter (dL) of blood. Those single values have been grouped in three different categories (Table 3) [24].

The last tuple to be created was for the ST depression in the electrocardiography test results of the patients. The grouping uses a similar logic to the age grouping. The first tuple goes from 0-0.4 millivolts, followed by a 1 millivolt step for each tuple (Table 4).
III. RESULTS

The trained model was tested with 303 patients. The cross validation was repeated 10 times on the training set. Table 5 shows the result of the confusion matrix including all correctly and wrongly classified heart diseases. Table 6 shows the performance of the model. The random forest reached an average accuracy of 84.448%.

The model was also tested without the 10 time cross-validation. Instead, the Random Forest was just trained and tested in one loop with 75% training data and 25% test data. The result was an accuracy of 82.895%, which is 1.553% less than executing the algorithm with a 10 times cross validation included.

<table>
<thead>
<tr>
<th>TABLE 5. CONFUSION MATRIX</th>
<th>Predicted</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Heart Disease</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>No Heart Disease</td>
<td>118</td>
<td>20</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>27</td>
<td>138</td>
</tr>
</tbody>
</table>

TABLE 6. METHOD PERFORMANCE

<table>
<thead>
<tr>
<th>Performance indicator</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>84.448%</td>
</tr>
<tr>
<td>Correct classified</td>
<td>256</td>
</tr>
<tr>
<td>Wrong classified</td>
<td>47</td>
</tr>
<tr>
<td>Error</td>
<td>15.512%</td>
</tr>
<tr>
<td>Cohen’s Kappa</td>
<td>0.689</td>
</tr>
</tbody>
</table>

Analyzing the decisions taken in the model shows the important variables. Serum cholesterol in mg/dl, the measure indicating if fasting blood sugar is higher than 120 mg/dl, Maximum Heart Rate achieved, and the Resting Electrocardiographic Measurement seem to be less important:

On the other hand, the four different classifications of chest pain (typical anginal, atypical anginal, nonanginal and asymptomatic), the number of major vessels and the hearts status (normal, fixed defect or reversible defect) seem to be very important for the classification of Heart Disease (see Fig. 5).

IV. DISCUSSION

An early detection of heart disease is very important to save lives. Therefore, a fast and reliable diagnosis is essential. As mentioned at the beginning, Machine Learning and Data Mining can be powerful support instruments by providing the necessary information and classification to diagnose a disease, based on the given input data. The Random Forests algorithm used in this paper has shown that it can be used as such an instrument and that it has achieved the original target of this paper: Supporting doctors (or medical staff in general) in diagnosing heart diseases and providing an efficient solution for classification, whether a person is sick or not.

V. CONCLUSION

This paper presented a new approach to heart disease classification, using the Random Forest machine learning algorithm and attributes based on clinical data and patient test results. It reached an overall accuracy of 84.448%. The highest accuracy was reached while using an additional 10 times cross-validation in the process and it outperforms other machine learning techniques using the same database [25]. Using the Random Forests algorithm without the cross validation secured an overall accuracy of 82.895%.

Limitations and future research

While we intensively evaluated other traditional machine learning approaches such as clustering [27] and also most modern convolutional neural networks, which are outstanding in other domains such as image recognition [28-30], we achieved the best results here with decision trees. However, the method of choice always limits scientific understanding. Hence our study has these limitations:

To further improve the accuracy of the algorithm, updating the database with more information and attributes could help to increase the level of accuracy already achieved. As mentioned in the introduction to this paper, it is already known that Age, Gender, Body Mass Index, Smoking Condition, Homocysteine, Reactive C-Protein, Fibrinogen, Previous familiar cases, Diet, Cholesterol HDL Triglycerides Lipoprotein, Sedentary Condition, Glucose Tolerance and Metabolic System, High blood pressure are risk factors for heart attacks. But some of that relevant information is not available in the used data set [20]. For example, beside the information about the Serum cholesterol in mg/dl, there is no information about the constitution of the patient. Adding information like weight or the Body Mass Index (BMI) could increase the information level of the database. Also, the information

Figure 5. Most predictive attributes [%]
about the Smoking Condition is missing. Both, tobacco use and an unhealthy diet are significant reasons for heart attacks and strokes, based on the information provided by World Health Organizations key messages.

Other important information which is missing from the database is a patient’s family history and other individual differences such as personality [31, 32].

Although all the listed reasons for heart diseases are known, their epidemiological relevance is different from case to case. Therefore, the attributes probably need to be weighed correctly. In order to estimate the risk of suffering from a heart disease, a global evaluation should be added to the information in the database as well. One solution could be the Anderson Table [26].

In future work we will triangulate simple ECG sensor data with other physiological sensor data (i.e., heart rate variability [33], electroencephalography [34], electrodermal activity [35], eye fixation [36-38], eye pupil diameter [39-42]). Furthermore, we will experimentally evaluate whether our novel approach is also robust under various conditions of a user's cognitive workload [43-46], concentration [47], and mindfulness [48, 49]. In addition, we will report common method bias evaluations [50, 51] and the results of transferring our novel spectral method [52] to ECG, where we already achieved outstanding results in predicting diseases such as schizophrenia [53, 54], epilepsy [55], and sleep disorder [56, 57] based on electroencephalographic data. Finally, we will conduct an empirical implementation study to evaluate acceptance [58-60] and trust [61] by physicians and patients and if the automated approach improves the coordination [62-63] between physicians more efficiently.

ACKNOWLEDGMENT

We thank the reviewers, who provided very helpful feedback on the refinement of the paper. This work is partly funded by the German Federal Ministry of Education and Research (13FH4E031A, 13FH176PX).

REFERENCES


