Coronary Heart Disease predicting using machine learning algorithms

Aladin Kapić   
Faculty of Engineering and   
Natural Sciences  
International Burch University, IBUSarajevo, Bosnia and Herzegovina  
aladin.kapic@stu.ibu.edu.ba

Jasmin Kevrić  
Faculty of Engineering and   
Natural Sciences  
International Burch University, IBUSarajevo, Bosnia and Herzegovina  
jasmin.kevric@ibu.edu.ba

*Abstract* – Framingham Heart Study dataset was used in this particular article. The credibility of data collected over the years, found its purpose in machine learning algorithms. Especially because of modern life, chronic heart disease (CHD) has a great presence. To summarize causes, explain reasons and predict diseases, machine learning algorithms found their purpose in these sphere too. Many questions were answered during the last century, but combining them together to accomplish greater goal, is modern time problem. Unknown values in dataset required preprocessing of data, later on, imbalanced dataset caused application of SMOTE algorithm to achieve better results. After many attempts, conclusion is to make almost complete artificial training set and give it a little go on real world data. Results are more than better and far away than expected, around 97 percent of positive classification of patients that have Coronary Heart Disease. Of course, single classification algorithm was not good enough so we had to use combination of Random Forest and KStar so combined achieved results at an enviable level.

Keywords - CHD, heart disease, artery disease, FHS, heart attack, arrythmia, heart failure, SMOTE, Random Forest, KStar, meta.vote

# Introduction

Considering only heart, people discovered many types of heart disease and each one is kind a unique with its own symptoms and treatment. People started seriously researching causes of them since beginning of twentieth century. Framingham Heart Study (FHS) was founded in 1948 and main goal was to find common factors or characteristics that contribute to cardiovascular disease. [1] That information’s were unknown at that time, so people were wondering about main reasons of heart failure. FHS had over 14,000 people from three generation, since its founding. Now, the data collected is used in combination with machine learning algorithms to analyze body condition, predict diseases like CAD, CHD, etc. With that aspect of view, our data is based on Coronary Heart Disease (CHD) and in next chapter we are going to talk about causes, treatments and algorithms that provides either adequate support for doctor’s opinions or exact answers of such disease. Since 1948 till now, it appears to be the most famous and influential investigation in cardiovascular disease epidemiology.

Coronary heart disease (CHD) is disease with characteristics of atherosclerotic plaques. Plaque is waxy substance that builds up inside coronary arteries. In that case, it comes to narrowing coronary arteries or even completely stopping flow of oxygen-rich blood. Heart has its own blood supply channels, called left and right coronary arteries. They are branching in lot of small branches and that allows blood flow to every heart cell. If plaque occurs over time, heart loses its own oxygen – rich blood supply which means, angina or even heart attack can occur. From that reasons, this simple disease can make so much of trouble or even death. Main symptom of CHD is strong pain in the chest but unfortunately CHD can also be a silent killer, which comes unexpected. Common CHD consequence is arrhythmia which is also dangerous to heart, especially in early fazes of developing heart attack but also after it. [2] To prevent that, we struggle with understanding exact causes, trying to predict disease based on historical data.

Main goal of this study is, besides helping doctors reading results, implementation in particular smart device that can process in real time all mayor factors that can and would lead to any sort of heart disease, especially to the Coronary Heart Disease. Real time reading data will process on FPGA device and obtain all results compared to training. According to training results and given real time results, it’ll be possible to diagnose state of patient and prevent any additional problems and health issues. Measuring blood pressure, systolic and diastolic BP, pulse rate and knowing other not fast changing or never changing factors as overweight, age, gender and cholesterol, also using as a part physical condition – is body on move or at state of rest, we should be able to determine possibility of Coronary heart disease.

At this time, we are not able to conclude nature of superior algorithm suitable to give best result but till the end of year, we would determine best way to obtain willing result and integration in device.

# Literature review

First published study was in 1957, 9 years after foundation of Framingham Heart Study and taking first participants. First accomplishment was finding that a nearly 4 times increase in coronary heart disease per 1000 persons occurred with the presence of high blood pressure >= 160/95 mmHg. [3] Beside these early achievements, people still believed that a suitable systolic blood pressure can be calculated as SBP = 100 + (age of participants), in mmHg. [4] In 1971, after 14 years of analyzing data and speaking about blood pressure, investigators and physicians finally demonstrated risk of coronary heart disease. According to previous data, they found stronger link between coronary heart disease and systolic pressure which leaves diastolic pressure innocuous. Later, in two other studies, they found that elevated systolic blood pressure was a predictor of heart failure. [5][6]

By the early 20th century through autopsy, cholesterol was linked to cardiovascular disease. Ancel Keys found high percentage of cholesterol among coronary heart disease patients. [7] In 1997, Framingham investigators reported relationship between HDL and LDL. [8] Also, in that year, investigators from other studies in United States reported that individuals with coronary disease had lower HDL in their bodies then healthy participants. Since that, cholesterol can be considered as a huge risk factor of coronary heart disease.

Also, huge part in metabolic risk factors is taken by diabetes. Lot of effort has been invested in finding a link between diabetes and coronary heart disease. First study was carried out from autopsies. Couple years later, considering clinical data and also years later Framingham, link was finally found. Just looking in column with diabetes, cardiovascular mortality was 3-fold higher for patients with diabetes then patients without. [9] Diabetes was nearly connected with higher risks for heart failure and considered as a one of the most significant factors. [10]

According to previous years of researching, most of risk factors have been found. Not only hypertension, cholesterol and disturbed level of sugar in blood, but also obesity and cigarette consumption can lead to coronary heart disease. Obesity continues to be a growing issue, especially in United States which raises chances of CHD and even heart attack. Framingham participants less than 50 years had 2-3 times higher chances of heart failure. [11] That was concluded with paper from updated Framingham, published by Kenchaiah in 2002. As we already mentioned, tobacco smoking is also one of the causes that leads to CHD. Content of tobacco has everything but no positive effect. It leads to acute increase in blood pressure and coronary vascular resistance. Basically, it clogs vein and coronary arteries and reduces oxygen-rich blood delivery to heart or the whole system. Physician strongly advice stop smoking to prevent any kind of heart disease. [12]

Talking about prediction of different kind heart diseases, machine learning algorithms have huge cut of pie in this whole story. Each data mining technique serves a different purpose. In Multi Model Data Mining Approach for Heart Failure Prediction [13] was discussion about is patient going to live or die. Data was separated as 75% training data and 25% test data. In this paper, 25 model were trained and compare with proposed multi model. Best model was Regression with predictive efficiency of 82.42%. Compared to multi model predictors efficiency with 85.87%, it has slightly lower efficiency. Results show that the proposed model is able to provide better accuracy than best model approach.

According to Chronic Disease Risk Monitoring Based on an Innovative Predictive Modelling Framework [14], published in 2017. K-Nearest neighbor and random forest learning methods turn out to be best performers in terms of activity classification accuracy and confusion matrix but not the best option in execution speed. Another two algorithms, Logistic egression and Decision tree were solid in bot fields. In fact, final prediction value will be passed from FACT machine that uses facts related to risk of heart diseases. For an example, if a person of age above 50 years has increased BMI (more than 29) and has high BP and/or HC, that person is at risk of heart disease. This framework can give better results about prediction heart disease risk.

Using only 10 parameters, in Development of Health Parameter Model for Risk Prediction of CHD Using SVM. [15], data was randomly divided into two subsets – training and test data. Training data consisted approximately 80% and test data around 20% (510 participants). The confusion matrices show that the number of false positives when prediction was performed using SVM was 57, and using Logistic regression analysis was 68

In Intelligent Heart Disease Prediction System Using Data Mining Techniques [16], published 2008, we were shown that Neural Network with 15 attributes has the highest accuracy, round 100%. Decision Tree with 15 attributes also had good performance with 99.62% accuracy. Luckily in combination with Genetic Algorithm and 6 attributes accuracy has fallen just for 0.42%.

In Data Mining Approach to Detect Heart Diseases [17], different classifiers are used and studied for optimum classifier for predicting heart disease. Naïve Bayes is one of the best performing data mining techniques used in diagnosis and prediction heart disease patients. Naïve Bayes classifiers have works well in complex real world and real time situations. In theory, this classifier has minimum error rate but it may not be case always. Going through analysis with Naïve Bayes highest accuracy accomplished is 85.03% and the lowest is 82.31%. Since not only accuracy plays role in this analysis, duration time of building model is significant factor. Speaking about naïve Bayes, time needed to build model is barely 0.02 second.

Summarizing collected info about predicting heart disease of patients, couple of algorithms outperformed but also some of them totally failed. In combination with genetic algorithm data scope is reduced to optimal subset of attributes slightly reducing accuracy for maximum 5%.

Leaving that on side, today, 1,300 scientific papers later, the risk factors behind cardiovascular disease – especially dietary intake – are common knowledge, thanks to the Framingham study. Not only scientists and doctors are up for this study by the time, but most regular people that prefer and practice healthy life.

# WORK WITH DATASET

Our dataset, Framingham Heart Study dataset contains 4240 rows which means 4240 instances. There are 15 attributes and one single class – patient has or does not have CHD. Dataset is separated in 4 main groups as demographic risk factors, risk factors associated with smoking, medical history risk factors and risk factors from the physical examination of the patient. Attributes are presented in nominal and numeric values. For an example gender, smoking habits, BP medicaments, did patient have a stroke or was hypertensive or have diabetes are nominal type and rest of them like age of patient, cigarettes per day, cholesterol, systolic and diastolic blood pressure, body mass index, heart rate and glucose level are numeric type. Since, there are rows with unknown values, we had to initialize them with “our will” values. Education attribute was displayed in 4 values: 1 is coded for high school, 2 for high school diploma, 3 for some college and 4 for a college degree. Dataset had missing values, so certain light preprocessing was needed. There were 105 cells with unknown values, so instead of cut them off we gave them values of 0 which represents “no idea of education” value. It is integer and can be processed. Not only education attribute, but cigarettes per day, BP medicaments, cholesterol, body mass index, heart rate, and glucose level also had missing values. These instances were removed from the database. Now, dataset is shortened by 28 rows due to Na values in cigarettes per day attribute, 52 rows due to Na in BP medicaments, 48 due to Na in cholesterol level, 17 due to Na in body mass index, 1 in heart rate attribute and 338 in glucose level. Going that way, now, our dataset contains 3769 fully optimized instances.

Class TenYearCHD , as we said, has only two values, true – has CHD and false does not have CHD. In this particular dataset, there is 3179 false values and 572 true values which represents really bad balanced dataset. In past time, people were dealing with imbalanced dataset such as fraudulent telephone calls (1996), telecommunication management (1996) etc. The performance of MLAs is typically evaluated using predictive accuracy. However, this is not the best way if the dataset is imbalanced – as it is our case. Nature of our application requires high rate of detection sick people and small error in majority class. History was teaching us to try many ways even if it seems impossible in that moment. So, people were trying to balance dataset properly. One way is to resample dataset, either by over-sampling the minority class or under-sampling the majority class. Here, we will try another way which means that, we are going to under-sample majority class and over-sample minority class on some special form, crate artificial samples that will guide us to better performance of algorithm giving us better prediction accuracy of whether patient will have CHD in ten years or not. Performance is evaluated by a confusion matrix that gives us dependency of predicted negative, predicted positive, actual negative and actual positive classifiers. So far research’s discussed over-sampling with replacement and has noted that it isn’t really improve minority class recognition. That lead us to consider new ways of creating synthetic artificial examples rather than by over-sampling with replacement.   
  
 SMOTE: Synthetic Minority Over-sampling Technique is proposed for creating synthetic examples. This approach is inspired by a technique that proved successful in handwritten character recognition in 1997. They actually created extra training samples by performing certain operations on real data they had. Here, we are going to create synthetic artificial samples and operate in feature space rather than in data space. Class that represents sick people is over-sampled by taking each minority class sample and introducing synthetic samples along nearest neighbors. Depending upon the number of over-sampled data, neighbors from the *k*-nearest neighbors are randomly chosen. Synthetic samples are generated as it follows: Just take two feature vectors (samples), one under consideration and its nearest neighbor. After that, multiply their difference by a random generated number between zero and one, and add it to the chosen sample. This will result with selection of a random point along the line between two features, chosen one and its nearest neighbor. Pseudo code of SMOTE algorithm was described in [18].

# RESULTS

Achieved results without SMOTE are not promising. Best performed algorithms gave us average total success percentage around 82%, but really low FP rate for TRUE class instance. Sensitivity of false results which can lead to the patient death in this situation is highly recommended to avoid. Problem is not correctly classified instances but low ratio of correctly classified patients that would have Chronical Heart Disease in next ten years. Let’s separate a few best of them. On first place is BayesNet with percentage of CCI of 80.43% and TP rate of 0.329 for true class instance. After BayesNet are following NaiveBayes and NaiveBayesUpdateable with TP rate of 0.280, RandomTree with TP of 0.276 and RandomizableFilteredClassifier with TP rate of 0.227. Also, it’s important to mention that we used 20 Cross-validation folds which slightly, but very significantly, raised the TP rate. After many attempts, results still were unusable and that was the point where SMOTE found its role in this CHD with FHS story.

We were able to see, the performance of machine learning algorithms greatly increased when we included SMOTE. One of the outputs, we also have, the Receiver Operating Characteristic Area which is between 0 and 1, closer to the 1, better. The Receiver Operating Characteristic (ROC) curve is a standard technique for summarizing classifier performance over a range of tradeoffs between true positive and false positive error rates (Sweets, 1988). The Area under Curve (AUC) is an accepted traditional performance metric for a ROC curve (Hart & Stork 2001). The ROC convex hull can also be used as a robust method of identifying potentially optimal classifiers. (Provost & Fawcett, 2001). That means if a line passes through a point on the convex hull then this is the line with the largest (TP) intercept.

As we mentioned in last part, total success was quite good, but TP rate of sick patients was low. Since, goal is to correctly classify sick people and save their lives, that achievement was disappointing. After that, we decide to play a little bit with SMOTE and synthetic samples. For the first time using SMOTE, our dataset was percently separated in two groups, one that contains 90% of data and the other one that contains rest of it. In groups, data was separated equally. That means, in first group we have 90% of healthy and 90% of unhealthy patients. On that way formed set, we applied SMOTE on training set and tested on the rest of set (last 10% of data). Number of correctly classified instances increased for around 1%. That would be great if TP rate of unhealthy people was not 0.03, which was horrible. Conclusion, urgently need to change approach.  
  
Later on, we finally got great achievement. Not only total number of correctly classified instances increased but also individual class classification of healthy and unhealthy patients made huge progress. Finally, we balanced it and the best thing was, both TP rate were over 0.9. Decision was to make completely new training dataset based on synthetic artificial samples and test it on data from Framingham Heart Study dataset. First experiments were based on the equalization of class samples. As we mentioned before, best performing algorithms in that stage were BayesNet. Now, with that kind of samples, results are not to boast. But, Random Forest and KStar performed well and turned out to be good candidates for further consideration. The individual use of algorithms has yielded good results in the classification of one class, whether of sick or healthy patients. Random Forest showed as good candidate for classification of healthy people. At the other hand, lazy KStar gave better results of classification unhealthy patients. Combining them together, individual TP rate of healthy and unhealthy patients slightly decreased.

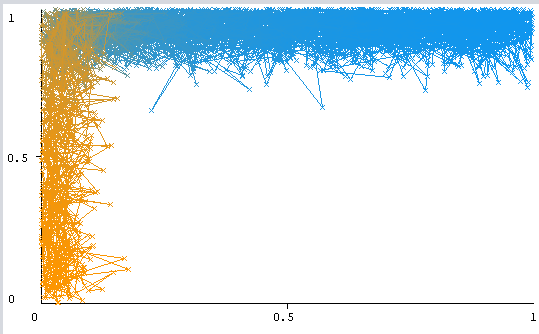
Applying SMOTE on our dataset, more precisely creating 400% more samples of unhealthy people, and after that 290% on unhealthy people, gives dataset of 12242 patients that do have CHD and 12554 patients that do not have CHD. Dataset was after that resampled at its 50% and finally we have 6277 instances of healthy and 6121 of people with CHD. Synthetic dataset was completely jackpot for Framingham dataset, and supplying Framingham dataset to the synthetic training data, gives more than good results. Random Forest as individual algorithm classified unhealthy patients with TP rate of 0.954 and total correctly classified instances of 94.00 %. Lazy KStar classified healthy patients better than unhealthy with TP rate of 0.991 for healthy patients and total correctly classified instances of 95.26%. Combined, their results are bit better since their TP rates are almost the same: 0.955 for unhealthy and 0.976 for healthy patients. Total number of correctly classified instances in this particular case is 95.84% which brings this combination to the top. Root mean squared error is 0.1874 and relative absolute error is 23.81%. All data described here will be given in table representation below:

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest | | | |
| Summary | TP Rate | ROC Area | Total Accuracy |
| Class True | 0.861 | 0.969 | 94.00 % |
| Class False | 0.954 | 0.969 |

|  |  |  |  |
| --- | --- | --- | --- |
| KStar | | | |
| Summary | TP Rate | ROC Area | Total Accuracy |
| Class True | 0.991 | 0.997 | 95.26% |
| Class False | 0.946 | 0.997 |

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest and Lazy KStar | | | |
| Summary | TP Rate | ROC Area | Total Accuracy |
| Class True | 0.976 | 0.995 | 95.84% |
| Class False | 0.955 | 0.995 |

This way, combining Lazy KStar and Random Forest algorithm, we got really useful results. Area under ROC (=0.995) is graphically represented below :



# CONCLUSION

Finally, at the end, we can say goal is more than accomplished. Good training dataset gives as opportunity to train well our algorithms and have pattern to recognize, classify and decide medical treatments. With almost 96% of successfully classified patients, we can help doctor to decide or even manage further actions on patient treatments.

# REFERENCES

1. National Heart, Lung and blood institute
2. Textbook of Medical Physiology, Arthur C. Guyton and John E. Hall
3. Dawber TR, Moore FE, Mann GV. Coronary heart disease in the Framingham study. Am J Public Health Nations Health. 1957;47:4–24
4. Kannel WB. Fifty years of Framingham Study contributions to understanding hypertension. J Hum Hypertens. 2000;14:83–90.
5. Kannel WB, Wolf PA, Verter J, McNamara PM. Epidemiologic assessment of the role of blood pressure in stroke. The Framingham study. JAMA. 1970;214:301–10.
6. Kannel WB, Castelli WP, McNamara PM, McKee PA, Feinleib M. Role of blood pressure in the development of congestive heart failure. The Framingham study. N Engl J Med. 1972;287:781–7.
7. Keys A, Fidanza F. Serum cholesterol and relative body weight of coronary patients in different populations. Circulation. 1960;22:1091–106
8. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. High density lipoprotein as a protective factor against coronary heart disease: The Framingham study. The American Journal of Medicine. 1977;62:707–14
9. Kannel W, McGee D. Diabetes and cardiovascular risk factors: the Framingham Study. Circulation
10. Partamian JO, Bradley RF. Acute myocardial infarction in 258 cases of diabetes. Immediate mortality and five-year survival. N Engl J Med. 1965;273:455–61
11. Hubert HB, Feinleib M, McNamara PM, Castelli WP. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. Circulation
12. Cigarette smoking and coronary heart disease : risks and management. Cardiol Clin 1996
13. Multi Model Data Mining Approach for Heart Failure Prediction ; Priyanka HU and Vivek R, 2016
14. Chronic Disease Risk Monitoring Based on an Innovative Predictive Modelling Framework ; Nitten S. Rajliwall Girija Chetty, Rachel Davey, 2017
15. Intelligent Heart Disease Prediction System Using Data Mining Techniques, Sellappan Palaniappan, Rafiah Awang 2008 IEEE
16. Data Mining Approach to Detect Heart Diseases; Vikas Chaurasia Saurabh Pal
17. SMOTE: Synthetic Minority Over-sampling Technique, Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer