How do different parameters affect the detection of coronary artery disease?

Daniel Arad
Ben Gurion University,
Department of Biomedical Engineering
Beer Sheva, Israel
araddani@post.bgu.ac.il

Gil Akler
Ben Gurion University,
Department of Biomedical Engineering
Beer Sheva, Israel
aklergil@post.bgu.ac.il

***Abstract* — coronary artery disease (CAD) is the most common cause of death in the general population. For many people, the first clue that they have CAD is a heart attack that can lead to cardiac arrest and even death. Therefore, the project's motivation is to find the most significant features that can lead to early detection of CAD using four algorithms of supervised learning (SVC, Random Forest, Gradient Boosting Classifier and Neural Networks classifier). The Random Forest algorithm yielded the best accuracy and sensitivity in any model with all the features. Therefore, we can conclude that this is the best algorithm approach for detecting CAD. Moreover, we found out that the most significant features are ST Slope, exercise angina and chest pain. For early detection of CAD, we saw that without the features exercise angina and chest pain we got results of less than 90% accuracy and sensitivity.**

***Key words* – Heart Disease, Coronary artery disease, heart disease dataset, machine learning, supervised learning.**

I. INTRODUCTION

Coronary Artery Disease (CAD) is the most common cause of death in the general population and specifically in the United States[1]. CAD is caused by plaque buildup in the walls of the arteries that supply blood to the heart (called coronary arteries) and other parts of the body. Plaque is made up of deposits of cholesterol and other substances in the artery. Plaque buildup causes the inside of the arteries to narrow over time, which can partially or totally block the blood flow [2]. This process is called atherosclerosis.
The decline in age-standardized mortality rates and in incidence of CAD in many countries illustrates the potential for prevention of premature deaths and for prolonging life expectancy.[3]

Most symptoms of CAD may go unrecognized at first, or they may only occur when the heart is beating hard like during exercise. As the plaque builds up in the coronary arteries and they continue to narrow, less and less blood gets to the heart and symptoms can become more severe or frequent. For many people, the first clue that they have CAD is a heart attack that can lead to cardiac arrest and even death. Many risk factors are linked to CAD and include age, sex, diet, smoking, cholesterol level and even other medical conditions. Therefore, the motivation of this study is to find the most significant risk factors and symptoms that can diagnose CAD prior to a lethal heart attack. [3]

The dataset used in this study is from a previous study and was taken from the IEEE data port website [4]. The dataset contains 11 features on 1190 patients. The features include numeric, nominal, and binary data classified by medical professionals as healthy and sick with CAD. The features include – age, sex, chest pain type, resting blood pressure, serum cholesterol levels, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise angina, old peak=ST and the slope of the peak at exercise – ST segment.

In this work we plan to assess the different features using different machine learning methods to find the most significant features that can lead to early detection of CAD.

II. METHODS

The first step of this study was to go over all the data and clean it. The data set was mostly cleaned and we only needed to impute some missing data points with the avg values of each feature. Since the data was previously classified by medical professionals into healthy and sick patients, we used supervised learning.

The second step of this project was to statistically analyze the data. Numeric data was tested using a t-test between the healthy and sick groups of each numeric feature. We also created a correlation matrix of all the features in the dataset.

The third step of this study was to split the data into test and train sets for evaluations. To avoid over-fitting, which is an undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for new data, we decided to use a function called cross validation [5]. In cross validation the data is split into k-folds, where one of the folds is marked as the validation set, and the other k-1 folds are marked as the training sets. Training proceeds on the training set, after which evaluation is done on the validation set. The training and validation sets are switched k times to find the mean accuracy of all the data folds. The final evaluation can be done on the test set.

The fourth step of this project was hyper-parameters tuning. In this step we create different supervised learning algorithm and change the parameters to evaluate the data. Our first analysis of the data was done in all the features. The first approach was to use linear SVM. In order to approve the results, we later decided to change the

polynomial degree of the data using the polynomial space. We decided to also test this data using the Random Forest algorithm.

The fifth step of this study was to change the number of features tested each time and randomly select them out of all the features in the dataset. For each set of features selected we calculated the chi-2 value and f-value. In each set of features, we ran four supervised learning models (SVC, Random Forest, Gradient Boosting Classifier and Neural Networks known as Multi-layer Perceptron classifier), to find the most significant features that can lead to early detection of CAD. We also used a function called feature selection. Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. This test was done using an ANOVA to compute the f-value – the highest score are the most significant features.

The last step of this project was to analyze the results in order to determine the best features which can lead to early detection of CAD. We measure the success of this project by measuring the sensitivity. We prefer to have more healthy patients misdiagnosed with CAD, than to misdiagnose sick patients as healthy – therefore we prefer to have a high true positive rate.

III. RESULTS

For the second part of this project, we got that the results of the t-test are all with a p-value < 0.05 between the healthy and sick group of each numeric feature. We can also see in figure 1 (appendix section) that the features with the highest correlation to the target are ST Slope, exercise angina and chest pain.

For the third and fourth step of this project we first create the cross-validation function to split the data into 5-folds of train and test sets. We ran the classification learner using linear SVC. We can see the results in table I (appendix section) that the accuracy was 82.4% and the sensitivity was 82.5%. Since the results had a low detection rate compared to previous studies (where the accuracy was around 93%)[6] we decided to change the polynomial degree of the data. As the degree of the data was higher the accuracy and sensitivity were also higher, but not as high as previous studies.

Our next algorithm was random forest, using different parameters such as criteria (Gini or Entropy) and number of estimators in the forest. The results here exceeded our expectations as we got an accuracy of above 90%, see table II. The highest accuracy was 93.5% using a criterion of Gini and 300 estimators. Using this specific forest, we also got a high sensitivity of 94.3%.

The fifth step of this study was to change the number of features tested each time and randomly select them out of all the features in the dataset. For each set of features selected we calculated the chi-2 value and f-value. In each set of features, we ran four supervised learning models mentioned in the method section. As we can see in table III the random forest algorithm got the 7 highest accuracies out of all the different combinations for detecting CAD. The highest accuracy was 93.1% using all the features and using chi-2 feature selection and the sensitivity was 93.5%. The second algorithm was Gradient Boosting Classifier that got the highest results after the random forest boosting with accuracy of 88% to all the features.

The last step was using the feature selection to determine the most significant features. We can see in table IV, that these features were ST Slope, exercise angina and chest pain.

IV. DISCUSSION, CONCLUSIONS AND FUTURE WORK

We can see from our initial statical analysis of the numeric features that since we got a p-value<0.05 we can conclude that there is a significant difference between the healthy and sick groups of those features. We can also see from our correlation matrix that the highest correlation to the target were ST Slope, exercise angina and chest pain. We can see that the features of chest pain and exercise angina are symptoms of early heart attack – therefore they are not features that would help early detection of CAD.

We saw that early detection is possible with high accuracy results (93.5%) when using all the features in the dataset using the random forest algorithm. The Random Forest algorithm yielded the best accuracy and sensitivity in any model – therefore we can conclude that this is the best algorithm approach for detecting CAD.

When trying to classify the most significant features for early detection of CAD, we saw that without the features exercise angina and chest pain we got results of less than 90% accuracy and sensitivity. These results are not up to the standards of previous studies and our research question.

We also saw that the results of the feature selection function matched our original hypothesis established by the early statistical analysis we did on the dataset. Both the feature selection algorithm and the correlation matrix had the ST Slope, exercise angina and chest pain features as the most significant for CAD detection.

For future work we believe that the features of exercise angina and chest pain need to be removed – as they are symptoms of heart attack. Instead of these features add fatigue, smoking and shortness of breath features.

ACKNOWLEDGMENT

We would like to thank Manu Siddhartha from Liverpool John Moore's University for tagging this dataset into healthy and sick patients. The dataset was taken from the website of IEEE data port.

REFERENCES

[1] P. A. McCullough, "Coronary Artery Disease," *Clinical Journal of the American Society of Nephrology*, vol. 2, no. 3, 2007, [Online]. Available: https://journals.lww.com/cjasn/Fulltext/2007/05000/Coronary_Artery_Disease.30.aspx

[2] J.-C. Tardif, "Coronary artery disease in 2010," *European Heart Journal Supplements*, vol. 12, no. suppl_C, pp. C2–C10, Aug. 2010, doi: 10.1093/eurheartj/suq014.

[3] P. T. Costa, "Influence of the normal personality dimension of neuroticism on chest pain symptoms and coronary artery disease," *The American Journal of Cardiology*, vol. 60, no. 18, pp. J20–J26, 1987, doi: https://doi.org/10.1016/0002-9149(87)90679-5.

[4] M. Siddhartha, "Heart Disease Dataset." IEEE DataPort, Nov. 05, 2020. doi: 10.21227/DZ4T-CM36.

[5] M. L. Schwabbauer, "Use of the latent image technique to develop and evaluate problem-solving skills," *Am J Med Technol*, vol. 41, no. 12, pp. 457–462, Dec. 1975.

[6] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution.," *Future Sci OA*, vol. 7, no. 6, p. FSO698, Mar. 2021, doi: 10.2144/fsoa-2020-0206.

DIVISION OF WORK

All the project was done by both of us together. We sat together on both algorithms development and writing the paper.

APPENDIX

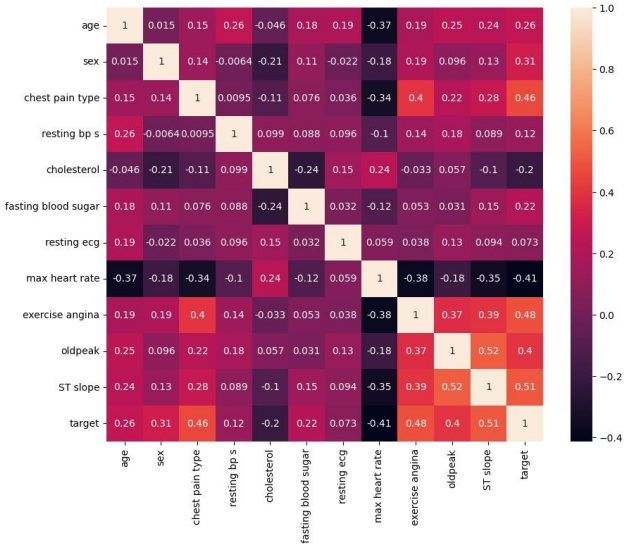

FIGURE I CORRELATION MATRIX

TABLE I SVC RESULTS

Degree	Mean Accuracy	Std Accuracy	Precision	Specificity	Sensitivity
4	0.852941	0.023916	0.873355	0.862745	0.844197
3	0.830252	0.036591	0.856427	0.846702	0.81558
2	0.828571	0.03963	0.854758	0.84492	0.81399
1	0.82437	0.034587	0.839806	0.823529	0.825119

TABLE III RANDOM FOREST RESULTS

mean ACC	*std ACC*	*estimators*	*criterion*	*Precis.*	*SP*	*SE*
0.935	0.035	300	*gini*	0.926	0.916	0.942
0.934	0.027	200	*entropy*	0.931	0.921	0.944
0.932	0.031	300	*entropy*	0.930	0.921	0.941
0.931	0.034	400	*gini*	0.923	0.912	0.942
0.931	0.036	200	*gini*	0.920	0.909	0.937
0.930	0.033	500	*gini*	0.925	0.914	0.942
0.929	0.033	400	*entropy*	0.929	0.919	0.937
0.928	0.032	100	*entropy*	0.929	0.919	0.936
0.928	0.034	500	*entropy*	0.925	0.914	0.942
0.927	0.034	100	*gini*	0.930	0.921	0.936

TABLE III FULL RESULTS – FIFTH STEP

k	Feature selection	classifier	Mean ACC	Std ACC	Prec.	SP	SE
11	*chi2*	RFC	0.931	0.032	0.930	0.921	0.934
11	*f_classif*	RFC	0.928	0.035	0.918	0.907	0.936
9	*f_classif*	RFC	0.921	0.036	0.925	0.916	0.931
9	*chi2*	RFC	0.919	0.041	0.923	0.912	0.939
7	*f_classif*	RFC	0.912	0.049	0.923	0.914	0.915
7	*chi2*	RFC	0.902	0.049	0.913	0.903	0.899
5	*f_classif*	RFC	0.889	0.045	0.893	0.882	0.883
11	*chi2*	GBC	0.880	0.022	0.887	0.873	0.887
11	*f_classif*	GBC	0.880	0.022	0.887	0.873	0.888
9	*chi2*	GBC	0.873	0.024	0.884	0.871	0.875
9	*f_classif*	GBC	0.873	0.024	0.884	0.871	0.875
7	*f_classif*	GBC	0.867	0.025	0.878	0.864	0.869
7	*chi2*	GBC	0.860	0.028	0.863	0.844	0.874
5	*f_classif*	GBC	0.854	0.026	0.857	0.837	0.869
5	*chi2*	GBC	0.851	0.048	0.847	0.823	0.875
5	*chi2*	RFC	0.841	0.048	0.841	0.816	0.866
5	*chi2*	MLPC	0.838	0.046	0.845	0.825	0.850
7	*chi2*	SVC	0.838	0.042	0.835	0.809	0.864
11	*chi2*	SVC	0.837	0.030	0.851	0.836	0.839
11	*f_classif*	SVC	0.837	0.030	0.851	0.836	0.839

TABLE IV FEATURE SELECTION RESULTS

Feature	Score
ST slope	408.001
exercise angina	358.493
chest pain type	319.073
max heart rate	244.704
oldpeak	224.118
sex	127.450
age	87.580
fasting blood sugar	58.533
cholesterol	48.662
resting bp s	17.775
resting ecg	6.375

LINK GIT

HTTPS://GITHUB.COM/GILAKLER1/CORONARYHEARTDISEASES