Implementation of various classification techniques for the prediction of Coronary Artery Disease

Amit Mishra[1], Ritwika Gautam[1], Varun Sapra[1]
[1]*University of Petroleum and Energy Studies, Dehradun*
(E-mail: amit581h@gmail.com, ritwikagautam@gmai.com, varun.sapra@gmail.com)

***Abstract*—** Abstract—Cardiovascular diseases (CVD) are escalating at an alarming rate. The numerous deaths caused by coronary artery disease (CAD) have disconcerted a lot of people. Thus, making the diagnosis and treatment of CAD vital. This paper compares different classification algorithms like Decision Trees, Random Forest and Logistic Regression, which are used to classify CAD and non-CAD patients from a given dataset. These classification algorithms will be applied to a small subset of features selected from the existing features in the original dataset, using a meta-heuristic correlative feature subset selection algorithm Cuckoo search. The results obtained will be compared to the result obtained from applying same classification algorithms on the dataset with all features. Logistic Regression clearly outperforms other methods with an accuracy of 89.01%.

Keywords— *Coronary Artery Disease, Decision Trees, Random Forest, Logistic, Regression*

I. INTRODUCTION

Every year 31% of all global deaths, that is 17.9 million deaths are caused due to cardiovascular diseases (CVD). Unhealthy lifestyle involving unbalanced diet, tobacco consumption, lack of physical activities, alcohol abuse trigger these diseases [1]. They are manifested primarily as heart attacks or strokes. Timely prognosis of the disease is essential to its treatment. For the prognosis of CVDs and coronary artery disease (CAD), there are both, invasive and non-invasive methods. The invasive methods of diagnosing the extent and severity of CAD (like angiography) are efficient and accurate, but also very painful and costly, not available and accessible to all [2]. It requires great medical proficiency, on behalf of the doctors, to be performed. This motivates the researchers for finding new non-invasive methods. This paper focuses on comparing the results for non-invasive methods using different machine learning classification algorithms.

Several other researchers have also proposed non-invasive methods for early diagnosis of CAD. Pouladian et al (2005) proposed using Arterio-Oscillo-Graphy (AOG) for diagnosing atherosclerosis, a kind of arteriosclerosis that causes coronary artery disease (CAD) in the form of sclerosis (endothelium hardening), stenosis (lumen decrease), and occlusion, depending on the severity of CAD. Given that early atherosclerosis involves the endothelium of many arteries [2], several noninvasive tests try to provide information about the structures and functions of peripheral arteries to detect CAD. Accuracy to detect CAD positively, achieved with AOG, was 73%. Amin et. al (2013) proposed a novel technique binary coordinate ascent based on the coordinate descent algorithm.[3] They considered the risk factors related with heart diseases and used them to predict CAD. They implemented genetic neural network to perform data mining.

The evaluation of the algorithm is carried out on the basis of the area under receiver-operating-characteristic (ROC) curve. The proposed algorithm showed a substantial improvement in terms of efficiency. Yanhua Wu et al (2013) proposed the use of magnetocardiography as a tool for detecting CAD non-invasively. Two dominant parameters--curvature of magnetic field zero line, and the area ratio of the extrema circle have been used. Data of 97 subjects collected by four sensors was arranged in a 2X2 array at 9 different positions. The result shows the sensitivity of 71.4% and specificity of 72% [4].

A machine learning framework was proposed by Chowdhuryy et al (2018). They have proposed the framework for portable ECG module implemented using convolutional network. Their method achieved an accuracy of 92.3% [5]. Dai et al (2015) explained the concept of supervised learning techniques in their study. They did that for the prediction of hospitalization of patients who are suffering from heart related diseases. SVM, Naïve Bayes and logistic regression have been used in the study [6]. Lin et al. (2015) presented a novel technique for reducing feature space. They used PSO and artificial bee colony algorithm (ABC) for the experimental purpose, which enhanced the capabilities of the model to reduce the feature set and the identification of significant features [7]. Another wrapper method for getting optimal brink by iteratively changing the feature subset was proposed by Manikandan, Susi and Abirami (2017) [8]. Verma et al. (2016) used non-invasive clinical attributes for the diagnosis of CAD. They used correlation based feature subset method for reducing the dimensionality and further used PSO search for optimization. Their approach enhanced the accuracy of diagnostic models [900].

This paper presents a comparison between various classification algorithms with and without using feature selection used to classify CAD and non-CAD patients.

II. METHODOLOGY

Data Collection: This is the most crucial part. The data that is being used should be veracious and unbiased. To ensure if the data is genuine data has been taken from the authentic source. In this case, we are using Z-Alizadeh Sani dataset from UCI, originally containing 303 instances and 56 features, including the one to be classified.

Data Preprocessing: The data in its raw form is not much useful. It needs to be refined or handling missing values. This could be achieved by ignoring the tuple or a specific data instance, or ignoring the column entirely. The missing values could also be replaced by arithmetic mean of the column, median of the column, or using a some learning algorithm like K Nearest Neighbour [10]. Remove outliers and smooth noisy data. The dataset could have some outliers or extreme values that does not fit with the rest of the data well, giving inaccurate results. These outliers need to be removed. Feature Selection: One of the major issues faced while working with data is the curse of dimensionality; datasets often have a lot of attributes, or features, that are irrelevant for the Machine Learning algorithms [10]. These superfluous features drastically increase the running time of the algorithm, thus reducing the efficiency of model. If these extra features are not trimmed, the problem of Over fitting may also occur. To rectify this, meta-heuristic algorithms are used for selecting relevant features from the dataset [11].

III. MATERIALS AND METHODS

The Z-Alizadeh Sani Dataset from UCI has been selected for the implementation of the classification model. The original dataset has 303 instances and 56 attributes (including predicted value). The total number of features makes it difficult to apply any classification algorithm. Even if the model is somehow made, it may be prone to over fitting, thus reducing the efficiency of the model. To tackle this, Meta heuristic algorithms for feature selection are used. One such algorithm, Cuckoo search yields 17 features from the dataset of 56 features.

Cuckoo search algorithm:

Objective function f(x), x = (x1, ..., xd)T

Generate initial population of n host nests xi

while (t <MaxGeneration) or (stop criterion)

Get a cuckoo randomly /generate a solution by Levy flights and then evaluate its quality/fitness Fi

Choose a nest among n (say, j) randomly

if (Fi > Fj),

Replace j by the new solution

end

A fraction (pa) of worse nests are abandoned and new ones/solutions are built/generated Keep best solutions (or nests with quality solutions) Rank the solutions and find the current best

end while

Features obtained from cuckoo are as follows:

Age, DM, HTN, CRF, CVA, BP, Typical Chest Pain, Dyspnea, Atypical, Nonanginal, St. Elevation, St Depression, Tin Version, ESR, EF-TTE, Region, RWMA

Table 1: Reduced feature subset

	Age	DM	HTN	CRF	CVA	BP	Typical Chest Pain	Dyspnea
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	58.897690	0.297030	0.590759	0.980198	0.983498	129.554455	0.541254	0.557756
std	10.392278	0.457706	0.492507	0.139550	0.127605	18.938105	0.499120	0.497475
min	30.000000	0.000000	0.000000	0.000000	0.000000	90.000000	0.000000	0.000000
25%	51.000000	0.000000	0.000000	1.000000	1.000000	120.000000	0.000000	0.000000
50%	58.000000	0.000000	1.000000	1.000000	1.000000	130.000000	1.000000	1.000000
75%	66.000000	1.000000	1.000000	1.000000	1.000000	140.000000	1.000000	1.000000
max	86.000000	1.000000	1.000000	1.000000	1.000000	190.000000	1.000000	1.000000

Atypical	Nonanginal	St Elevation	St Depression	Tinversion	ESR	K	EF-TTE	Region RWMA
303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
0.693069	0.947195	0.046205	0.234323	0.297030	19.462046	4.230693	47.231023	0.620462
0.461983	0.224015	0.210275	0.424276	0.457706	15.936475	0.458202	8.927194	1.132531
0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	3.000000	15.000000	0.000000
0.000000	1.000000	0.000000	0.000000	0.000000	9.000000	3.900000	45.000000	0.000000
1.000000	1.000000	0.000000	0.000000	0.000000	15.000000	4.200000	50.000000	0.000000
1.000000	1.000000	0.000000	0.000000	1.000000	26.000000	4.500000	55.000000	1.000000
1.000000	1.000000	1.000000	1.000000	1.000000	90.000000	6.600000	60.000000	4.000000

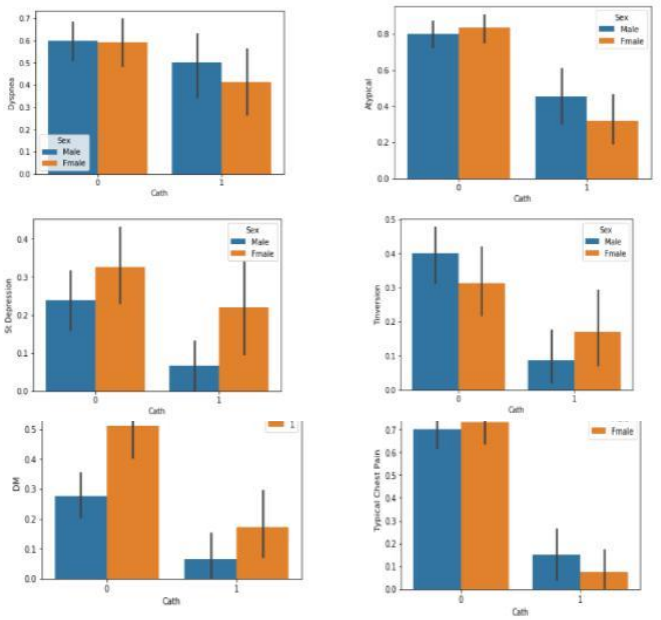

Fig 2: Deep Insight of data visualization for CAD patients with different attributes

For the implementation of the classification algorithms, Scikit-learn, a python based machine learning library is used.

A. DECISION TREE

A decision tree is a flowchart-like structure with internal nodes has a condition, and the response of the input to that condition determines the path forward. The topmost node is the root, following conditions are reached through branches and the leaf node or the terminal node represents the class that has been assigned to that input depending on the classification rules [12]. A decision tree consists of three types of nodes Decision nodes – typically represented by squares Chance nodes – typically represented by circles End nodes – typically represented by triangles. Decision trees are mostly used in decision-making situations when the decisions can be classified into a finite number of classes. They can be diversely used in business, health economics, and public health, and are examples of operations research or management science methods.

B. Logistic Regression

Logistic regression uses the logistic model of statistics. In its basic form, a logistic model uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression estimates the parameters of a logistic model; it is a form of binomial regression. A binary logistic model has two possible outcomes like yes/no, true/false, male/female, or positive/negative; these are represented by the labels "1" and "0". In the logistic model, the logarithm of the odds for the value labeled "1" gives a positive outcome and logarithm of the odds for the value labeled "0" gives a negative outcome [13].

C Random Forest

Random Forest, as the name suggests is a forest, made-up of trees, Decision tree. Random forests are an ensemble learning method for classification and regression problem. They construct a multitude of decision trees at the time of training and take the mode of the classification, or mean for regression, of all those trees to predict the output. This helps to avoid the overfitting problem of decision trees.

IV RESULTS

The results of different methods applied on the dataset are evaluated and the following performance measures are recorded.

Decision Tree

Table2 shows the results of Decision Tree (DT) with accuracy achieved 76.9% and Table 3 shows the confusion matrix for DT.

Table 2: Performance parameters recorded for Decision Tree

Precision	recall	F1-score	Support
0.89	0.80	0.84	70
0.50	0.67	0.57	21
0.80	0.77	0.78	91

Table 3: Confusion matrix

	Positive	Negative
Positive	56	14
Negative	7	14

Accuracy Score: 0.7692307692307693

Logistic regression

Table 4 and Table 5 show the recorded performance measures of logistic regression, and confusion matrix respectively.

Table 4: Performance attributes recorded for Logistic Regression

Precision	recall	F1-score	Support
0.94	0.91	0.93	70
0.74	0.81	0.77	21
0.89	0.89	0.89	91

Table 5: Confusion matrix

	Positive	Negative
Positive	64	6
Negative	4	17

Accuracy Score: 0.8901098901098901

Random Forest

Table6 and Table7 show the recorded performance measures of Random Forest and confusion matrix respectively.

Table 6: Performance measures for Random Forest

Precision	recall	F1-score	Support
0.89	0.91	0.90	70
0.68	0.62	0.65	21
0.84	0.85	0.84	91

Confusion matrix

	Positive	Negative
Positive	64	6
Negative	8	13

Accuracy Score: 0.8461538461538461

The same methods have been applied to the complete dataset and the recorded performance measures clearly

indicated the increase in accuracy with reduced feature subset. Table8 describes the summary of all three classification methods on the complete data set and Table9 indicates the performance measures of all three methods with reduced features.

Authors	Input data	Mean se	Mean sp
Pal et al. [11]	Patient metadata	0.94	0.39
Banerjee et al. [12]	PPG	0.62	0.82
Schmidt et al [13]	PCG	0.47	0.82
Banerjee et al. [14]	PCG	0.79	0.77
Choudhary et al [15]	PPG, PCG	0.82	0.83

Algorithm	precision	recall	F1score	accuracy	MSE
DT	0.75	0.76	0.75	0.758	0.241
RF	0.80	0.79	0.77	0.791	0.208
LR	0.79	0.79	0.79	0.791	0.208

DISCUSSION

In this paper we have compared different classification algorithms like Decision Trees, Random Forest and Logistic Regression, to classify CAD and non-CAD patients from a given dataset. The dataset was reduced using cuckoo search method and the reduced dataset gives better results as compared to the complete dataset. The study proved that logistic regression yields the best results for the prediction of coronary artery disease

REFERENCES

[1] W. H. Organisation et al., "Cardiovascular Disease (cvds):Fact sheet no.317, 2011" 2011

[2] Pouladian, M., Golpayegani, M., Abbaspour Tehrani-Fard, A. and Bubvay-Nejad, M. (2005). Noninvasive Detection of Coronary Artery Disease by Arterio-Oscillo-Graphy. IEEE Transactions on Biomedical Engineering, 52(4), pp.743-747

[3] Amin, S. U., Agarwal, K., & Beg, R. (2013). Genetic neural network based data mining in prediction of heart disease using risk factors. In Information Communication Technologies (ICT), 2013 IEEE Conference on (pp. 1227-1231).

[4] Wu, Y., Gu, J., Chen, T., Wang, W., Jiang, S., & Quan, W. (2013, July). Noninvasive diagnosis of coronary artery disease using two parameters extracted in an extrema circle of magnetocardiogram. In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE (pp. 1843-1846). IEEE.

[5] Chowdhuryy, M. H. I., Sultana, M., Ghosh, R., Ahamed, J. U., Mahmood, M. A. I. (2018, February). AI Assisted Portable ECG for Fast and Patient Specific Diagnosis. In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

[6] Dai, W., Brisimi, T. S., Adams, W. G., Mela, T., Saligrama, V., Paschalidis, I. C. (2015). Prediction of hospitalization due to heart diseases by supervised learning methods. International journal of medical informatics, 84(3), 189-197.

[7] Lin, K. C., Hsieh, Y. H. (2015). Classification of medical datasets using SVMs with hybrid evolutionary algorithms based on endocrine-based particle swarm optimization and artificial bee colony algorithms. Journal of medical systems, 39(10), 119.

[8] Manikandan, G., Susi, E., Abirami, S. (2017, February). Feature Selection on High Dimensional Data. Using Wrapper Based Subset Selection. In Recent Trends and Challenges in Computational Models (ICRTCCM), 2017 Second International Conference on (pp. 320-325). IEEE.

[9] Verma, L., Srivastava, S., Negi, P. C. (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. Journal of medical systems, 40(7), 178.

[10] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer methods and programs in biomedicine, 141, pp.19-26.

[11] D. Pal, K. Mandana, S. Pal, D. Sarkar, C. Chakraborty, Fuzzy expert system approach for coronary artery disease screening using clinical parameters, Knowledge Based Systems, vol. 36, no. Supplement C, pp. 162-174, 2012

[12] R. Banerjee, R. Vempada, K. M. Mandana, A. D. Choudhury, A. Pal, Identifying coronary artery disease from photoplethysmogram, Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pp. 1084-1088, 2016.

[13] S. E. Schmidt, J. Hansen, H. Zimmermann, D. Hammershoi, E. Toft, J. J. Struijk,;Coronary artery disease and low frequency heart sound signatures, Computing in Cardiology 2011. IEEE, pp. 481-484, 2011.

[14]R. Banerjee, A. D. Choudhury, P. Deshpande, S. Bhattacharya, A. Pal, K. Mandana, A robust dataset-agnostic heart disease classifier from phonocardiogram, Engineering in Medicine and Biology Society (EMBC) 2017 39th Annual International Conference of the IEEE, pp. 4582-4585, 2017.

[15] A. D. Choudhury, R. Banerjee, A. Pal, K. M. Mandana, "A fusion approach for non- invasive detection of coronary artery disease", Proceedings of the 2017 EAI International Conference on Pervasive Computing Technologies for Healthcare, 2017.