A HYBRID CARDIOVASCULAR DISEASE DIAGNOSIS AND PREDICTION SYSTEM USING MACHINE LEARNING APPROACH

Shakeel Juman TP
Faculty in Computer Science and Application

ABSTRACT:

The biggest cause of deaths worldwide is the cardiovascular disease and nowadays, its prediction at an early stage is of great importance. In this paper, Cardiovascular disease prediction is done by adopting Supervised Learning Algorithms using the patient's medical record and the comparison of results are done with the known supervised classifiers Decision tree, Random forest, Logical regression,K nearest neighbor classifier and Naïve bayes. Classifier is used to classify the patient record information. To determine the risk of cardiovascular disease,attributes are given as input to the classifier in the classification stage 14. The Physicians can diagnose the disease in a more effective way using this proposed system. The record collected from 303 patients is used to test the efficiency of the classifier. The results shows that the prediction of the likelihood of cardiovascular patients can be done using various classifiers in the most efficient way.

KEYWORDS:

Cardiovascular, dataset, feature extraction, decision tree, random forest, logical regression, k nearest neighbor classifier and Naive bayes.

I. INTRODUCTION

One of the most critical human disease in the world is Cardiovascular disease and it has a very severe impact on human life. Heart becomes insufficient to push the required amount of blood to other parts of the body in Cardiovascular disease. Heart failure prevention and treatment can be carried out only by an accurate, on-time diagnosis and prediction at an early stage of the disease. Traditional medical history has been considered to be not reliable in many aspects for the diagnosis and prediction of cardiovascular disease. Non-invasive based methods such as

machine learning are reliable and efficient to classify the healthy peopleand the diseased people.A machine-learning-based prediction system on the basis of diagnosing the human condition for heart disease by using heart disease dataset is proposed in this study[1].Popular machine learning algorithms,feature extraction,feature selection and classifiers performance evaluation metrics such as classification accuracy, specificity, sensitivity, correlation coefficient and execution time are used in this study. Identification and classification of people with heart disease from healthy people can be easily carried out using this proposed study. We have already discussed about the various classifiers, feature selection algorithms, pre-processing methods, validation method and classifiers performance evaluation used in this. On a full set of features the performance of the proposed system can be validated.In terms of accuracy and execution time of classifiers, the features reduction has an impact on classifiers performance.

So, in the efficient diagnosis of the heart patients, this proposed machine-learning-based decision support system will surely be a great aid for the doctors.

II. DATASET

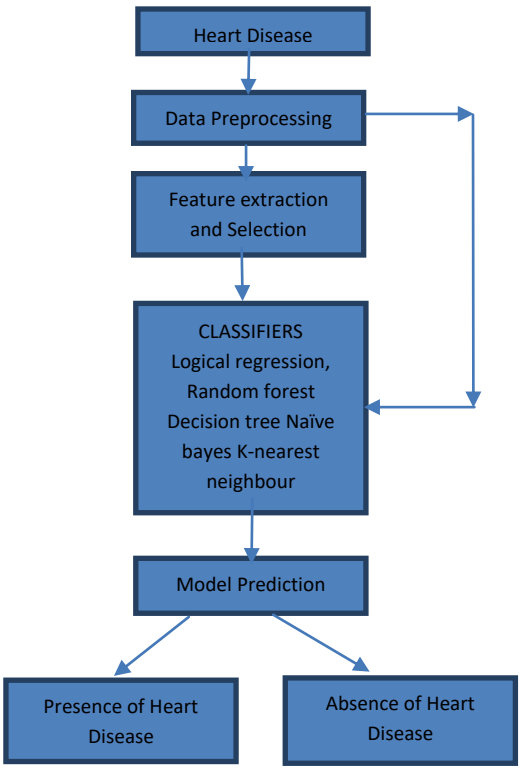
In this paper,the “Cleveland heart disease dataset 2016” is used. It can be accessed from online data mining repository of the University of California, Irvine (UCI). In this research study, this dataset was used for designing machine-learning-based system for heart disease diagnosis and prediction. The Cleveland heart disease dataset comprises a sample size of 303 patients, 76 features and values. For analysis, the sample size is sought as 303 with 13 features more appropriate independent input features, and target output label was extracted and used for diagnosing the heart disease.

III. METHODOLOGY

The aim of this proposed system is to differentiate the people with cardiovascular disease from healthy people. For the diagnosis of cardiovascular disease, the full and selected features of the performance of different machine learning predictive models were tested. The performance of the classifiers and feature selection algorithms was tested and reduced dimensionality of dataset comprising data correlated with each other is used in this. The popular machine learning classifiers, random forest, logical regression, k-nearest neighbor, decision tree, and naïve bayes

are used in the system. In this research paper, fold cross-validation (CV) method and four performance evaluation metrics is used. The proposed system's methodology is carried out in five stages:

1. Pre-processing of dataset.
2. Feature extraction.
3. Feature scaling.
4. Machine learning classifiers, and
5. Classifier's performance evaluation methods.



1. Data pre-processing:

For efficient representation of data and machine-learning classifier, the preprocessing of data is necessary which should be trained and tested in an effective manner. For the effective use in the classifiers,pre-processing techniques such as removing of missing values and standard scalar have been applied. From the dataset, missing value feature row is just deleted. In this research, all these data preprocessing techniques are used.

2. Feature extraction and selection:

Feature extraction is an attribute reduction process. Feature extraction actually transforms the attributes, unlike feature selection which ranks the existing attributes according to their predictive significance. The linear combinations of the original attributes are the transformed attributes, or features. Sometimes irrelevant features affect the classification performance of the machine learning classifier, so feature selection is necessary for the machine learning process (4). Feature selection reduces the model execution time and improves the -classification accuracy. Important feature for the selection of feature selection is to reduce the dimensionality of features while using on full feature working. Standard scalar is used for the feature scaling.

3. Machine learning classifiers:

Machine learning classification algorithms are used to classify the heart patients and healthy people. In this paper, some popular classification algorithms and their theoretical background are discussed briefly. The class of given data points is predicted by the process of classification.

Classes are sometimes called as targets or labels or categories. The task of approximating a mapping function (f) from input variables(X) to discrete output variables (y) is done by classification predictive modeling. Classification belongs to the category of supervised learning where the targets also provided with the input data. Some classifiers are used in this project. They are K-Nearest –neighbours-classifier, decision-tree, naïve bayes, logical-regression and random-forest classifier.

- **K-Nearest Neighbour classifier**

K-NN is a supervised learning classification algorithm.K-NN algorithm [2] predicts the class label of a new input: K-NN utilizes the similarity of new input to its input samples in the training set. The K-NN classification performance is not good if the new input is same the samples in the training set. Let (x,y) be the training observations and the learning function h:x⟶y, so that given an observation x,h(x) can determine y value. It is a lazy learning algorithm which stores all instances correspond to training data points in re-dimensional space.

When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors.[5]

In the distance-weighted nearest neighbor algorithm, it weighs the contribution of each of the k neighbors according to their distance using the following query giving greater weight to the closest neighbors. In this paper we use K-Nearest Neighbor algorithm, a non-parametric method used for classification and regression. To find a predefined number of training samples closest in distance to the new point and predict the label from these dataset is the principal behind the nearest neighbor methods. Train accuracy: 78.21% and Test accuracy 63.81% is produced by this. In this we take 8 as neighbors and it turns out that default value of n neighbors[4] is optimal.

- **Decision tree classifier**

A decision tree is a supervised machine learning algorithm [7]. It is just a tree where every node is a leaf node or decision node. The techniques of the decision tree are simple and easily understandable for how to take the decision. A decision tree contained internal and external nodes linked with each other. The internal nodes are the decision-making part and the child node to visit the next nodes. The leaf node on the other hand has no child nodes and is associated with a label.

Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting a termination condition. The tree is constructed in a top-down recursive divide and conquer manner. All the attributes should be categoricalotherwise, they should be discretized in advance. attributes In the top of the tree have more impacts towards India classification and they are identified using the information gain concept. DT algorithm creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. it is simple to understand and interpret and it's possible to visualise how important a particular feature was for our tree. it produces train accuracy 100% and test accuracy 78.82%.

we can ensure variable 'tha'l turns out to be a significantly important feature. Remembermy hypothesis that fasting "blood sugar" is a very weak feature? table graph confirms this clearly. decision tree model learns that train set perfectly, and at the same time is entirely over-fitting the data, what results in poor production. Other values of "Max-depth" that parameter need to be tired out. with max-depth set as 6, the scorer went to almost 84 percentage.by now, KNN outperforms decision tree. by now , KNN outperforms decision tree.it has 82.3% in training accuracy after pruning of tree.

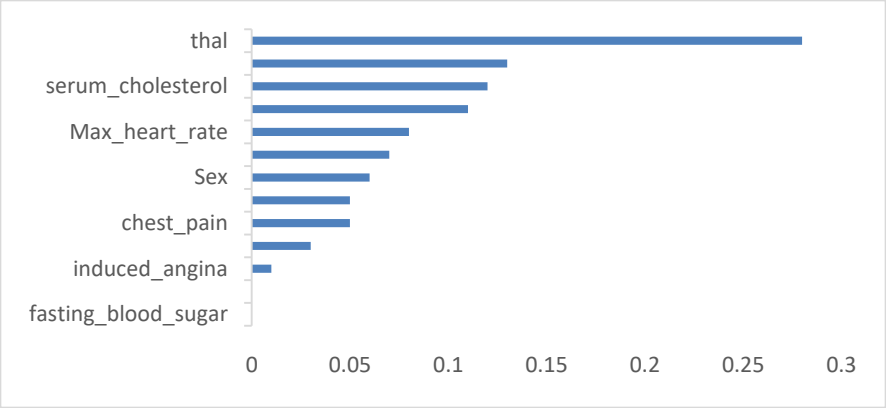


Figure 1 : The plotting of decision tree classifier result

- **Logistic regression**

Logistic regression is basic technique in statistical analysis that attempts to predict a data value based on your observations. A logistic regression algorithm looks at the relationship between a dependent variable and one or more dependent variables. A logistic regression is a classification algorithm [3]. For binary classification problem, in order to predict the value of predictive variable y when y Є [0,1],0 is negative class and 1 is positive class. It also uses multi-classification to predict the value of y when y Є [0,1,2,3]

In order to classify two classes O and I, a hypothesish(∂) = ∂^T X will be designed and threshold classifier output is h(∂) at 0.5. If the value of hypothesis h∂(x)>=0.5, it will predict y = 1 which mean that the person has heart disease and if value of (x) 0.5, then predict y - 0 which shows that the person is healthy[6].

Hence, the prediction of logistic regression under the condition is done. Logistic regression sigmoid function can be written as follows:

$$h\partial(x)=g(\partial^T X),$$

where g(z)-I/(1-X^{-z})

It produces Train accuracy 84.85% and Test accuracy : 85.71%. Negligible variation between train and test score indicates us as that model performs at the optimal level. Although the result itself is slightly lower than KNN, yet is still satisfactory.

- **Naivebayes**

In machine learning, native bayes classifiers are a family of simple probabilistic classifiers based on applying based Baye's theorem with strong (naive) independence assumptions between the features. The NB is a classification supervised theorem to determine

the class of a new feature vector. The NB uses the training dataset to find out the conditional probability value of vectors for a given class. After completing the probability conditional value of each vector, the new vectors class is computed based on its conditionality probability. NB is used for text-concerned problem classification. It produces Train accuracy 83.38% and test accuracy 85.81%. This model produced the better result than KNN algorithm. While it slightly under-fits the data, this model doesn't offer any hyper-parameters for tuning and improve overall performance.

- **Random Forests**

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and computing the class that is the mode the classes (classification) or mean production (regression) of the individual trees.

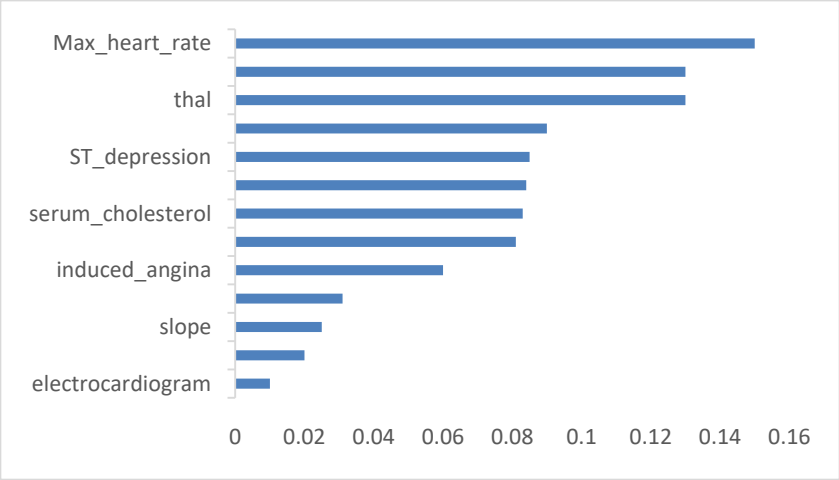


Figure 2. Features of random forest

Random forest can be used for both classification and regression problems, by using random forest regressor we can use random forest on regression problems. It has best result out of the classifier, it produces train accuracy : 100% and test accuracy 88.8%, but it considers as overfitting so we prune the trees after that it produces better result that is Train accuracy : 87.6% and Test accuracy 87.8% that brings efficient prediction.

IV PERFORMANCE EVALUATION METRICS

For checking the performance of the classifiers, different performance evaluation metrics were used in this research. We used confusion metrix, every observation in the testing set is predicted exactly in one box. It is 2x2 matrix because there are 2 repose classes. Moreover, it gives two types of correct prediction of the classifier and two types of classifier of incorrect prediction. Table shows the confusion matrix.

	Predicted HD patient (1)	Predicted healthy person (0)
Actual HD patient (1)	TP	FN
Actual healthy person (0)	FP	TN

Table 1. Confusion matrix

From confusion matrix, we compute the following:

TP*:* predicted output as true positive (TP), we concluded that the HD subject is correctly classified and subjects have heart disease.

TN: predicted output as true negative (TN), we concluded that a healthy subject is correctly classified and the subject is healthy.

FP: predicted output as false positive (FP), we concluded that a healthy subject is incorrectly classified that they do have heart disease [a type 1 error].

FN: predicted output as false negative (FN), we concluded that a heart disease is incorrectly classified that the subject does not have heart disease as the subject is healthy (a type 2 error)

I shows that positive case means diseased, and 0 shows that a negative case means healthy.

- Classification accuracy : accuracy shows the overall performance of the classification system as follows:

$$\text{classification accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{x}100\%$$

- Classificationerror: it is the overall incorrect classification of the classification model which is calculated as follows:

$$\text{Classification error} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{x}100\%$$

- Sensitivity: It is the ratio of the recently classified heart patients to the total number of

heart patients. The sensitivity of the classifier for detecting positive instances is known as "true positive rate". In other words, we can say that sensitivity (true positive fraction) confirms that if a diagnostic test is positive and the subject has the disease. It can be written as follows:

$$\text{Sensitivity (Sn) /recall/true positive rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{x}100\%$$

- Specificity : a diagnostic test is negative and the person is healthy and is mathematically written as follows :

$$\text{Specificity (Sp)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \text{x}100\%$$

- Precision : the equation of precision is given as follows :

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{x}100\%$$

The goal of the project was to compare different machine learning algorithms and predict if a certain person, given different personal characteristics and symptoms, will get heart disease or not. Here are the final results.

	accuracy
KNN	0.688525
Decision Trees	0.819672
Logistic Regression	0.852459
Naive Bayes	0.852469
Random Forests	0.885246

Figure 3. Accuracy of classifiers

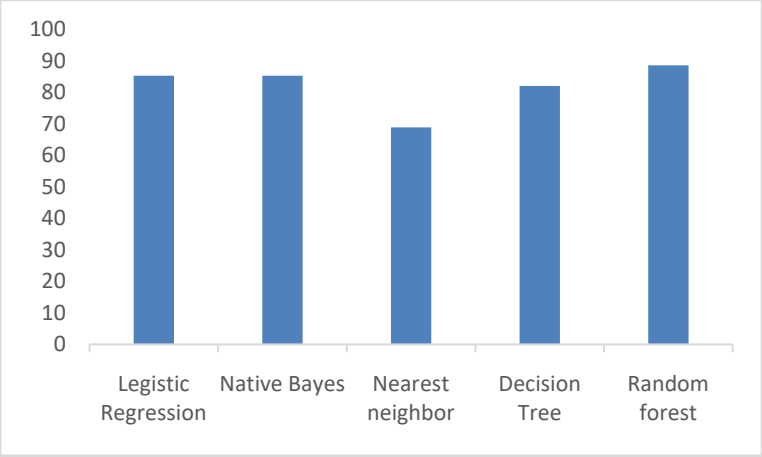


Figure 4. Graph of classifier's accuracy

V.CONCLUSION

Various machine learning techniques are used to predict that whether the heart disease is present or not in this proposed system. The accuracy and performance of proposed system is analyzed using various tools. The result can be concluded that the more complex algorithms like decision tree and random forests generated better results compared to the basic ones. In most cases it is worth to emphasize that hyper-parameter tuning is essential to achieve robust results out of these techniques. Simpler methods have also proved to be useful as well by producing decent results. In medical field, machine learning had absolutely bright future. In a place where heart disease experts are not available, we may quite accurately predict whether a disease will occur or not with just basic information about a certain patient's medical history.

REFERENCES

[1] G Biau, "Analysis of a random forests model," J.Mach.Learn. Res., Vol.13.pp.1063-1995,2012

[2] K. Srinivas, "Analysis of coronary cardio vascular disease and prediction of heart attack in coal mining regions using data mining techniques, " IEEE Trans. Comput. Sci. Educ. (ICCSE). P. p(1344-1349),2010

[3] S. Thirumuruganathan. " A detailed Introduction to K-Nearest Neighbor (KNN) Algorithm " [Online]

[4] R. Jing and Y. Zhang. "A View of Support Cector Machines Algorithm on Classification Problems". In : Proc. Of 2010 International Conference on Multimedia Communications, pp.13-16.2010

[5], S.Kiruthika Devi et al. – Prediction of Heart Disease using Data Mining Techinques, Indian Journal of Scine and Technology, Vo 9(39(, October 2016

[6] V.Chauraisa and S.Pal. "Data Mining Approach to Detect Cardiac Vascular Diease," Int. J. AdvComput Sci. Inf. Technol, Vol. Vol 2, no. No. 4, p.pp 56-66., 2013

[7] Q.J "Induction of decision trees. "Machine Learn. Vol. 1, p.81-106, 1986