International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 9 Issue 8 ǁ 2021 ǁ PP. 18-23

Automated Heart Disease Recogniton System

Punam S. Patil
PG Student
SSVPS'S B.S. Deore Collage of Engineering, Dhule,4245005, India

Abstract
Machine learning is one of the most widely sought after technol- ogy in to- day's world for solving real life problems. Automated disease diagnosis is one of the vital applications of machine learning which has the ability to revolutionize the health care industry. We as machine learning engineers look forward to- wards experimenting with and analyzing medical data to pro- duce substantial results. To begin with, we have worked over the fam- ous Cleveland dataset for heart disease diagnosis. It included experi- menting with various machine learning algo- rithms as well as neural net- works over the dataset. The results obtained and further analysis of it has been presented in this report.
Keywords *– Logistic Regression, Naive Bayes , Decision Tree, Random Forest, SVM,K-NN, Heart Disease Prediction.*

Date of Submission: 23-07-2021 Date of acceptance: 08-08-2021

I. Introduction

Heart disease has been the significant cause of death in the world for the last 10 years. Mil- lions of people die every year because of heart disease and large population of People suf- fers from heart disease. A need to develop such a medical diagnosis system arises day by day. The important key points of such medical diagnosis systems are reducing cost and obtain- ing more accurate rate efficiently. Developing a medical diagnosis system based on machine learning for prediction of heart disease provides more accurate diagnosis than traditional wa y and reduces cost of treatment.

In this paper, prediction of heart disease by an automated medical diagnosis system based on machine learning is proposed to satisfy this need. Machine Learning (ML) which is subfield of data mining handles large scale wellformatted dataset efficiently. In the medi cal field, machine learning can be used for diagnosis, detection and prediction of various d iseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage. This in turn will help to provide effective treatment to patients and avoid se- vere consequences. ML plays a very important role to detect the hidden discrete patterns a nd thereby analyse the given data. After analysis of data ML techniques help in heart dis- ease prediction and early diagnosis.

This paper provides a comparison of different machine learning classifica- tion techniques, such as Logistic Regression, Decision Tree (DT), Naïve Bayes (NB), K-

Nearest Neighbor (KNN), Random Forest and Support Vector Machine (SVM), and of the ir use in combination, through bagging, boosting and stacking on a heart disease data set using 10Fold Cross Validation as the data portioning model. The dataset used is the Clevel and Heart Disease data set taken from the University of California, Irvine (UCI) learni ng data set repository, donated by Detrano.

II. Literature Survey

1. Analysing and improving the diagnosis of ischaemic heart disease with machine learning: Kukar etal. Conducted many experiments with various learning algo- rithms and achieved the performance level comparable to that of clinicians. Also extended the algorithms to deal with non-uniform misclassifi- cation costs in order to per- form ROC analysis and control the trade-off between sen- sitivity and specificity. The ROC analysis showed signifi- cant improvements of sensitivity and specificity com- pared to the performance of the clinicians.

2. Diagnosis of Heart Disease using Datamining Algorithm: Rajkumar and Reena In this paper the data classification is based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm. Tanagra tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are com- pared.

Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients: Parthiban and Srivatsa Successfully employed Machine learning methods such as Na¨ıve Bayes and Support Vector Machines for the classi- fication purpose. Support vector machines are a modern technique in the field of machine

learning and have been successfully used in different fields of application., the system exhibited good accuracy and predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patientget- ting a heart disease.

3. Diagnosing Coronary Heart Disease Using Ensemble Ma- chine Learning: Miao etal.
In this research, an advanced ensemble machine learning technology, utilizing an adaptive Boosting algorithm, is developed for accurate coronary heart disease diagnosis and outcome predictions. The de- veloped ensemble learn- ing classification and prediction models were applied to 4 different data sets for coronary heart disease diagno- sis, including patients diagnosed with heart disease from Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiol- ogy (HIC), Long Beach Medical Center (LBMC), and Switzerland University Hospital (SUH). The testing results showed that the de- veloped ensemble learning clas- sification and prediction models achieved model accuracies of 80.14% for CCF, 89.12% for HIC, 77.78% for LBMC, and 96.72% for SUH, exceeding the accuracies of previ- ously pub- lished research.

4. Heart Disease Diagnosis Using Machine Learning Algo- rithm: Ghumbre and Ghatol
In this paper, India centric dataset is used for Heart disease diagnosis. The correct diagnosis performance of the auto- matic diagnosis sys- tem is estimated by using classification accuracy, sensitivity and spe- cificity analysis. The study shows that, the SVM with Sequential Mi- nimization Opti- mization learning algorithm have better choice for medical disease diagnosis application.

5. Diagnosing Coronary Heart Disease Using Ensemble Ma- chine Learning: Miao etal.
In this research, an advanced ensemble machine learning technology, utilizing an adaptive Boosting algorithm, is developed for accurate coronary heart disease diagnosis and outcome predictions. The de- veloped ensemble learn- ing classification and prediction models were applied to 4 different data sets for coronary heart disease diagno- sis, including patients diagnosed with heart disease from Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiol- ogy (HIC), Long Beach Medical Center (LBMC), and Switzerland University Hospital (SUH). The testing results showed that the de- veloped ensemble learning clas- sification and prediction models achieved model accuracies of 80.14% for CCF, 89.12% for HIC, 77.78% for LBMC, and 96.72% for SUH, exceeding the accuracies of previ- ously pub- lished research.

6. Heart Disease Diagnosis Using Machine Learning Algo- rithm: Ghumbre andGhatol
In this paper, India centric dataset is used for Heart disease diagnosis. The correct diagnosis performance of the auto- matic diagnosis sys- tem is estimated by using classification accuracy, sensitivity and spe- cificity analysis. The study shows that, the SVM with Sequential Mi- nimization Opti- mization learning algorithm have better choice for medical disease diagnosis application.

III. Research Methodology

A. The Cleveland Data Set

The data set used in the current research contains 303instances with a total number of 76 attributes. However, themajority of the studies use a maximum of 14 attributes as these are closely linked to heart disease. The featuresincluded are age, sex, chest pain type, resting blood pressure,cholesterol, fasting blood sugar, resting ECG, maximum heartrate, exercise induced angina, oldpeak, slope, number of vessels colored and thalassemia, respectively. The main class has two values, “False” and “True”, corresponding to the absence or presence, respectively, of any heart disease.

B. Machine Learning Techniques

The attributes mentioned in above Table are provided as input to the different ML al- gorithms such as Random Forest, Decision Tree, Logistic Regression and Naive Bayes clas- sification techniques. The input dataset is split into 80% of the training dataset and the re- maining 20% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analysed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further. The different algo- rithms explored in this paper are listed as below.

1) Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation be- tween 0 and 1. Ther are 13 independent variables

which makes logistic regression good for classification.

2) Multinomial Naive Bayes

Multinomial Naive Bayes algorithm is a probabilistic learning method. The algorithm is based on the Bayes theorem and predicts the label for given input. It calculates the probabil- ity of each class for a given sample and then gives the class with the highest probability as output.

Bayes theorem, formulated by Thomas Bayes, calculatesthe probability of an event occurring based on the priorknowledge of conditions related to an event. It is based on the following formula:

$P(A|B) = P(A) * P(B|A)/P(B)$

3) Gaussian Naive Bayes

The Gaussian Naive Bayes approach builds upon the Naive Bayes algo- rithm but considers the probabilistic distribution curve to be Gaussian in nature.

4) Bernoulli Naive Bayes

This method is also similar to Naive Bayes algorithm but uses a Bernoulli distribution func- tion instead:

$p(k, m) = (mk)q^k(1 - q)^{n-k}$

5) Support Vector Machine

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal
hyperplane (MMH).

6) Decision Tree:

A Decision Tree is a flow chart- like structure that includes a root node, branches, and leaf nodes. The dataset attributes are defined through the internal nodes. The branches are the outcome of each test against each node. It is a popular classifier because it is simple, fast, and easy to interpret, explain and implement. It requires no do- main knowledge or parameter setting.

7) Random Forest

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets a class expectation and the class with most votes turns into a model's forecast. In the ran- dom forest classifier, the more the number of trees higher is the accuracy It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable

8) K-Nearest Neighbors (K-NN):

K-Nearest Neighbors classifies an object by the majority vote of its closest neighbors. In other words, based on some distance metrics, the class of a new instance will be predicted. The distance metric used in nearest neighbor methods for numerical attributes can be a simple Euclidean distance

***C.* Accuracy calculation**

Detecting presence of disease

The first model we design is based on binary classification where the output classes 1,2,3,4 are effectively considered as 1 denoting presence and output class 0 denotes absence. For a new test case, the model will predict if the patient has a heart disease or not.

Accuracy of the algorithms are depends on four values namely true positive(TP), false positive(FP), true negative(TN) and false negative(FN).

Accuracy= (FN+TP) / (TP+FP+TN+FN)

The numerical value of TP, FP, TN, FN defines as: TP= Number of person with heart diseases.
TN= Number of person with heart diseases and no heart diseases. FP= Number of person with no heart diseases.
FN= Number of person with no heart diseases and with heart diseases.

Algorithm	Accuracy(%)
Logistic regression	76.31
Multinomial Naive Bayes	72.37
Gaussian Naive Bayes	84.21
Bernoulli Naive Bayes	77.63
Linear SVC	89.47
Decision tree classifier	65.79
Random forest classifier	84.21
K Neighbors Classifier	84.21

Thus it can be seen that Linear SVC is able to detect presence of heart disease with an accuracy of 89% which is the highest among its counterparts.

Multiclass classification of disease

The model designed for this type of classification outputs a dis- crete value between 0 and

2. Similar to the previous model, 0 indicates complete absence of heart disease and 1,2,3,4 indicate presence of heart disease but with increasing severity of the same. Thus a patient can get a better insight into their health status by such labels.

As a result, it is a challenging task to achieve higher accu- racy in such problems. We have however given our best efforts to produce superior re- sults. The performance metric for this model is same as earlier wiz. accu- racy. The results we ob- tained are as follows:

Algorithm	Accuracy(%)
Logistic regression	55.26
Multinomial Naive Bayes	53.95
Gaussian Naive Bayes	38.16
Bernoulli Naive Bayes	51.32
Linear SVC	53.95
Decision tree classifier	52.63
Random forest classifier	55.26
K Neighbors Classifier	57.9

Thus it can be seen that KNN is the best algorithm and it is able to deliver an accuracy of around 58% on the test dataset.

Performance on other met- rics

As a part of the ablation studies, we also show the performance of the system on three other metrics:

1. Precision: How correct the system has been in predicting the true positive samples.

Precision =TP/ TP+FP

2. Recall :How many of the true positive samples was the system able to correctly predict

Recall =TP/TP+FN

3. F1 score: It is the harmonic mean of the precision and recall of the algorithm

F1Score = 2PR/ P+R

We first present the results of all the algorithms on these metrics for the pres- ence detection i.e. binary task, weighted to their representation.

Algorithm	Precision	Recall	F1 score
Logistic regression	0.76	0.76	0.76
Multinomial NB	0.72	0.71	0.72
Gaussian NB	0.85	0.84	0.84
Bernoulli NB	0.78	0.78	0.78
Linear SVC	0.89	0.87	0.88
Decision tree classifier	0.66	0.63	0.64
Random forest classifier	0.83	0.81	0.81
KNN classifier	0.80	0.84	0.82

It can be seen that the Linear SVC algorithm retains its performance even on the other three metrics. This could be attributed to the fact that as a binary classification task, the dataset is fairly distributed and hence the accuracy perfor- mance carries on to other metrics.
Next, we calculate the performance of these algorithms for the multiclass classifica- tion task. Note that the metric values are weighted to their class representation

Algorithm	Precision	Recall	F1 score
Logistic regression	0.27	0.27	0.24
Multinomial NB	0.61	0.25	0.24
Gaussian NB	0.21	0.28	0.19
Bernoulli NB	0.24	0.25	0.24
Linear SVC	0.25	0.30	0.24
Decision tree classifier	0.30	0.33	0.31
Random forest classifier	0.25	0.28	0.26
KNN classifier	0.32	0.30	0.30

It can be seen that no single algorithm dominates all the three metrics. Further the performance has also dropped for the multiclass classification task. Decision tree produces the best F1 and recall values while Multinomial NB was morepre- cise in its predictions.

IV. Results

For both of these models, we take into consideration the highest accuracy obtained during cross validation as those weightscan then be further used in the future. The results obtained are as follows:

Model	Accuracy(%)
Presence of disease	91.8
Multiclass classification	58.1

Result analysis

Thus it can be seen that neural networks outperforms reg- ular machine learning algorithms in both uniclass as well as multiclass classification. For detecting presence or absence of heart disease, an accuracy of around 92% is achieved which is more than the logistic regression algorithm. In case of multi- class classification, neural networks match the performance of other algorithms to achieve an accuracy of 58%. Further work involves tuning the hyperparameters to achieve even better re- sults, working on addition of features as well as normalization of data to speed up the com- putation.

References

[1]. Classification algorithms in machine learning:https://medium.com/@sifium/machine- learning-types-of-classification 9497bd4f2e14
[2]. Scikit learn: https://scikit- learn.org
[3]. Keras: https://www.keras.io
[4]. Multiclass classification using Keras: https://machinelearningmastery.com/multi-class-classification- tutorial-ke-ras-deep-learning-

library/
[5]. Kukar, M., Kononenko, I., Gro˘selj, C., Kralj, K. and Fet- tich, J., 1999. Analysing and improving the diagnosis of ischaemic heart dis- ease with machine learning. Artificial intelligence in medicine, 16(1), pp.25-50.
[6]. Rajkumar, A. and Reena, G.S., 2010. Diagnosis of heart disease using datamining algorithm. Global journal of computer science and tech- nology, 10(10), pp.38-43.
[7]. Parthiban, G. and Srivatsa, S.K., 2012. Applying machine learning methods in diagnosing heart disease for diabetic patients. Interna- tional Journal of Applied Information Systems (IJAIS), 3(7), pp.25- 30.
[8]. Miao, K.H., Miao, J.H. and Miao, G.J., 2016. Diagnosing coronary heart disease using ensemble machine learning.