International Journal of Control and Automation
Vol. 13, No. 1s, (2020), pp. 31-37

Optimal Machine Learning Classifiers for Prediction of Heart Disease

Rahul Kumar Jha[1], Dr Santosh Kumar Henge[1], Dr Ashok Sharma[1]
[1]*School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India*

Abstract

Cardiovascular Disease (CVDs) is the major cause of mortality. Heart disease is one of the serious and prevalent diseases which is impacting millions of people across the globe and needs special attention and cure. Many researchers are working together to find feasible solution for prediction of heart disease and for that they are involving popular technology i.e. Artificial Intelligence (AI) and Machine Learning (ML). There are many popular classifiers in machine learning like Decision Tree (DT), Support Vector Machine, Naive Bayes (NB), KNN, Logistic Regression (LR), Artificial Neural Network, Deep Neural Network, and many more which has been used by researchers to find a solution for many health related disease. In this paper, an experiment has been conducted for prediction of heart disease using popular classification models like SVM, DT, NB, Random Forest, DNN and KNN and to find best model suitable for this purpose using ML.

Keywords: *Cardiovascular Disease (CVDs), Heart Disease (HD), Prediction of Heart Disease (HDP), Decision Tree (DT), Genetic Algorithm (GA), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naïve Bayes (NB), Logistic Regression (LR), Recurrent Fuzzy Neural Networks (RFNN).*

1. Introduction

Cardiovascular Disease (CVDs) or Heart Disease (HD) is the number one cause of death on the globe and rate of death due to heart disease is more than any other disease. According to World Health Organization (WHO) survey, 17.7 million death occurred worldwide because of CVDs among which 85% death was because of heart attack and most impacting age with this disease is 70's [1], 7 million people die every year because of smoke. Smoke thicken the blood vessels, raises triglycerides - a type of fat and ultimately cause heart attack [2]. This needs special attention and cure and prediction of heart disease in early stage is very much important. Many researcher across the glove are working day and night to find a solution to prevent heart disease and this has been a most interesting and challenging subject for many of them. Artificial Intelligence (AI) with Machine Learning (ML) has emerged as a life saver and this technology has spread its existence in nearly all the fields like automation, robotics, image processing, voice recognition, health domain and many more. The technology has proved itself so prominent that today every researcher has included this technology in their research work and working toward finding solution of many serious problems in health and other domain and contributing in social cause; few of their research has been considered in this proposed study and included as literature review in section 2. Machine learning with its popular classification models like DT, SVM, NB, KNN, Logistic Regression (LR), ANN, DNN, Fuzzy System and hybrid system contributed a lot in finding solution for many social problems and its contribution in health disease prediction like cancer and heart disease has been remarkable and been helpful in lowering down the death rate all over the globe. In this paper an experiment has been carried out using few of the popular ML techniques. Remaining part of the paper is structured as: section 2 contains literature review of many related articles, we will review existing study done by many researchers; section 3 contains dataset information to be used in the experiment; section 4 has detailed discussion on the proposed experiment following with conclusion and future work in section 5 and reference in section 6.

ISSN: 2005-4297 IJCA

2. Comparative Analysis of Past Designed Approaches

Many researchers have done some remarkable work toward finding solution for HDP and achieved wonderful results using some popular ML techniques like DT, NB, KNN, ANN, DNN and many more. Many researchers have implemented hybrid system in their research work which outperformed better than other traditional algorithms; use of feature selection has helped in taking out surprising results. Model generated by their research have really helped society in prediction of disease and helped in decreasing death rate in society and ultimately creating a better tomorrow.

The author G. S. Santhana Krishnan J presented the study of "Prediction of Heart Disease Using Machine Learning Algorithms" in a conference to showing experiment using python programming and ML Algorithm namely DT and NB Algorithm and to find the best among these in predicting HD [3]. The author U. H. Memon has proposed the approach of "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", it has developed a ML-based diagnosis system for HDP such as Seven popular ML algorithms, three feature selection algorithms, the cross-validation method were experimented to compare the accuracy and to find the best model [4]. The authors Beulah and Jeeva proposed the approach of "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", it has described the study to advance the performance of fragile classification algos, to implement them with a medical dataset and to show its usefulness for prediction of disease at an premature stage. Experiment showed an improvement of 7% in weak algorithms experimented in this paper [5].

The authors A. H. S. R. G. W. Ravindhar NV has proposed the approach of "Intelligent Diagnosis of Cardiac Disease", it has the study which was presented taking four ML algorithm with Neural Network to demonstrate the performance and to predict the heart disease. Result showed an accuracy of 98% [6]. The author T. S. K. R. S. S. M. B. Praveen Kumar Reddy M has proposed the approach of "Heart Disease Prediction Using Machine", it has presented and shown the experiment with ML Algorithm SVM and DT to find the best among these in predicting HD [7]. The author K. S. P. R. S. Purushottam has proposed "Efficient Heart Disease Prediction System A framework", which is used to build a system to enable technical and non-technical doctors to predict heart disease. This could help doctors in making correct decision about the HD risk [8].

C. T. P. D. M. G. F. A. V. Sabrina Mezzatesta has proposed the approach of "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis", it has described the scenario which was discovered where patients with End_Stage_Kidney_Disease (ESKD) faced CVD risk in parallel. To overcome this issue a study was presented aiming at prediction of disease with a certain precision, death and CVD in dialysis patients [9]. The author C. P. Shashikant R has projected the ""Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter", which has presented a study to predict cardiac arrest in smokers using ML classifiers and Heart_Rate_Variability (HRV) parameters. Study compared many popular ML classifiers and result showed that Random_Forest (RF) model accomplished the best result with ACC of 93.61%, precision of 94.59%, SENS of 92.11%, the SPEC of 95.03% [10].

The author M. F. O. U. Kevin Buchan has projected the approach of ""Automatic prediction of coronary artery disease from clinical narratives". This system was developed that was capable to predict coronary artery disease in patient based on their narrative medical histories, result showed 77.4% F1-Score [11]. The author E. P. E. G.-O. Juan-Jose Beunza has proposed the approach of "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)". This analysis based study was presented to compare ML algorithms to diagnose HD in patient. Several popular ML algorithms were experimented to compare the accuracy and to find the best model [12]. The author V. S. Divya Jain has presented a review of "Feature selection and

classification systems for chronic disease prediction: A review". It has described the performance of popular ML classification and feature selection algorithms to predict heart disease. The experiment with various methods with cross-validation method was reviewed to show the comparison of ACC, SENS and SPEC [13].

The author A. I. Kaan Uyar has proposed the approach of "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks". This analysis was proposed having GA based trained RFNN to diagnosis of heart diseases. Model was trained using Cleveland dataset and achieved ACC of 97.78% [14]. The author T. M. S. H. H. K. Rajesh Nichenametla has proposed the approach of "Prediction of Heart Disease Using Machine Learning Algorithms". It has presented a study showing experiment with ML Algorithm NB and DT to find the best among these in predicting HD. NB showed best result among two when dataset is large, DT is good for small dataset [15].

3. Dataset

For this research work Cleveland heart disease dataset has been used. This data is freely available in UCI repository and has been commonly used by many researchers performing similar study using ML. Dataset contains 303 heart patient records having 76 attributes including personal and clinical data out of which 6 records has been deleted during pre-data processing. For experiment purpose, 14 attribute has been chosen among which, 13 attributes has been used as feature and one attribute (num) used as the output for HDP. This attribute has value lie between 0-4 where 0 means no heart disease and numbers 1 to 4 means that patient has heart disease.

3.1 Dataset attributes

Below table 1 lists the 14 attributes from Cleveland dataset [16].

Table 1: Cleveland dataset attributes

Age of patient (AGE)	Angina due to Exercise (EIA)
Patient gender (SEX)	Old Peak – ST depression by exercise (OPK)
Chest pain type (CPT)	Slope of peak exercise ST (PES)
Resting Blood Pressure (BP)	Major vessel number (VCA)
Fasting Blood Sugar (FBS)	Thallium Scan (THA)
Resting ECG Result (RES)	Cholesterol (SCH)
Max Heart Rate (MHR)	Heart Disease Diagnosis (NUM)

3.2 Distribution of dataset

The data has been distributed considering heart disease diagnosis classification. It has been distributed between 0 to 4 where 164 records are related to patient who don't have heart disease and rest records are classified into 1-4 based on the diagnosis [16]. Below graph shows the distribution of records based on num attribute.

ISSN: 2005-4297 IJCA

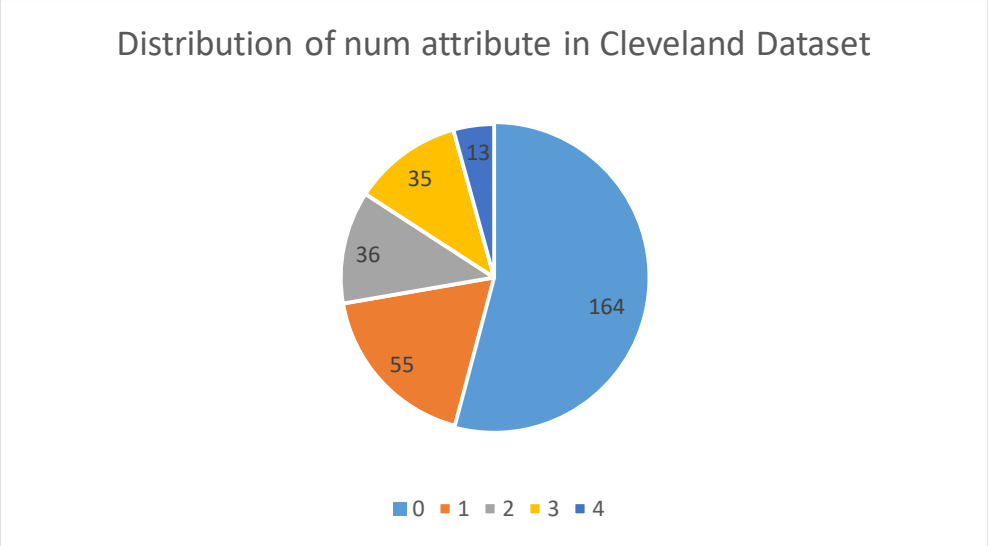

Figure 1: Graph showing distribution of num attribute in Cleveland dataset

4. Experiment and Result

Experiment has been done on various popular ML classification methods like SVM, DT and etc. using Rapid Miner tool; model optimizing has been done several times to achieve considerable result; feature selection has been introduced to get better result; result has been analysed on the basis of various parameters like accuracy (ACC), sensitivity (SENS), specificity (SPEC), AUC and F-Measure.

4.1 Experiment workflow

Cleveland dataset [16] was inputted into Rapid Miner tool; data was pre-processed before applying feature selection; feature selection was applied with reason that many articles which were reviewed during experiment, demonstrated outperformed result in terms of accuracy matrix and time elapsed in training model when applying feature selection; various ML classifiers were experiment to compare the model output results. Below figure explains the workflow which was followed during the experiment.

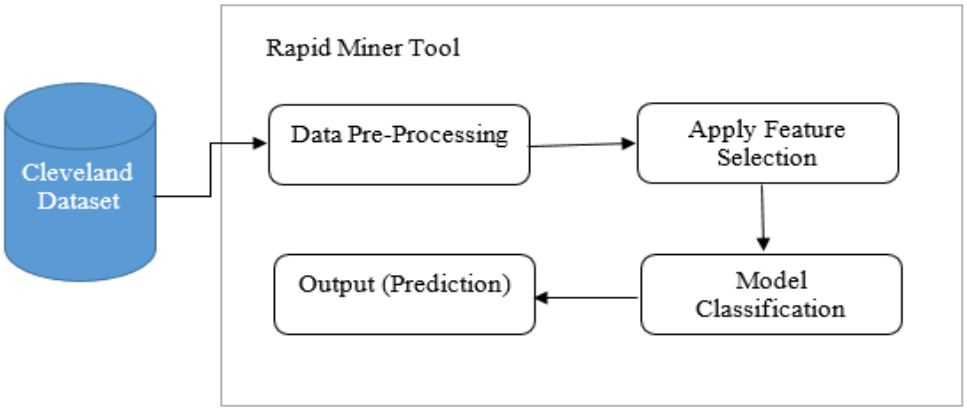

Figure 2: Experiment workflow using Rapid Miner Tool

ISSN: 2005-4297 IJCA

4.2 Data Pre-Processing

This dataset [16] have 303 records out of which 6 records are found as missing data and has been deleted and final 297 records were taken forward for experiment purpose.

4.3 Performance Evaluation Matrix

Further, to evaluate the performance of the generated classification models, 2X2 confusion matrix has been created which was based on true and false prediction combination. Below table shows the confusion matrix used for calculation.

Table 2: Confusion matrix showing values that has been used for further computation.

Actual result	Predicted HD (1)	Predicted No HD (0)
Has heart disease (1)	TP	FN
No heart disease (0)	FP	TN

According to above matrix, it could be concluded that:
TP (True Positive) – if a patient has HD and he is predicted as HD positive
TN (True Negative) – if a patient has no HD and he is predicted as no HD
FP (False Positive) – if a patient has no HD but he is predicted as HD (Also known as type 1 error)
FN (False Negative) – if a patient has HD but he is predicted as no HD (it is type 2 error)
Based on above conclusion, let's calculate performance matrix:

$$\text{Accuracy (or overall performance of model)} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \text{X } 100\%$$

$$\text{Sensitivity (has heart disease)} = \frac{\text{TP}}{\text{TP+FN}} \text{X } 100\%$$

$$\text{Specificity (no heart disease)} = \frac{\text{TN}}{\text{TN+FP}} \text{X } 100\%$$

$$\text{Classification Error (incorrect classification)} = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} \text{X } 100\%$$

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

4.4 Feature Selection

Below table 3 shows the feature selection applied to different classifiers.

Table 3: Feature selection applied to different models

Model	**Feature name**
Decision Tree	AGE, CPT, BP, PES
KNN	AGE, SEX, BP, PES, OPK
SVM	AGE, RES, PES, BP, CPT
DNN	SEX, CPT, BP, FBS, RES, PES, THA
Random Forest	AGE, BP, CPT
Naïve Bayes	AGE, THA, SCH, OPK

ISSN: 2005-4297 IJCA

4.5 Performance Result

Based on above equations, performance of different model were calculated and enlisted in below table 4. Each model were trained for 10-15 minutes each and repeated epoch were performed.

Table 4: Result analysis of different model

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	F-Measure (%)
Decision Tree	73.3	78.8	54.1	76.2	81.6
KNN	76.7	72.4	74.3	78.4	73.2
SVM	88.6	83.8	81.1	92.3	91.4
DNN	93.4	91.6	88..4	92.1	87.9
Random Forest	82.2	74.4	76.7	83.6	80.3
Naïve Bayes	83..6	87.8	78.4	83.9	82.6

5. Conclusion and Future Work

With Machine learning, it is now possible to diagnose heart disease before the condition became abnormal. Many classification model has proved its relevance in making better solution for HDP. In this study, many poplar ML classifiers i.e. DT, SVM, Random Forest, DNN, NB, KNN etc. has been selected to compare the performance and find the best optimized model to be used for prediction of HD. From experiment output, it is clear that though SVM performed better than rest of the model with ACC of 88.6% but DNN from Neural Network family outperformed over other traditional classification models with ACC, SENS and SPEC as 93.4%, 91.6% and 88.4% respectively, same was observed during literature review too. In future, hybrid system using neural network with feature selection algorithm could perform better in terms of model performance and time elapsed.

References

[1] WHO, "Cardiovascular Diseases," who, [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases/.

[2] "Smoking and Heart Disease," cardiosmart, [Online]. Available: https://www.cardiosmart.org/Healthy-Living/Stop-Smoking/Smoking-and-Heart-Disease.

[3] G. S. Santhana Krishnan J., "Prediction of Heart Disease Using Machine Learning Algorithms," IEEE, June 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8741465.

[4] u. H. Memon, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," hindawi, October 2018. [Online]. Available: https://www.hindawi.com/journals/misy/2018/3860146/.

[5] L. S. C. J. C. Beulah Christalin, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," sciencedirect, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235291481830217X.

[6] A. H. S. R. G. W. Ravindhar NV, "Intelligent Diagnosis of Cardiac Disease," ijitee, September 2019. [Online]. Available: https://www.ijitee.org/wp-content/uploads/papers/v8i11/J97650881019.pdf.

[7] T. S. K. R. S. S. M. B. Praveen Kumar Reddy M, "Heart Disease Prediction Using Machine," ijitee, August 2019. [Online]. Available: https://www.ijitee.org/wp-content/uploads/papers/v8i10/J93400881019.pdf.

[8] K. S. P. (. R. S. Purushottam, "Efficient Heart Disease Prediction System," sciencedirect, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091630638X.

[9] C. T. P. D. M. G. F. A. V. Sabrina Mezzatesta, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis," sciencedirect, August 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0169260718317188.

[10] C. P. Shashikant R., "Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter," sciencedirect, June 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210832719301048.

[11] M. F. O. U. Kevin Buchan, "Automatic prediction of coronary artery disease from clinical narratives," sciencedirect, August 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046417301466.

[12] E. P. E. G.-O. Juan-Jose Beunza, "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)," sciencedirect, September 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1532046419301765.

[13] V. S. Divya Jain, "Feature selection and classification systems for chronic disease prediction: A review," sciencedirect, November 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866517300294.

[14] A. I. Kaan Uyar, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," sciencedirect, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091732495X.

[15] T. M. S. H. H. K. Rajesh Nichenametla, "Prediction of Heart Disease Using Machine Learning Algorithms," researchgate, May 2018. [Online]. Available: https://www.researchgate.net/publication/326733163_Prediction_of_Heart_Disease_Using_Machine_Learning_Algorithms.

[16] M. P. Robert Detrano, "UCI Heart Disease Data Set," V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

ISSN: 2005-4297 IJCA