Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 85 (2016) 962 – 969

Procedia Computer Science

Efficient Heart Disease Prediction System

Purushottam[a,c]* , Prof. (Dr.) Kanak Saxena[b], Richa Sharma[c]

[a]Research Scholar , R.G.P.V. Bhopal(M.P),India.
[b]Prof & Head, Department of Computer Application S.A.T.I. Vidisha (M.P), India
.[a] [c]Assistant Prof, Amity University Uttar Pradesh, Noida, India.

Abstract

Cardiovascular sickness is a major reason of dreariness and mortality in the present living style. Distinguishing proof of cardiovascular ailment is an imperative yet an intricate errand that should be performed minutely and proficiently and the right robotization would be exceptionally attractive. Each individual can't be equivalently skilled thus as specialists. All specialists can't be similarly talented in each sub claim to fame and at numerous spots we don't have gifted and authority specialists accessible effortlessly. A mechanized framework in therapeutic analysis would upgrade medicinal consideration and it can likewise lessen costs. In this exploration, we have planned a framework that can proficiently find the tenets to foresee the risk level of patients in view of the given parameter about their health. The main contribution of this study is to help a non-specialized doctors to make correct decision about the heart disease risk level. The rules generated by the proposed system are prioritized as Original Rules, Pruned Rules, Rules without duplicates, Classified Rules, Sorted Rules and Polish . The execution of the framework is assessed as far as arrangement precision and the outcomes demonstrates that the framework has extraordinary potential in anticipating the coronary illness risk level all the more precisely.

Keywords: Heart disease prediction System , Polish, CVD, CAD ,C4.5.

1. Introduction

In today's opportunity at numerous spots clinical test outcomes are regularly made in light of specialists' instinct and experience as opposed to on the rich data accessible in numerous expansive databases. Numerous a times this procedure prompts inadvertent predispositions, lapses and a tremendous medicinal expense which influences the nature of administration gave to patients.

Today numerous doctor's facilities introduced some kind of quiet's data frameworks to man-age their social insurance or patient information. These data frameworks commonly produce a lot of information which can be in distinctive organization like numbers, content, diagrams and pictures yet sadly, this database that contains rich data is once in a while utilized for clinical choice making.

Like business knowledge and examination, the term information mining can mean diverse things to distinctive individuals. In exceptionally straightforward way we can characterize information mining as this is the investigation of substantial information sets to discover examples and utilize those examples to foresee or fore-cast the probability

Peer-review under responsibility of the Organizing Committee of CMS 2016
doi:10.1016/j.procs.2016.05.288

of future occasions. The motivation to do this problem comes from World Health Organization estimation. According to the World Health Organization estimation till 2030, very nearly 23.6 million individuals will pass on because of Heart malady. So to minimize the danger, expectation of coronary illness ought to be finished. Analysis of coronary illness is typically in view of signs, manifestations and physical examination of a patient. The most troublesome and complex assignment in medicinal services area is finding of right ailment. This colossal entirety huge of rough data is the rule resource that can be capably pre-taken care of and inspected for key information extraction that direct or by suggestion influences the remedial society for cost sufficiency and reinforce decision making. Authentic determination of coronary sickness can't be possible by using simply human understanding. There are heaps of parameters that can impacts the accu-rate conclusion like less exact results, less experience, time subordinate execu-tion, data up degree and whatnot. Packs of headway and examination happened in this field using multi-parametric qualities with nonlinear and direct parts of Heart Rate Variability (HRV).A novel framework was proposed by Heon Gyu Lee et al. [4]. To fulfill this, various experts have used various classifiers e.g. CMAR (Classification Multiple Association Rules), SVM (Support Vector Machine), Bayesian Classifiers and C4.5). A latest's rate techniques in this field depicted in [8].Some plausible strategies and technique we recommended incorporates the clinical information institutionalization, examination and the information sharing over the related industries to improve the precision & viability of information mining applications in social insurance. [5] It is likewise prudent to investigate the utilization of content digging and picture digging for extension the nature and extent of information mining applications in medicinal services part. Information mining application can likewise be investigated on computerized indicative pictures for application viability. Some advancement has been made in these areas. [6][7].

There is a lot of data put away in stores that can be utilized viably to guide a medical practitioners in decision making in human services. This brings up an essential issue:

"By what means would we be able to transform information into helpful data that can empower medicinal services practitioners to settle on viable clinical decision?" This is the primary goal of this research.

2. Background

In late time, numerous associations in human services division utilizes data mining applications seriously and broadly on substantial scale. In information mining we can utilize diverse master cess and innovation to change this colossal measures of information into helpful data for solid and exact choice making. Another reason is that the social insurance exchanges created by this part are excessively voluminous and perplexing, making it impossible to be broke down and prepared by customary systems. Choice using so as to make can be enhanced majorly by using mining applications in finding patterns and examples in substantial volumes of ordinary data.[1] In late patterns investigation on these extensive dataset has gotten to be fundamental because of monetary weights on medicinal services commercial enterprises. This separated data can be utilized for choices making taking into account the relapse examination of restorative and money related information. Learning extraction can impact industry working proficiency, income and expense maintaining so as to utilize learning disclosure from database with at most care[2].Research demonstrates that on the off chance that we utilizes information mining applications as a part of social insurance organizations then these associations would be in better position to meet their fleeting objectives and long haul needs, Benko and Wilson argue.[3] We can get extremely valuable results from human services crude information by changing crude information into helpful data. An extraordinary reason that empowers analysts in this field is that this is exceptionally helpful for all partner included in the human services division. Like, in the event that we consider Insurance supplier, they can identify misuse and extortion, expert in human services can pick up help with choices making, similar to in client relationship administration. Social insurance suppliers (doctor's facilities, doctor, test research centres and patient and so forth.) can likewise utilize information mining applications in their separate master zone for master choice finding so as to make for instance, best practices and right & compelling medicines.

3. Heart Disease risk level prediction

The Heart disease database contains the screening clinical information of heart patients. At first, the database pre-processed to make the mining handle more able.

Database Details

(a) Database Creators: V.A. Therapeutic Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
(b) Database Donor: David W. Aha (aha@ics.uci.edu) (714) 856-8779
The registry contains a database related with coronary disease. Data can be collected from uci.
Cleveland Clinic Foundation (Cleveland. Data) [12].

Inputs attributes

Age, Sex, Chest Pain, Resting blood pressure, Serum cholesterol, Fasting blood sugar, Resting electrocardiographic results, Maximum heart rate achieved, Exercise induced angina, ST depression, Slope of the peak exercise ST segment, Number of major vessels colored by fluoroscopy and thal.

Outputs class attribute

num (the predicted attribute)

The proposed study used covering rules model for classification (taking into account decision trees) as C4.5Rules [10], [11], [13] on the pre-processed database and discover the created rule sets with various need. Additionally pruned and ordered standards are also calculated.
We have utilized WEKA device [15] for dataset examination and KEEL [13],[14] to discover the order choice principles.

4. KEEL Experiment Implementation:

KEEL (Knowledge Extraction based on Evolutionary Learning) is being utilized for implementation. KEEL is an open source (GPLv3) Java programming apparatus to implement developmental process for Data Mining issues.
In the proposed study an implementation is being done using the dataset from Cleveland [12].In the pre-processing stage an AllPossible-MV [13][14] algorithm used for calculation to fill the missing values in the data set.
Missing Values Handling-for each instance in the data set, the presence of missing data is tested, and if exists any, all seen values of the attribute are imputed, resulting in 1 or more instances

5. Classification Decision Rules generated in our experiments:

The decision tree is constructed top-down. In each step a test for the actual node is chosen (starting with the root node), which best separates the given examples by classes.

A hill climbing algorithm is then performed in order to find the best subset of rules (according to the MDL heuristic).

PARAMETERS

- Confidence: is the confidence level. It is a float value that determines what is the minimal confidence that must has a leaf in order to be considered in the tree. Confidence value for our study is 0.25

- MinItemsets: is the minimum number of item-sets per leaf. It is an integer value that determines how much data instances must contain a leaf in order to be created. This value is 2 for our study

- Threshold: determines which algorithm to use in order to find the best subset of rules. For rule sets with sizes under the threshold, an exhaustive algorithm is performed; for sets above the threshold, a hill climbing algorithm is used. We have set the Threshold value to 10.

The rules generated by our experiments:

5.1 Original Rules:

i. If MHR<=3 and ST depression <=2.1 and Chest Pain <=3 ➔ 0
ii. If (MHR<=3 and ST depression <=2.1 and Chest Pain >3 and ca <=0 ➔ 0
iii. If (MHR<=3 and ST depression <=2.1 and Chest Pain >3 and ca >0 and serum cholesterol <=282) ➔ 1

iv. If (MHR<=3 and ST depression <=2.1 and Chest Pain >3 and ca >0 and serum cholesterol >282) ➔ 0
v. If (MHR<=3 and ST depression >2.1 and exercise induced angina <=0 and slope of the peak exercise ST segment <=2) ➔ 1
vi. If (MHR<=3 and ST depression >2.1 and exercise induced angina <=0 and slope of the peak exercise ST segment >2) ➔ 0
vii. If (MHR<=3 and ST depression >2.1 and exercise induced angina >0 and ca <=0) ➔ 0
viii. If (MHR<=3 and ST depression >2.1 and exercise induced angina >0 and ca >0) ➔ 3
ix. If (MHR>3 and Sex <=0 and serum cholesterol <=295) ➔ 3
x. If (MHR>3 and Sex <=0 and serum cholesterol >295) ➔ 1
xi. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR<=6 and resting blood pressure <=135) ➔ 0
xii. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR<=6 and resting blood pressure >135) ➔ 2
xiii. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol <=205 and Age<=54) ➔ 3
xiv. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol <=205 and Age>54) ➔ 2
xv. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting blood pressure <=114 and Age<=53) ➔ 1
xvi. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting blood pressure <=114 and Age>53) ➔ 2
xvii. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting blood pressure >114 and serum cholesterol <=266 and ca <=0) ➔ 0
xviii. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting blood pressure >114 and serum cholesterol <=266 and ca >0 and Age<=61) ➔ 1
xix. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting blood pressure >114 and serum cholesterol <=266 and ca >0 and Age>61) ➔ 0
xx. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting blood pressure >114 and serum cholesterol >266 and resting electrocardiographic<=1) ➔ 1
xxi. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting blood pressure >114 and serum cholesterol >266 and resting electrocardiographic>1) ➔ 3

5.2 Pruned Rules:

i. If (MHR<=3 and ST depression <=2.1 and Chest Pain <=3) ➔ 0
ii. If (MHR<=3 and ST depression <=2.1 and ca <=0) ➔ 0
iii. If (MHR<=3 and ST depression <=2.1 and Chest Pain >3 and ca >0 and serum cholesterol <=282) ➔1
iv. If (MHR<=3 and ST depression <=2.1) ➔ 0
v. If (MHR<=3 and ST depression >2.1 and exercise induced angina <=0 and slope of the peak exercise ST segment <=2) ➔ 1
vi. If (MHR<=3 and exercise induced angina <=0) ➔ 0
vii. If (MHR<=3 and ca <=0) ➔ 0
viii. If (MHR<=3 and ST depression >2.1 and exercise induced angina >0 and ca >0) ➔ 3
ix. If (MHR>3 and Sex <=0 and serum cholesterol <=295) ➔ 3
x. If (MHR>3 and serum cholesterol >295) ➔1
xi. If (ST depression <=2.4 and fasting blood sugar <=0 and MHR<=6) ➔ 0
xii. If (MHR>3 and fasting blood sugar <=0 and MHR<=6 and resting blood pressure >135) ➔ 2
xiii. If (fasting blood sugar <=0 and MHR>6 and serum cholesterol <=205 and Age<=54) ➔ 3

xiv. If (Sex >0 and fasting blood sugar <=0 and MHR>6 and serum cholesterol <=205 and Age>54) ➔ 2
xv. If (ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and Age<=53) ➔ 1
xvi. If (ST depression <=2.4 and MHR>6 and resting blood pressure <=114 and Age>53) ➔2
xvii. If (ST depression <=2.4 and resting blood pressure >114 and serum cholesterol <=266 and ca <=0) ➔0

5.3 Rules without duplicates:

i. If (MHR<=3 and ST depression <=2.1 and Chest Pain <=3) ➔ 0
ii. If (MHR<=3 and ST depression <=2.1 and ca <=0) ➔0
iii. If (MHR<=3 and ST depression <=2.1 and Chest Pain >3 and ca >0 and serum cholesterol <=282) ➔ 1
iv. If (MHR<=3 and ST depression <=2.1) ➔ 0
v. If (MHR<=3 and ST depression >2.1 and exercise induced angina <=0 and slope of the peak exercise ST segment <=2) ➔ 1
vi. If (MHR<=3 and exercise induced angina <=0) ➔ 0
vii. If (MHR<=3 and ca <=0) ➔ 0
viii. If (MHR<=3 and ST depression >2.1 and exercise induced angina >0 and ca >0) ➔ 3
ix. If (MHR>3 and Sex <=0 and serum cholesterol <=295) ➔ 3
x. If (MHR>3 and serum cholesterol >295) ➔ 1
xi. If (ST depression <=2.4 and fasting blood sugar <=0 and MHR<=6) ➔ 0
xii. If (MHR>3 and fasting blood sugar <=0 and MHR<=6 and resting blood pressure >135) ➔ 2
xiii. If (fasting blood sugar <=0 and MHR>6 and serum cholesterol <=205 and Age<=54) ➔ 3
xiv. If (Sex >0 and fasting blood sugar <=0 and MHR>6 and serum cholesterol <=205 and Age>54) ➔ 2
xv. If (ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and Age<=53) ➔ 1
xvi. If (ST depression <=2.4 and MHR>6 and resting blood pressure <=114 and Age>53) ➔ 2
xvii. If (ST depression <=2.4 and resting blood pressure >114 and serum cholesterol <=266 and ca <=0) ➔ 0
xviii. If (MHR>3 and ST depression <=2.4 and fasting blood sugar <=0 and serum cholesterol >205 and resting blood pressure >114 and serum cholesterol <=266 and ca >0 and Age<=61) ➔ 1
xix. If (ST depression <=2.4 and serum cholesterol <=266 and Age>61) ➔ 0
xx. If (ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting electrocardiographic<=1) ➔ 1
xxi. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and serum cholesterol >266 and resting electrocardiographic>1) ➔3
xxii. If (MHR>3 and Sex >0 and ST depression >2.4 and MHR>124 and Chest Pain >3 and exercise induced angina >0) ➔ 2
xxiii. If (MHR>3 and Sex >0 and MHR>124 and Chest Pain >3 and resting electrocardiographic>1 and resting blood pressure <=142) ➔1

5.4 *Classified Rules:*

Rule Set 0:
i. If (MHR<=3 and ST depression <=2.1 and Chest Pain <=3) ➔ 0 ➔t:10.1312574862847
ii. If (MHR<=3 and ST depression <=2.1 and ca <=0) ➔0 ➔t:10.1312574862847
iii. If (MHR<=3 and ST depression <=2.1) ➔ 0 ➔t:6.762569063545247
iv. If (MHR<=3 and exercise induced angina <=0) ➔0 ➔t:3.5897011441351196
v. If (MHR<=3 and ca <=0) ➔ 0 ➔t:4.1217701197746
vi. If (ST depression <=2.4 and fasting blood sugar <=0 and MHR<=6) ➔0 ➔t:10 31158272161852

vii. If (ST depression <=2.4 and resting blood pressure >114 and serum cholesterol <=266 and ca <=0) ➔ 0 ➔t:16.840367272392715
viii. If (ST depression <=2.4 and serum cholesterol <=266 and Age>61) ➔0 ➔t:13 9746836919829
ix. If (fasting blood sugar >0 and exercise induced angina <=0 and resting blood pressure <=156) ➔ 0 ➔t:9.75855933260464

Rule Set 1:

i. If (MHR<=3 and ST depression <=2.1 and Chest Pain >3 and ca >0 and serum cholesterol <=282) ➔ 1 ➔t:19.648509913498945
ii. If (MHR<=3 and ST depression >2.1 and exercise induced angina <=0 and slope of the peak exercise ST segment <=2) ➔ 1 ➔t:12.888640199119738
iii. If (MHR>3 and serum cholesterol >295) ➔ 1 ➔t:8.130109337967596
iv. If (ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and Age<=53) ➔ 1
v. ➔t:20.28699049907931
vi. If (MHR>3 and ST depression <=2.4 and fasting blood sugar <=0 and serum cholesterol >205 and resting blood pressure >114 and serum cholesterol <=266 and ca >0 If (ST depression <=2.4 and fasting blood sugar <=0 and MHR>6 and serum cholesterol >205 and resting electrocardiographic<=1) ➔1 ➔t:19.575519434255018

Rule Set 2:

i. If (MHR>3 and fasting blood sugar <=0 and MHR<=6 and resting blood pressure >135) ➔2➔t:12.619989401751669
ii. If (Sex >0 and fasting blood sugar <=0 and MHR>6 and serum cholesterol <=205 and Age>54) ➔2➔t:19.4758425613074
iii. If (ST depression <=2.4 and MHR>6 and resting blood pressure <=114 and Age>53) ➔ 2 ➔t:15.441499694596182
iv. If (ST depression <=2.4 and fasting blood sugar >0 and exercise induced angina >0) ➔ 2 ➔t:9.979316860749648
v. If (MHR>3 and Sex >0 and ST depression >2.4 and MHR>124 and Chest Pain >3 and exercise induced angina >0) ➔ 2 ➔t:21.681708712080138

Rule Set 3:

i. If (MHR<=3 and ST depression >2.1 and exercise induced angina >0 and ca >0) ➔3 ➔t:12.952068134551817
ii. If (MHR>3 and Sex <=0 and serum cholesterol <=295) ➔3 ➔t:12 26913357183703
iii. If (fasting blood sugar <=0 and MHR>6 and serum cholesterol <=205 and Age<=54) ➔ 3 ➔t:16.18425003790777
iv. If (MHR>3 and Sex >0 and ST depression <=2.4 and fasting blood sugar <=0 and serum cholesterol >266 and resting electrocardiographic>1) ➔ 3 ➔t:22.695159196669785
v. If (MHR>3 and resting blood pressure >156) ➔3 ➔t:6.612535146046203
vi. If (Sex >0 and ST depression >2.4 and MHR<=124) ➔ 3 ➔t:12 87169988355576

Rule Set 4:

i. If (MHR>3 and ST depression >2.4 and MHR>124 and Chest Pain <=3) ➔ 4 ➔t:15.570914546225168
ii. If (MHR>3 and Sex >0 and ST depression >2.4 and exercise induced angina <=0) ➔ 4 ➔t:12.823786323183375

5.5 Polish:

Class wise Rule distribution
0- Rule Set: 3

i. If (MHR>3 and Sex <=0 and serum cholesterol <=295) ➔ 3
ii. If (Sex >0 and ST depression >2.4 and MHR<=124) ➔ 3

1- Rule Set: 2

i. If (ST depression <=2.4 and fasting blood sugar >0 and exercise induced angina >0) ➔ 2

2- Rule Set: 4

i. If (MHR>3 and Sex >0 and ST depression >2.4 and exercise induced angina <=0) ➔ 4

3- Rule Set: 0

i. If (MHR<=3 and ST depression <=2.1 and Chest Pain <=3) ➔ 0

ii. If (MHR<=3 and ca <=0) ➔ 0

iii. If (fasting blood sugar >0 and exercise induced angina <=0 and resting blood pressure <=156) ➔ 0

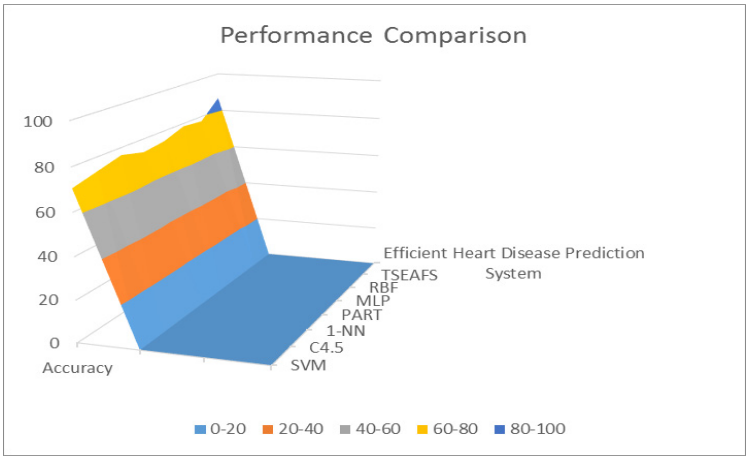
4- Rule Set: 1

i. If (MHR>3 and serum cholesterol >295) ➔ 1

6. Performance Evaluation

The performance of various well known algorithms on Heart Disease data set [12] is listed in Table 1 and it shows that Efficient Heart Disease Prediction System have the better accuracy than other given classifiers.

The Algorithm Used	SVM	C4.5	1-NN	PART	MLP	RBF	TSEAFS	**Efficient Heart Disease Prediction System**
Accuracy (%)	70.59	73.53	76.47	73.53	74.85	78.53	77.45	**86.7**



7. Conclusion

In this research paper, we have presented an Efficient Heart Disease Prediction System using data mining. This system can help medical practitioner in efficient decision making based on the given parameter. We have train and test the system using 10 fold method and find the accuracy of 86.3 % in testing phase and 87.3 % in training phase and because this model demonstrates the better results and helps the area specialists and even individual related with the field to get ready for a superior determine and give the patient to have early determination results as it performs sensibly well even without retraining.

References

1. Biafore, S. (1999). Predictive solutions bring more power to decision makers. Health Management Technology, 20(10), 12-14.
2

2. Silver, M. Sakata, T. Su, H.C. Herman, C. Dolins, S.B. & O'Shea, M.J. (2001). Case study: how to apply data mining techniques in a healthcare data warehouse. Journal of Healthcare Information Management, 15(2), 155-164.
3. Benko, A. & Wilson, B. (2003). Online decision support gives plans an edge. Managed Healthcare Executive, 13(5), 20
4. Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819:
5. Cody, W.F. Kreulen, J.T. Krishna, V. & Spangler, W.S. (2002). The integration of business intelligence and knowledge management. IBM Systems Journal, 41(4), 697-713
6. Ceusters, W. (2001). Medical natural language understanding as a supporting technology for data mining in healthcare. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), PhysicaVerlag Heidelberg, New York, 41-69.
7. Megalooikonomou, V. & Herskovits, E.H. (2001). Mining structure function associations in a brain image database. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), Physica-Verlag Heidelberg, New York, 153-180.
8. Chhikara, S & Sharma,P Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases, I JRASET 2014,PP 396-402.
9. 12Tallón-s, Antonio J., César Hervás- Martínez, JoséC. Riquelme, and Roberto Ruiz. (2013)"Feature selection to enhance a two-stage evolutionary algorithm in product unit neural networks for complex classification problems",Neurocomputing J.R. Quinlan. 1993, C4.5: Programs for Machine Learning. Morgan Kauffman Publishers, San Mateo-California.
10. J.R. Quinlan. 1995, MDL and Categorical Theories (Continued). In Machine Learning: Proceedings of the Twelfth International Conference. Lake Tahoe, California. Morgan Kaufmann, , 464-470.
11. Eibe Frank and Ian H. Witten. 1998 Generating accurate rule sets without global optimization. In Proc 15th International Conference on Machine Learning, Madison, Wisconsin, pages 144-151. Morgan Kaufmann.
12. Heart attack dataset from http://archive.ics.uci.edu/ml/datasets/Heart Disease
13. J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, 2009 KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 307-318
14. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3 (2011) 255-287
15. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten 2009; The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
16. Sharma Purushottam, Dr Kanak Saxena, Richa Sharma" Efficient Heart Disease Prediction System using Decision Tree" in IEEE International Conference on Computing Communication and Automation (ICCCA-2015),May 2015
17. Sharma Purushottam, Dr Kanak Saxena, Richa Sharma" Heart Disease Prediction System Evaluation Using C4.5 Rules and Partial Tree" in Springer, Computational Intelligence in Data Mining, 2015, pp-285-294, DOI 10.1007/978-81-322-2731-1_26.

applied sciences

MDPI

Article

Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators

Karna Vishnu Vardhana Reddy [1], Irraivan Elamvazuthi [1,*], Azrina Abd Aziz [1], Sivajothi Paramasivam [2], Hui Na Chua [3] and S. Pranavanand [4]

[1] Department of Electrical and Electronics Engineering, Universiti Teknologi PETRONS, Seri Iskandar 32610, Malaysia; vishnu_17009417@utp.edu.my (K.V.V.R.); azrina_aaziz@utp.edu.my (A.A.A.)
[2] School of Engineering, UOWM KDU University College, Shah Alam 40150, Malaysia; siva@kdu.edu.my
[3] Department of Computing and Information Systems, School of Engineering, and Technology, Sunway University, Petaling Jaya 47500, Malaysia; huinac@sunway.edu.my
[4] Department of E.I.E., VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad 500090, India; pranavanand_s@vnrvjiet.in
* Correspondence: irraivan_elamvazuthi@utp.edu.my

Abstract: Cardiovascular diseases (CVDs) kill about 20.5 million people every year. Early prediction can help people to change their lifestyles and to ensure proper medical treatment if necessary. In this research, ten machine learning (ML) classifiers from different categories, such as Bayes, functions, lazy, meta, rules, and trees, were trained for efficient heart disease risk prediction using the full set of attributes of the Cleveland heart dataset and the optimal attribute sets obtained from three attribute evaluators. The performance of the algorithms was appraised using a 10-fold cross-validation testing option. Finally, we performed tuning of the hyperparameter number of nearest neighbors, namely, 'k' in the instance-based (IBk) classifier. The sequential minimal optimization (SMO) achieved an accuracy of 85.148% using the full set of attributes and 86.468% was the highest accuracy value using the optimal attribute set obtained from the chi-squared attribute evaluator. Meanwhile, the meta classifier bagging with logistic regression (LR) provided the highest ROC area of 0.91 using both the full and optimal attribute sets obtained from the ReliefF attribute evaluator. Overall, the SMO classifier stood as the best prediction method compared to other techniques, and IBk achieved an 8.25% accuracy improvement by tuning the hyperparameter 'k' to 9 with the chi-squared attribute set.

Keywords: heart disease; data pre-processing; attribute evaluation; machine learning classifiers; hyperparameter tuning

Citation: Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. *Appl. Sci.* **2021**, *11*, 8352. https://doi.org/10.3390/app11188352

Academic Editor: Giancarlo Mauri

Received: 27 July 2021
Accepted: 1 September 2021
Published: 9 September 2021

1. Introduction

Cardiovascular disease (CVD) is the biggest concern in the medical sector at present. It is one of the most lethal and chronic diseases, leading to the highest number of deaths worldwide. From the recent statistics reported by World Health Organization (WHO), about 20.5 million people die every year due to cardiovascular disease, which is approximately 31.5% of all deaths globally. It is also estimated that the number of annual deaths will rise to 24.2 million by 2030. About 85% of cardiovascular disease deaths are due to heart attack and strokes [1]. A heart attack is mainly caused when the blood flow to the heart is blocked due to the build-up of plaque in the arteries. Stroke is caused by a blood clot in an artery within the brain, which cuts off blood circulation to the brain [2]. Heart disease is triggered mostly when the heart is unable to provide enough blood supply to parts of the body [3,4]. It results in early symptoms, such as an irregular heartbeat, shortness of breath, chest discomfort, sudden dizziness, nausea, swollen feet, and a cold sweat. The accurate prediction and proper diagnosis of heart disease in time are indispensable for improving the survival rate of patients. The risk factors that cause CVD include high BP, cholesterol,

alcohol intake, and tobacco consumption, as well as obesity, physical inactivity, and genetic mutations. The early detection of signs and changes in lifestyle, such as physical activity, avoiding smoking, and appropriate medical examination by clinicians, can help to reduce mortality [5].

The techniques that are currently used to predict and diagnose heart disease are primarily based on the analysis of a patient's medical history, symptoms, and physical examination reports by doctors. Most of the time, it is difficult for medical experts to accurately predict a patient's heart disease, where they can predict with up to 67% accuracy [6] because, currently, the diagnosis of any disease is done concerning the similar symptoms observed from previously diagnosed patients [7]. Hence, the medical field requires an automated intelligent system for the accurate prediction of heart disease. This can be achieved by utilizing the huge amount of patient data that is available in the medical sector, along with machine learning algorithms [8]. In recent times, data science research groups have paid much attention to disease prediction. This is owing to the rapid development of advanced computer technologies in the healthcare sector, as well as the availability of massive health databases [9]. The combination of new deep-learning and intelligent decision-making systems has great potential to improve healthcare assistance in our society [10]. Data is the most valuable resource for obtaining new or additional knowledge and collecting important information. There is an enormous amount of data (big data) in various sectors, such as science, technology, agriculture, business, education, and health. This is completely unprocessed data, either in a structured or unstructured form [11]. It is necessary to extract valuable information from big data to store, process, analyze, manage, and visualize this data via performing data analysis [12].

Currently, in the healthcare sector, the information that is related to patients with medical reports is readily available in databases and is growing rapidly day by day. This raw data is highly redundant and unbalanced. It requires pre-processing to extract important features, reduce the execution time of training algorithms, and improve the classification efficiency [13]. The latest advancements in computing capacities and reprogramming capabilities of machine learning improve these processes and open doors for research opportunities in the healthcare sector [14], especially regarding the early prediction of the diseases, such as CVD and cancer, to improve the survival rate. Machine learning is used in a wide range of applications, from identifying risk factors for disease to designing advanced safety systems for automobiles. Machine learning offers predominant prediction modeling tools to address the current limitations [15]. It has good potential for transforming big data for prediction algorithm development. It relies on a computer to learn complex and non-linear interactions between attributes by minimizing the error between the predicted and observed outcomes [16]. The machine learns patterns from the features that are available in the existing dataset and applies them to the unknown dataset to predict the outcome. One of the powerful machine learning techniques for prediction is classification. Classification is a supervised machine learning method that is effective at identifying the disease when trained using appropriate data [17].

The main contribution of this research work was to implement an intuitive medical prediction system for the diagnosis of heart disease using contemporary machine learning techniques. In this work, different kinds of machine learning classifier algorithms, such as naïve Bayes (NB), logistic regression (LR), sequential minimal optimization (SMO), instance-based classifier (IBk), AdaBoostM1 with decision stump (DS), AdaBoostM1 with LR, bagging with REPTree, bagging with LR, JRip, and random forest (RF) were trained to select the best predictive model for the accurate heart disease detection at an initial stage. Three attribute selection techniques, such as correlation-based feature subset evaluator, chi-squared attribute evaluator, and ReliefF attribute evaluator, were utilized to obtain the optimal set of attributes that greatly influenced the performance of the classifiers when predicting the target class. Finally, tuning the hyperparameter "number of nearest neighbors" in the IBk classifier was performed on both the full attribute set and optimal sets obtained from attribute evaluators.

2. Related Works

This section discusses the state-of-the-art methods for heart disease diagnosis using machine learning techniques that were accomplished by various effective research works.

R. Perumal et al. [18] developed a heart disease prediction model using the Cleveland dataset of 303 data instances through feature standardization and feature reduction using PCA, where they identified and utilized seven principal components to train the ML classifiers. They concluded that LR and SVM provided almost similar accuracy values (87% and 85%, respectively) compared to that of k-NN with 69%. C. B. C. Latha et al. [19] performed a comparative analysis to improve the predictive accuracy of heart disease risk using ensemble techniques on the Cleveland dataset of 303 observations. They applied the brute force method to obtain all possible attribute set combinations and trained the classifiers. They achieved a maximum increase in the accuracy of a weak classifier of 7.26% based on ensemble algorithm, and produced an accuracy of 85.48% using majority vote with NB, BN, RF, and MLP classifiers using an attribute set of nine attributes. D. Ananey-Obiri et al. [20] developed three classification models, namely, LR, DT, and Gaussian naïve Bayes (GNB), for heart disease prediction based on the Cleveland dataset. Feature reduction was performed using single value decomposition, which reduced the features from 13 to 4. They concluded that both LR and GNB had predictive scores of 82.75% and AUC of 0.87. It was suggested that other models, such as SVM, k-NN, and random forest, be included.

N. K. Kumar et al. [21] trained five machine learning classifiers, namely, LR, SVM, DT, RF, and KNN, using a UCI dataset with 303 records and 10 attributes to predict cardiovascular disease. The RF classifier achieved the highest accuracy of 85.71% with an ROC AUC of 0.8675 compared to the other classifiers. A. Gupta et al. [22] replaced the missing values based on the majority label and derived 28 features using the Pearson correlation coefficient from the Cleveland dataset and trained LR, KNN, SVM, DT, and RF classifiers using the factor analysis of mixed data (FAMD) method; the results based on a weight matrix RF achieved the best accuracy of 93.44%. M. Sultana et al. [23] explored KStar, J48, sequential minimal optimization (SMO), BN, and MLP classifiers using Weka on a standard heart disease dataset from the UCA repository with 270 records and 13 attributes; they achieved the highest accuracy of 84.07% with SMO.

S. Mohan et al. [24] developed an effective hybrid random forest with a linear model (HRFLM) to enhance the accuracy of heart disease prediction using the Cleveland dataset with 297 records and 13 features. They concluded that the RF and LM methods provided the best error rates. S. Kodati et al. [25] developed a heart disease prediction system (HDPS) with the Cleveland dataset of 297 instances and 13 attributes using Orange and Weka data mining tools, where they evaluated the precision and recall metrics for the naïve Bayes, SMO, RF, and KNN classifiers. A. Ed-daoudy et al. [26] researched the Cleveland dataset of 303 records and 14 attributes from UCI. They evaluated the performance of the four main classifiers, namely, SVM, DT, RF, and LR, using Apache Spark with its machine learning library MLlib.

I. Tougui et al. [27] compared the performances of LR, SVM, KNN, ANN, NB, and RF models to classify heart disease with the Cleveland dataset with 297 observations and 13 features using six data mining tools: Orange, Weka, RapidMiner, Knime, MATLAB, and Scikit-Learn. V. Pavithra et al. [28] proposed a new hybrid feature selection technique with the combination of random forest, AdaBoost, and linear correlation (HRFLC) using the UCI dataset of 280 instances to predict heart disease. Eleven (11) features were selected using filter, wrapper, and embedded methods; an improvement of 2% was found for the accuracy of the hybrid model. C. Gazeloglu et al. [29] projected 18 machine learning models and 3 feature selection techniques (correlation-based FS, chi-square, and fuzzy rough set) to find the best prediction combination for heart disease diagnosis using the Cleveland dataset of 303 instances and 13 variables.

N. Louridi et al. [30] proposed a solution to identify the presence/absence of heart disease by replacing missing values with the mean values during pre-processing. They trained three machine learning algorithms, namely, NB, SVM (linear and radial basis func-

tion), and KNN, by splitting the Cleveland dataset of 303 instances and 13 attributes into 50:50, 70:30, 75:25, and 80:20 training and testing ratios. M. Kavitha et al. [31] implemented a novel hybrid model on the Cleveland heart dataset of 303 instances and 14 features with a 70:30 ratio for training and testing by applying DT, RF, and hybrid (DT + RF) algorithms. B. A. Tama et al. [32] designed a stacked architecture to predict heart disease using RF, gradient boosting machine, and extreme gradient boosting with particle swarm optimization (PSO) feature selection using various heart disease datasets, including the Cleveland with 303 instances and 13 attributes.

From the experimental works, it is understood that data pre-processing and feature selection can substantially enhance the classification accuracy of machine learning algorithms. During pre-processing, most researchers [18,19,21,22,26,29–32] replaced the missing values, either by using the mean value or the majority mark of that attribute, to make sure the dataset was comprehensive. In some works [20,24,25,27], the missing valued instances were removed. Feature selection is a challenging task due to the large exploration space. It grows exponentially according to the number of features available in the dataset. To solve this issue, an effective comprehensive search technique is required during feature selection. Furthermore, some studies have employed ensemble models, which combine multiple basic learning algorithms to obtain a better prediction accuracy. However, the performance of these techniques can further be improved regarding accurately predicting disease.

3. Materials and Methods

This section discusses the proposed methodology, which comprises the dataset description, data pre-processing, machine learning classifiers, attribute evaluators, and performance metrics.

3.1. Proposed Research Methodology

The experimental workflow of the proposed methodology is shown in Figure 1. As a first step, we collected the Cleveland heart disease dataset in .csv format from the UCI machine learning repository. Then, we imported the dataset into the software tool and explored the attributes, types, value ranges, and other statistical information. The next step was pre-processing the data, which included tasks such as looking for the missing values in the dataset and replacing missing values, either with the user constant or mean value depending on the type of attribute, to make sure the machine learning classifiers provide better performance. Thereafter, classification was performed with cross-validation using several machine learning algorithms, such as NB, LR, SMO, IBk, AdaBooostM1 + DS, AdaBoostM1 + LR, bagging + REPTree, bagging + LR, JRip, and RF using the full set of attributes. Cross-validation is a resampling method that is used to assess the efficacy of the machine learning model by partitioning the original dataset into a training set to train the model and a test set to evaluate it. The observations in a dataset can be randomly split into k equal-sized groups. We then trained the model using $k-1$ folds and validated the models using the remaining kth fold. We repeated this step until all k folds served as a test set and took the average of the recorded values as the performance metric of the model. This work considered k = 10, i.e., a 10-fold cross-validation. Further, we applied attribute evaluators, such as correlation-based feature selection with the BestFirst search method, chi-squared attribute evaluation with Ranker, and ReliefF attribute evaluation with Ranker using a full training set to obtain the optimal set of attributes for predicting heart disease risk and trained the classifiers again using cross-validation. Finally, we tuned the hyperparameter 'k' in the IBk classifier for enhanced performance and analyzed the results.

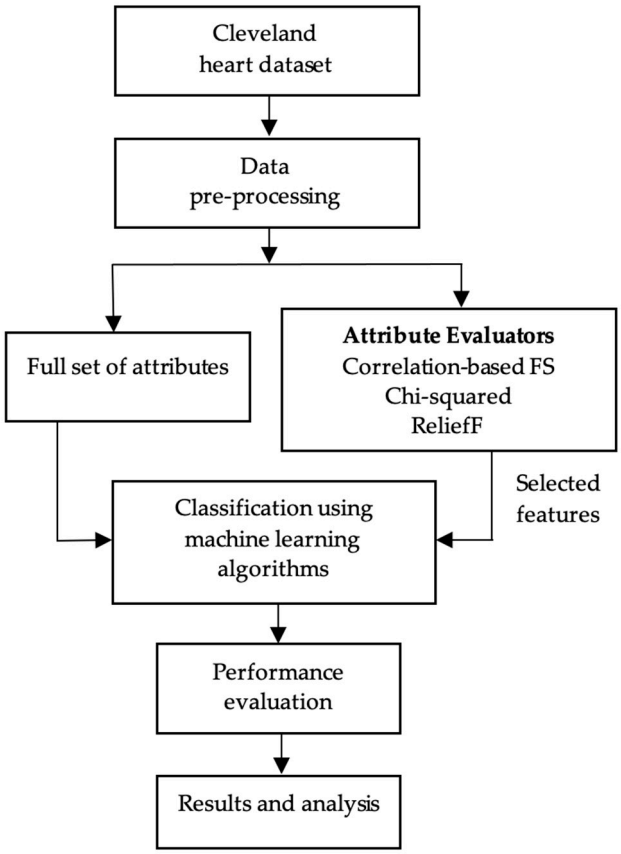


Figure 1. Experimental workflow of the proposed methodology.

3.2. Dataset Description and Statistics

The Cleveland heart dataset consists of 303 instances with 76 attributes, but only 14 attributes are considered more suitable for research experimental purposes. The attribute descriptions for the Cleveland heart dataset are given in Table 1.

Table 1. Attribute descriptions for the Cleveland heart dataset from the UCI machine learning repository [33].

Attribute	Description	Type of Attribute	Attribute Value Range
age	Age in years	Numeric	29 to 77
sex	Gender	Nominal	0 = female, 1 = male
cp	Chest pain type	Nominal	1 = typical angina, 2 = atypical angina, 3 = non-angina pain, 4 = asymptomatic
trestbps	Resting blood pressure in mm Hg on admission to the hospital	Numeric	94 to 200
chol	Serum cholesterol in mg/dL	Numeric	126 to 564
fbs	Fasting blood sugar > 120 mg/dL	Nominal	0 = false, 1 = true

Table 1. *Cont.*

Attribute	Description	Type of Attribute	Attribute Value Range
restecg	Resting electrocardiographic results	Nominal	0 = normal, 1 = ST-T wave abnormality, 2 = definite left ventricular hypertrophy by Estes' criteria
thalach	Maximum heart rate achieved	Numeric	71 to 202
exang	Exercise induces angina	Nominal	0 = no 1 = yes
oldpeak	ST depression induced by exercise relative to rest	Numeric	0 to 6.2
slope	The slope of the peak exercise ST segment	Nominal	1 = upsloping, 2 = flat, 3 = downsloping
ca	Number of major vessels colored by fluoroscopy	Nominal	0–3
thal	The heart status	Nominal	3 = normal, 6 = fixed defect, 7 = reversible defect
target	Prediction attribute	Nominal	0 = no risk of heart disease, 1 to 4 = risk of heart disease

The attributes with less than 10 classes are considered nominal or categorical types. The attribute 'sex' consists of two classes based on gender: 1 = male and 0 = female. The attribute 'cp' contains four classes of chest pain types: 1 = typical angina, 2 = atypical angina, 3 = non-angina pain, and 4 = asymptomatic. The attribute 'fbs' includes two classes regarding whether the fasting blood sugar >120 mg/dL: 1 = true and 0 = false. The attribute 'restecg' comprises three classes of resting electrocardiographic results: 0 = normal, 1 = abnormality in the ST-T wave, 2 = definite hypertrophy in the left ventricular. The attribute 'exang' consists of two classes based on exercise-induced angina: 1 = yes and 0 = no. The attribute 'slope' includes three classes of peak exercise ST segment slope: 1 = upslope, 2 = flat, and 3 = downslope. The attribute 'ca' comprises four classes based on the number of major vessels (0–3) that are colored using fluoroscopy. The attribute 'thal' contains three classes of heart status: 3 = normal, 6 = fixed, and 7 = reversible. The attribute 'target' consists of five classes of prediction: 0 = no risk of heart disease, and 1 to 4 = the risk of heart disease in various stages. Since the main purpose of this research work was to predict whether a patient was at risk of developing heart disease, the values in the range 1 to 4 were converted to 1. Therefore, the 'target' attribute consisted of only two classes: 0 and 1. The attributes 'age,' 'trestbps,' 'chol,' 'thalach,' and 'oldpeak' are considered as numeric/integer type attributes.

The statistical characteristics of the numeric attributes, such as the minimum, maximum, mean, standard deviation, missing, distinct, and unique values, are provided in Table 2(a) There are no missing values found in the numeric attributes of the Cleveland dataset.

Table 2. (a) The statistical outline of the numeric attributes. (b) The statistical outline of the nominal attributes.

(a)							
Attribute	**Min.**	**Max.**	**Mean**	**StdDev**	**Missing**	**Distinct**	**Unique**
age	29	77	54.439	9.039	0	41	4 (1%)
trestbps	94	200	131.69	17.6	0	50	17 (6%)
chol	126	564	246.693	51.777	0	152	61 (20%)
thalach	71	202	149.607	22.875	0	91	28 (9%)
oldpeak	0	6.2	1.04	1.161	0	40	10 (3%)

(b)					
Attribute	**Label**	**Count**	**Proportion**	**Missing**	**Distinct**
sex	0 1	97 206	32% 68%	0	2
cp	1 2 3 4	23 50 86 144	7.6% 16.5% 28.4% 47.5%	0	4
fbs	0 1	258 45	85.15% 14.85%	0	2
restecg	0 1 2	151 4 148	49.83% 1.32% 48.84%	0	3
exang	0 1	204 99	67.33% 32.67%	0	2
slope	1 2 3	142 140 21	46.86% 46.20% 6.93%	0	3
ca	0 1 2 3	176 65 38 20	58.08% 21.45% 12.54% 6.6%	4 (1.32%)	4
thal	3 6 7	166 18 117	54.79% 5.95% 38.6%	2 (0.66%)	3
target	0 1	164 139	54% 46%	0	2

Min.—minimum, Max.—maximum, StdDev—standard deviation.

The statistical characteristics of the nominal attributes, such as label, count, missing, and distinct values, are provided in Table 2(b). There are six (6) instances in total out of 303 that were found to have missing values, which accounted for 2% of the whole dataset: four (4) from the 'ca' attribute, and two (2) from the 'thal' attribute. The target class labels 0 (no risk of heart disease) consisted of 164 instances and label 1 (risk of heart disease) consisted of 139 instances, which accounted for 54% and 46% of the dataset, respectively.

3.3. Pre-Processing of Dataset

Having missing data means that the dataset is incomplete. In statistics, missing values or missing data occur when no data value is stored for the variable in an observation. These missing values are represented by blank/dashes. The main reason for having missing values is that respondents forget/refuse/fail to answer certain questions. Other reasons include sensor failure, loss of data while transferring, internet connection disruption, and wrong mathematical calculations, such as dividing by zero. It is always hard to predict

when missing values are present in the dataset because sometimes, they affect results and sometimes not. In a dataset, each variable may only have a small number of missing responses, but in combination, the missing data could be numerous. The analysis might run but the results may not be statistically significant because of the missing data. For research purposes, replacing missing values either by a user constant or the mean value will be more effective than removing those observations from the dataset. There are some missing values in the Cleveland heart dataset, namely, from the nominal attributes 'ca' and 'thal', which were replaced with the user constant based on the majority mark. The attribute 'ca' has four missing values and has the value 0 as the majority mark in 176 observations out of 299. Meanwhile, the attribute 'thal' has two missing values and has the value 3 as the majority mark in 166 observations out of 301. Therefore, to make sure the dataset is complete, the missing values in 'ca' and 'thal' were replaced by the corresponding majority marks 0 and 3, respectively. A visualization of all 14 attributes of the Cleveland heart dataset is presented in Figure 2.

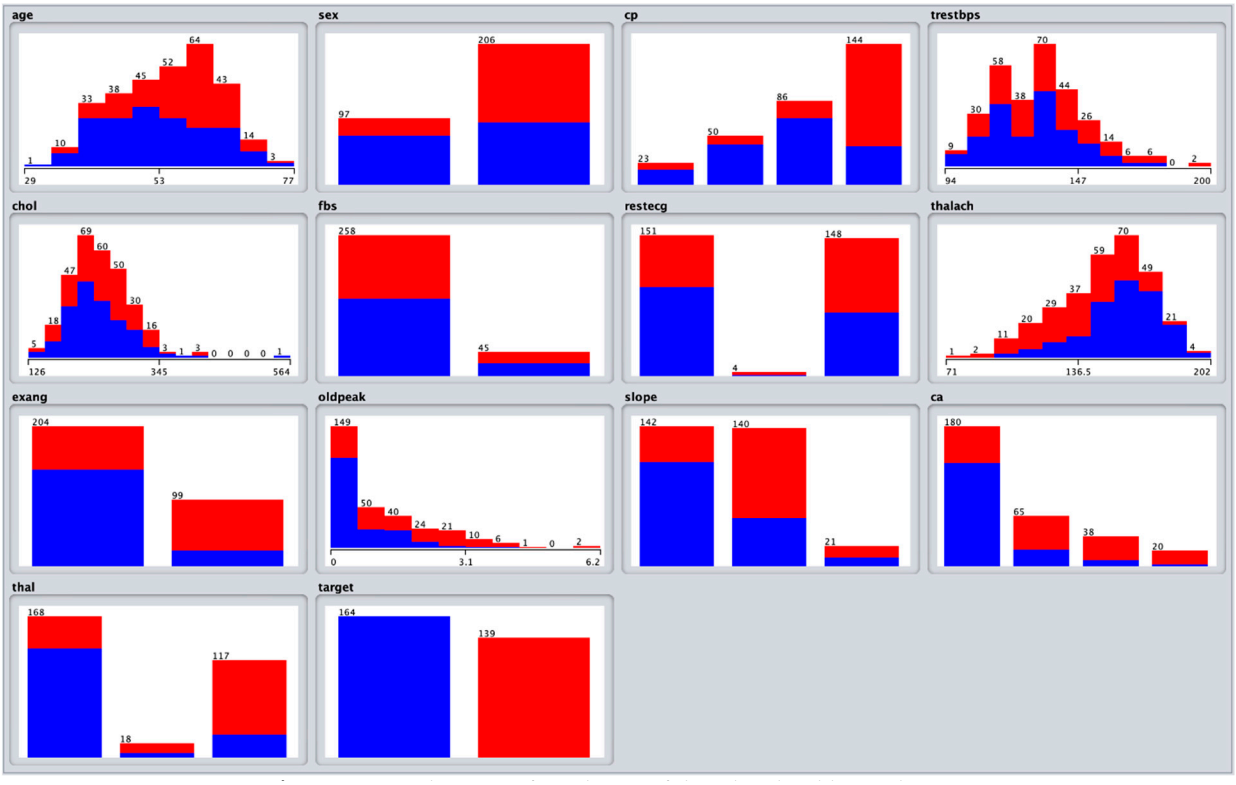


Figure 2. Visualization of attributes of the Cleveland heart dataset.

3.4. Machine Learning Models for Classification

Researchers have applied multiple supervised machine learning algorithms on a single dataset to identify the best classifier for disease prediction. This section discusses the various classifiers that were used in this work to predict heart disease risk.

Naïve Bayes (NB) is based on the Bayes theorem, which assumes that the training observations are samples from a set of statistical distributions. Each response class has its distribution. Each distribution in the model provides a probability that a new data point would be found at its location. For the normal distribution, the parameters are the mean and standard deviation [34].

Logistic regression (LR) is an equation where each predictor is multiplied by a coefficient and summed together. This sum becomes the argument for the logistic function to predict the class [35]. For a single observation x with n features the response y is given by

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (1)$$

Sequential minimal optimization (SMO) is an algorithm that is used to solve very large quadratic programming (QP) optimization problems quickly without any extra storage requirement while training a support vector machine (SVM) [23]. SMO selects two Lagrange multipliers and analytically finds the optimal values for these multipliers to solve the SVM QP problem [36].

The instance-based classifier IBk, also known as k-nearest neighbors, determines the class of observation by comparing it to nearby observations from the training data set. The distance measure that is used to determine the neighbors can be selected from a range of options. The model uses majority voting from the K nearest data points and assigns a class to the unknown observation [37].

Bagging, also known as the bootstrap aggregation technique, is a simple and powerful ensemble technique that is used to decrease the variance of the decision tree classifier [37]. It provides the learning algorithm with a training set consisting of a random sample of m training examples that are selected from the initial training set of m items on each run.

Boosting is an example of an ensemble technique that creates a robust classifier from several weak classifiers. Adaptive boosting (AdaBoostM1) is a successful boosting algorithm that was developed for binary classification and is used to boost the performance of any machine learning algorithm. The decision tree with one level or one decision for classification, called the decision stump, is the most suitable algorithm to work with AdaBoost [38].

JRip is a rule-based classifier that utilizes repeated incremental pruning to produce error reduction (RIPPER). It is a bottom-up approach for learning rules that treats specific judgments of examples in the training data as a class and finds a set of rules that covers all members of the class [39].

Random forest constructs a forest of random trees by creating a set of decision trees from a random sample of the training set to minimize the variance at the expense of a small increase in bias (controlling over-fitting) and results in a final prediction model that should be more accurate and reliable. While growing the trees, the random forest adds more randomness to the algorithms by using random thresholds for each attribute [40].

3.5. Attribute Evaluators

Three attribute evaluators correlation-based feature selection with the BestFirst search method, chi-squared attribute evaluation with Ranker, and ReliefF attribute evaluation with Ranker are used in this work.

The correlation-based feature selection technique considers the individual predictive capacity of each attribute, as well as the degree of redundancy between them when determining the value of a subset of attributes. The subsets of attributes with a low intercorrelation but high correlation with the class are preferred [29]. Table 3 shows the attribute set that was obtained from the correlation-based feature selection method.

Table 3. Attribute sets that were obtained from the correlation-based feature selection.

Attribute No.	Attribute Name
3	cp
7	restecg
8	thalach
9	exang
10	oldpeak
12	ca
13	thal

The chi-squared attribute evaluation technique is an attribute ranking filter that computes the value of the chi-squared statistic with respect to the class to determine the rank of an attribute using the Ranker search method [40,41]. The rank values of the Cleveland attributes using chi-squared techniques are shown in Table 4. We created an attribute space with the 10 best predictor attributes by removing the three least ranked ones, namely, fbs, trestbps, and chol, from the dataset and trained the machine learning algorithms.

Table 4. Attribute sets that were obtained from the chi-squared attribute evaluation.

Attribute No.	Attribute Name	Rank
13	thal	82.6845
3	cp	81.8158
12	ca	72.6169
10	oldpeak	61.5234
9	exang	56.5193
8	thalach	51.5870
11	slope	45.7846
1	age	24.8856
2	sex	23.2181
7	restecg	10.0515
6	fbs	0.1934
4	trestbps	0
5	chol	0

The ReliefF attribute evaluation technique is also an attribute ranking filter that evaluates the value of an attribute by sampling an instance many times and comparing the value of the supplied attribute for the closest instances of the same and different classes [37]. It can work with data from both discrete and continuous classes. This method utilizes all the instances while sampling, the number of nearest neighbors k = 10, and the Ranker search method to provide the rank values [42] of the Cleveland attributes, which are recorded in Table 5. The classifiers were trained with the top nine attributes by discarding the four lowest-ranked attributes, namely, age, trestbps, fbs, and chol, from the dataset.

Table 5. Attribute sets that were obtained from the ReliefF attribute evaluation.

Attribute No.	Attribute Name	Rank
12	ca	0.18812
3	cp	0.17789
13	thal	0.11452
2	sex	0.09307
11	slope	0.06898
9	exang	0.06667
7	restecg	0.05842
10	oldpeak	0.02350
8	thalach	0.02118
1	age	0.01786
4	trestbps	0.01577
6	fbs	0.01386
5	chol	0.00181

3.6. Performance Metrics

The performance metrics used in this research work, namely, accuracy, mean absolute error (MAE), sensitivity (recall), fallout, precision, F-measure, specificity, and ROC area, are discussed here. The confusion matrix shown in Table 6 depicts various performance metrics for evaluating a classifier. True positives are the responses equal to the positive class that are correctly predicted as positive. True negatives are the responses equal to the negative class that are correctly predicted as negative. False positives are the responses equal to the negative class but are predicted as positive. False negatives are the responses equal to the positive class but are predicted as negative.

Table 6. Confusion matrix.

		Predicted Class High Risk (1)	Low Risk (0)
Actual class	High risk (1)	True Positive (TP)	False Negative (FN)
	Low risk (0)	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\% \quad (2)$$

$$\text{MAE} = \frac{\sum |\text{Predicted value} - \text{Actual value}|}{\text{Number of predictions}} \quad (3)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \quad (5)$$

$$\text{Fallout} = \frac{\text{FP}}{\text{TN} + \text{FP}} \times 100\% \quad (6)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \quad (7)$$

$$\text{F} - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

ROC Area: The area under the ROC curve measures the quality of a model's predictions regardless of what classification threshold is chosen. The ROC curve represents the

true positive rate (sensitivity or recall) vs. the false positive rate (fallout) at every $0 \rightarrow 1$ threshold.

4. Results

The results of the machine learning classifiers using the full set of attributes and optimal set that was obtained from attribute evaluators, tuning the parameter 'k' in the IBk method, and comparison with related works are discussed in the following.

As shown in Table 7, the highest accuracy of 85.148% was attained with the sequential minimal optimization (SMO) algorithm, followed by logistic regression (LR) with an accuracy of 84.818% using the full set of attributes from the Cleveland dataset using the 10-fold cross-validation test option. The SMO algorithm also provided the best MAE of 0.148, sensitivity of 0.851, fallout of 0.157, precision of 0.852, F-measure of 0.851, and specificity of 0.90 compared to other machine learning algorithms. The meta classifier bagging with LR achieved a high ROC area value of 0.91, followed by the single classifier LR with a ROC area of 0.909. The LR and AdaBoostM1 with LR classifiers also reached a specificity of 0.90. NB provided the second-best MAE of 0.184. The visualization of the threshold curve for the target class provided an ROC area of 0.91 using the bagging with LR meta classifier using the full set of attributes, as shown in Figure 3. The bar plot of the performance metrics using the full set of attributes of the Cleveland heart dataset is shown in Figure 4.

Table 7. Performance of machine learning classifiers based on the full set of attributes using 10-fold cross-validation.

Classifier	Accuracy	MAE	Sensitivity	Fallout	Precision	F-Measure	ROC Area	Specificity
NB	83.828	0.184	0.838	0.167	0.838	0.838	0.907	0.870
LR	84.818	0.210	0.848	0.162	0.85	0.847	0.909	0.900
SMO	85.148	0.148	0.851	0.157	0.852	0.851	0.847	0.900
IBk/KNN	76.897	0.233	0.769	0.235	0.769	0.769	0.764	0.790
AdaBoostM1 + DS	82.838	0.227	0.828	0.178	0.829	0.828	0.888	0.870
AdaBoostM1 + LR	84.818	0.204	0.848	0.162	0.850	0.848	0.860	0.900
Bagging + REPTree	80.858	0.290	0.809	0.198	0.809	0.809	0.878	0.850
Bagging + LR	84.488	0.214	0.845	0.162	0.845	0.845	0.910	0.880
JRip	74.917	0.325	0.749	0.256	0.749	0.749	0.755	0.780
RF	81.848	0.276	0.818	0.188	0.818	0.818	0.897	0.850

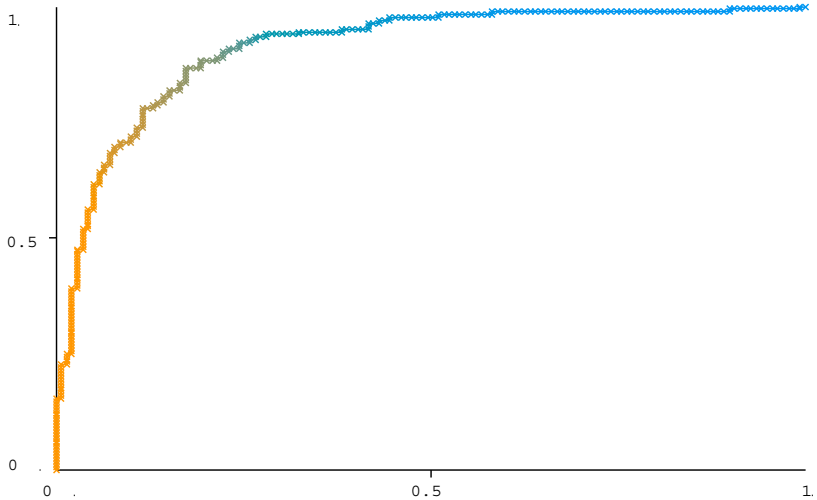


Figure 3. ROC curve of bagging with LR meta classifier using the full set of features.

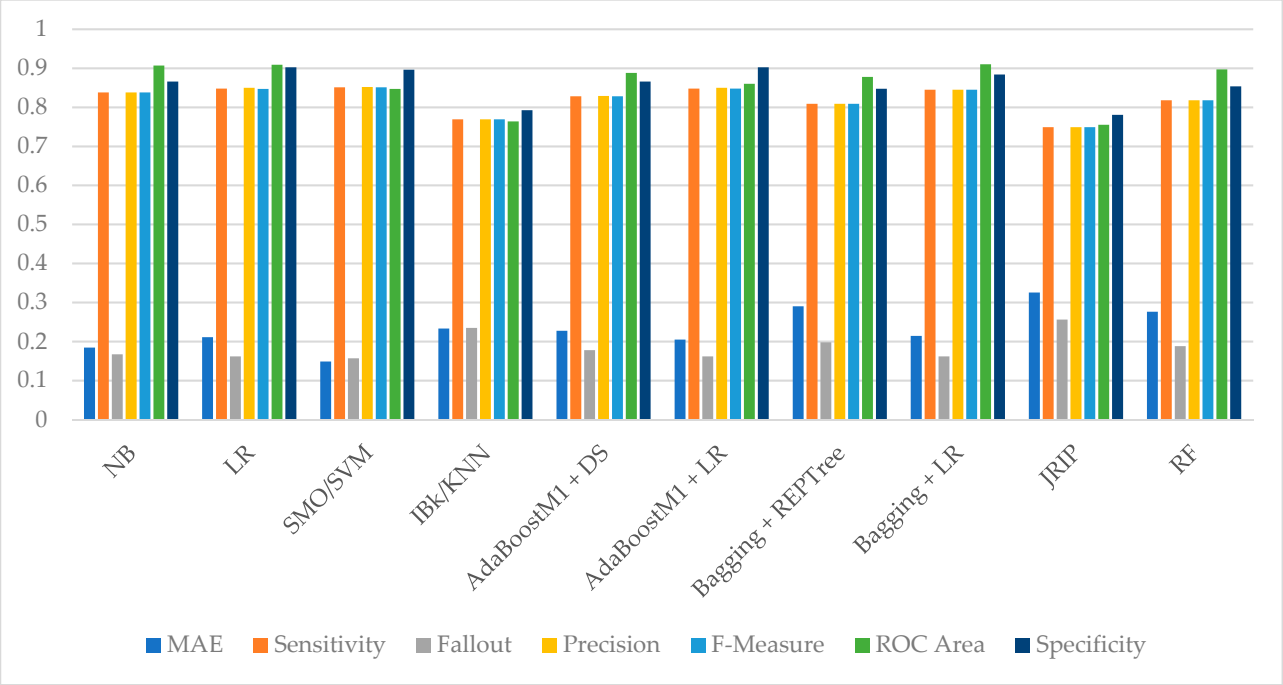


Figure 4. Performance metrics using the full set of attributes from the Cleveland heart dataset.

Table 8 shows that the highest accuracy of 84.158% was achieved with the naïve Bayes (NB) algorithm, followed by the SMO, AdaBoostM1 with decision stump (DS), and bagging + LR algorithms, with an accuracy of 83.828% when using the optimal attribute set that was obtained from the correlation-based feature selection method. The SMO algorithm provided the best MAE of 0.161, the NB classifier produced a high sensitivity of 0.842, precision of 0.843, F-measure of 0.841, ROC area of 0.905, and specificity of 0.90 compared to other algorithms. The meta classifier bagging with LR achieved the best fallout value of 0.167. The performance metrics of the ML classifiers with the optimal set obtained using correlation-based feature selection are graphically presented in Figure 5. The ROC curve of the naïve Bayes classifier using the correlation-based feature selection set provided with an area of 0.905 is shown in Figure 6.

Table 8. Performance of the machine learning classifiers using the optimal attribute set found based on the correlation-based feature selection technique.

Classifier	Accuracy	MAE	Sensitivity	Fallout	Precision	F-Measure	ROC Area	Specificity
NB	84.158	0.190	0.842	0.168	0.843	0.841	0.905	0.900
LR	83.168	0.227	0.832	0.176	0.832	0.831	0.902	0.870
SMO	83.828	0.161	0.838	0.169	0.839	0.838	0.835	0.880
IBk/KNN	78.877	0.218	0.789	0.219	0.789	0.788	0.781	0.830
AdaBoostM1 + DS	83.828	0.228	0.838	0.169	0.839	0.838	0.900	0.880
AdaBoostM1 + LR	83.168	0.236	0.832	0.176	0.832	0.831	0.817	0.870
Bagging + REPTree	81.188	0.276	0.812	0.197	0.812	0.811	0.886	0.860
Bagging + LR	83.828	0.229	0.838	0.167	0.838	0.838	0.902	0.870
JRip	74.257	0.332	0.743	0.269	0.743	0.741	0.750	0.800
RF	79.538	0.255	0.795	0.210	0.795	0.795	0.882	0.820

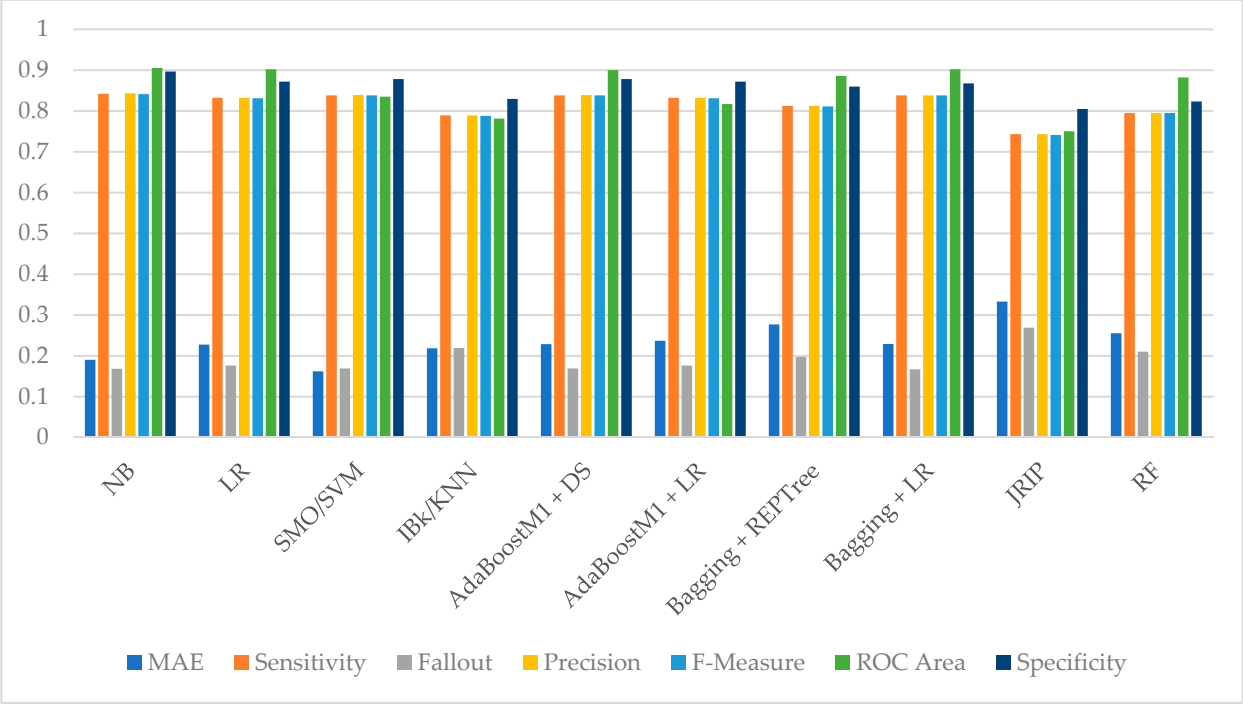


Figure 5. Performance metrics using the optimal set obtained from the correlation-based feature selection technique.

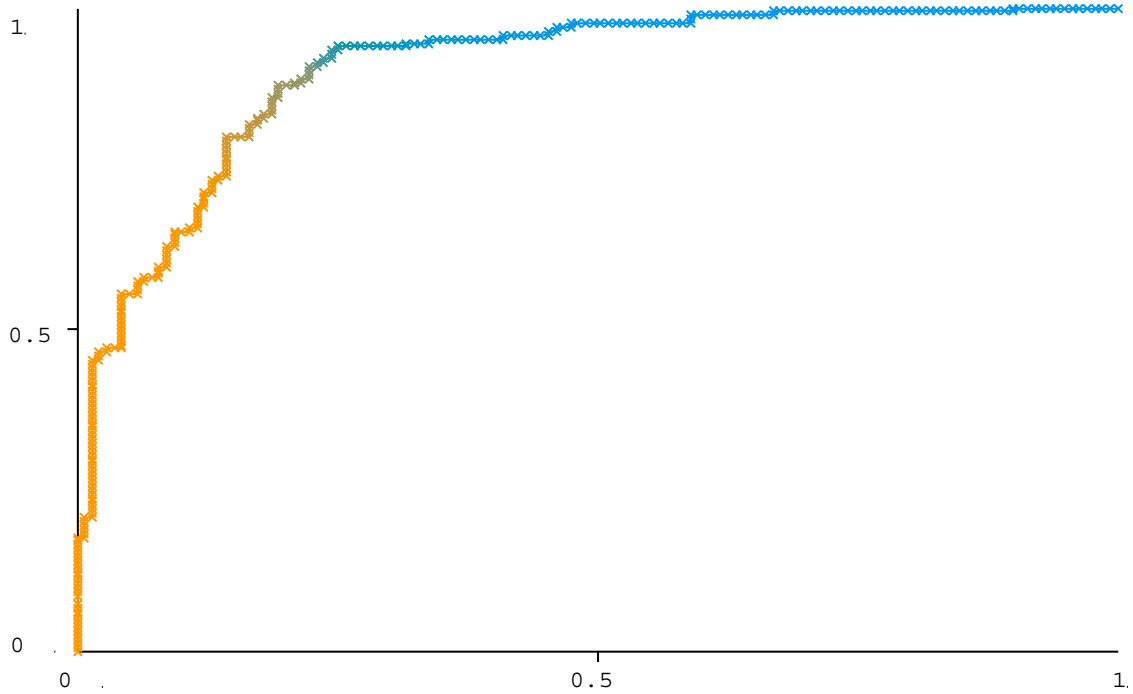


Figure 6. ROC curve of an NB classifier with a correlation-based feature selection set.

As shown in Table 9, the maximum accuracy of 86.468% was attained with the sequential minimal optimization (SMO) algorithm, followed by bagging with an LR classifier, with a maximum accuracy of 85.478% using the optimal attribute set obtained from chi-squared attribute evaluation technique. The SMO algorithm also offered the best MAE of 0.135, sensitivity of 0.865, fallout of 0.142, precision of 0.865, F-measure of 0.864, and specificity of 0.90 relative to other classifiers. Both the naïve Bayes and logistic regression classifiers

achieved a high ROC area value of 0.909. The graphical representation of performance metrics using the chi-squared attribute evaluation method is shown in Figure 7. The ROC curve with an area of 0.909, which was found using the naïve Bayes and logistic regression models using the chi-squared attribute evaluation set, are shown in Figure 8a,b respectively.

Table 9. Performance of the machine learning classifiers based on the optimum attribute set found using the chi-squared attribute evaluation technique.

Classifier	Accuracy	MAE	Sensitivity	Fallout	Precision	F-Measure	ROC Area	Specificity
NB	83.498	0.183	0.835	0.171	0.835	0.835	0.909	0.870
LR	84.488	0.212	0.845	0.159	0.845	0.845	0.909	0.870
SMO	86.468	0.135	0.865	0.142	0.865	0.864	0.861	0.900
IBk/KNN	77.887	0.223	0.779	0.226	0.779	0.779	0.775	0.800
AdaBoostM1 + DS	83.498	0.224	0.835	0.173	0.836	0.834	0.899	0.880
AdaBoostM1 + LR	84.488	0.214	0.845	0.159	0.845	0.845	0.854	0.870
Bagging + REPTree	82.178	0.282	0.822	0.188	0.823	0.821	0.883	0.880
Bagging + LR	85.478	0.217	0.855	0.152	0.855	0.854	0.908	0.890
JRip	76.897	0.319	0.769	0.235	0.769	0.769	0.765	0.790
RF	83.168	0.257	0.815	0.193	0.816	0.814	0.904	0.880

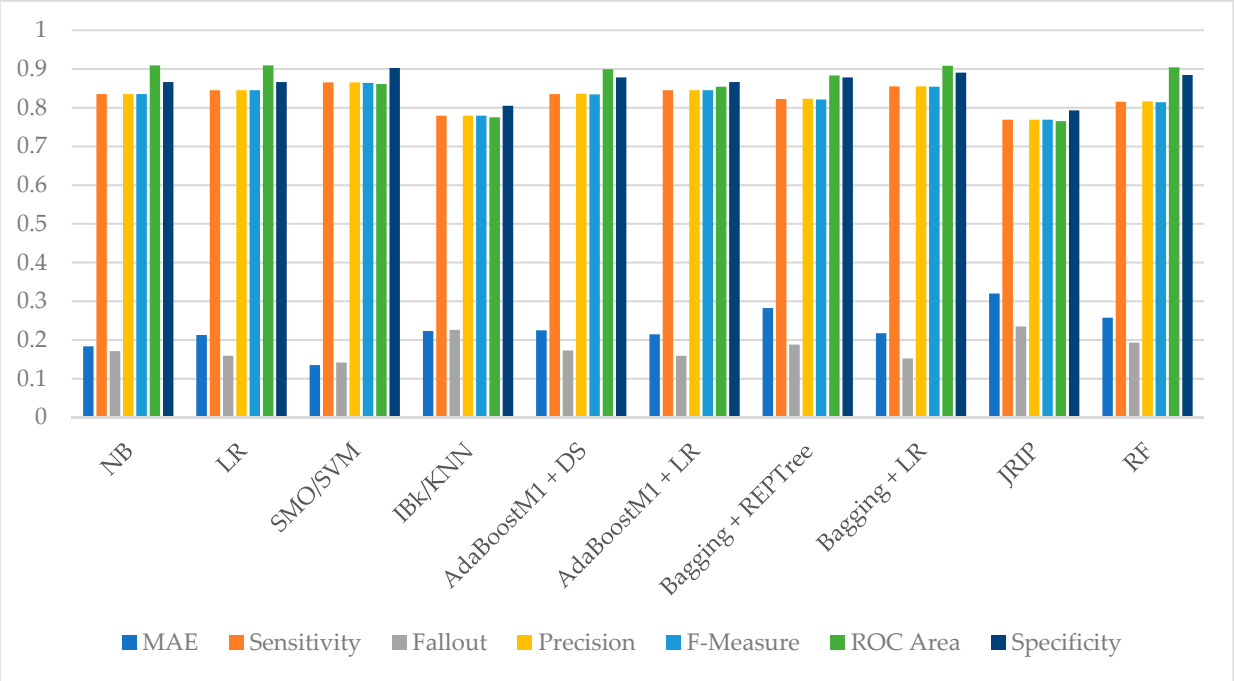
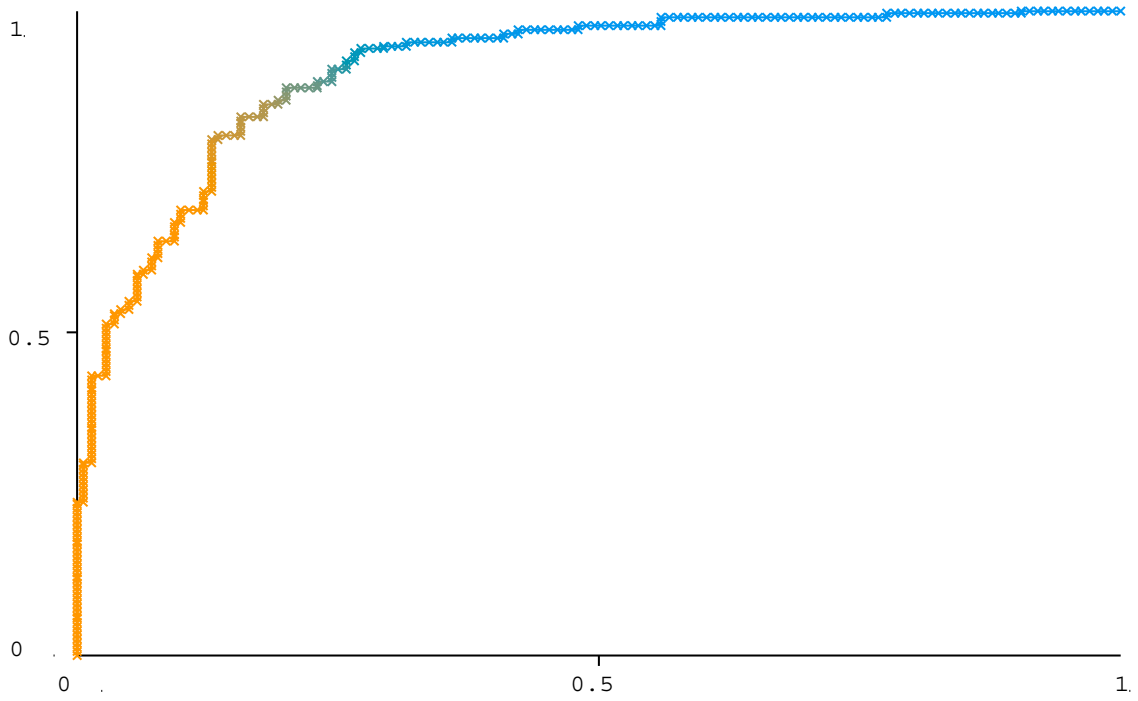
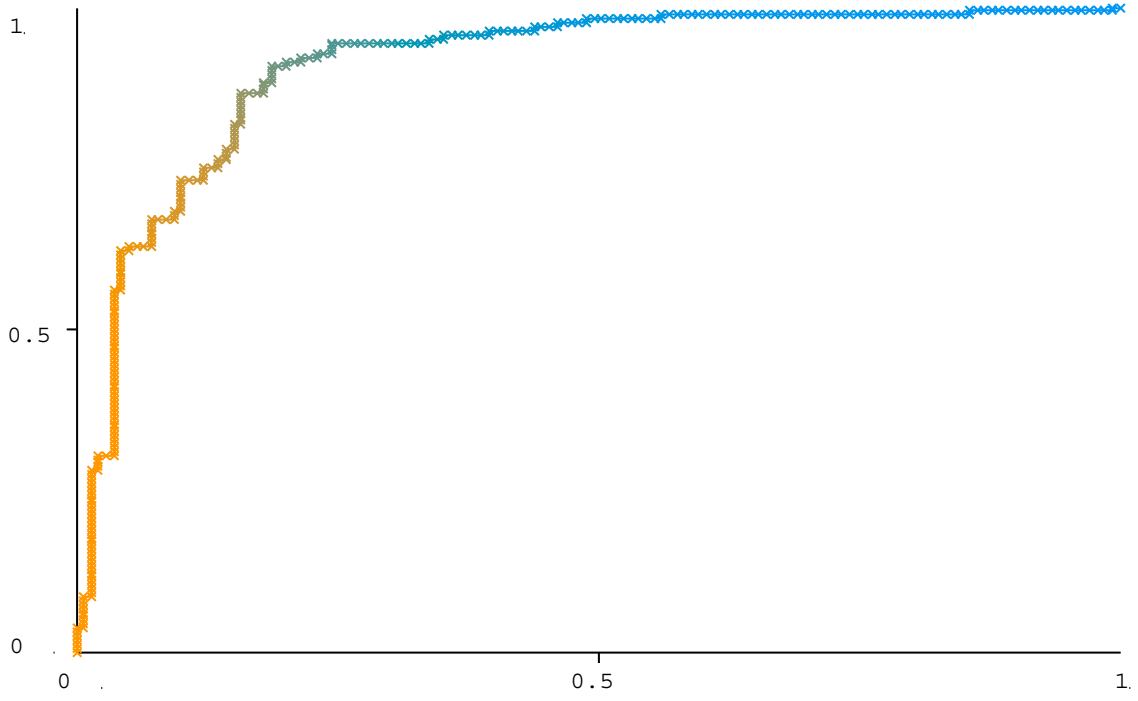


Figure 7. Performance metrics based on the optimal set obtained using the chi-squared attribute evaluation technique.



(**a**)



(**b**)

Figure 8. ROC curves of the optimal set obtained using chi-squared attribute evaluation: (**a**) naïve Bayes classifier and (**b**) logistic regression classifier.

As shown in Table 10, the highest accuracy of 86.138% was achieved with the sequential minimal optimization (SMO) algorithm, followed by the bagging with LR algorithm, with the highest accuracy of 85.148% based on the optimum attribute set, which was obtained using the ReliefF attribute evaluation method. Furthermore, the SMO algorithm produced the best MAE of 0.138, sensitivity of 0.861, fallout of 0.145, precision of 0.862,

F-measure of 0.861, and specificity of 0.90 compared with the other machine learning classifiers. The meta classifier bagging + LR achieved a high ROC area value of 0.91. The visualization of the threshold curve for the target class produced an ROC area of 0.91 using bagging with LR meta classifier, which is shown in Figure 9. Figure 10 shows a graphical representation of performance metrics using the ReliefF attribute evaluation approach.

Table 10. Performance of machine learning classifiers based on the optimal attribute set using the ReliefF attribute evaluation technique.

Classifier	Accuracy	MAE	Sensitivity	Fallout	Precision	F-Measure	ROC Area	Specificity
NB	84.488	0.180	0.845	0.160	0.845	0.845	0.909	0.870
LR	84.488	0.212	0.845	0.159	0.845	0.845	0.909	0.870
SMO	86.138	0.138	0.861	0.145	0.862	0.861	0.858	0.900
IBk/KNN	78.547	0.218	0.785	0.220	0.785	0.785	0.777	0.820
AdaBoostM1 + DS	83.828	0.225	0.838	0.169	0.839	0.838	0.901	0.880
AdaBoostM1 + LR	84.488	0.211	0.845	0.159	0.845	0.845	0.867	0.870
Bagging + REPTree	83.828	0.278	0.838	0.170	0.839	0.838	0.890	0.880
Bagging + LR	85.148	0.216	0.851	0.154	0.852	0.851	0.910	0.880
JRip	74.587	0.333	0.746	0.261	0.745	0.745	0.753	0.790
RF	81.188	0.252	0.812	0.195	0.812	0.811	0.895	0.850

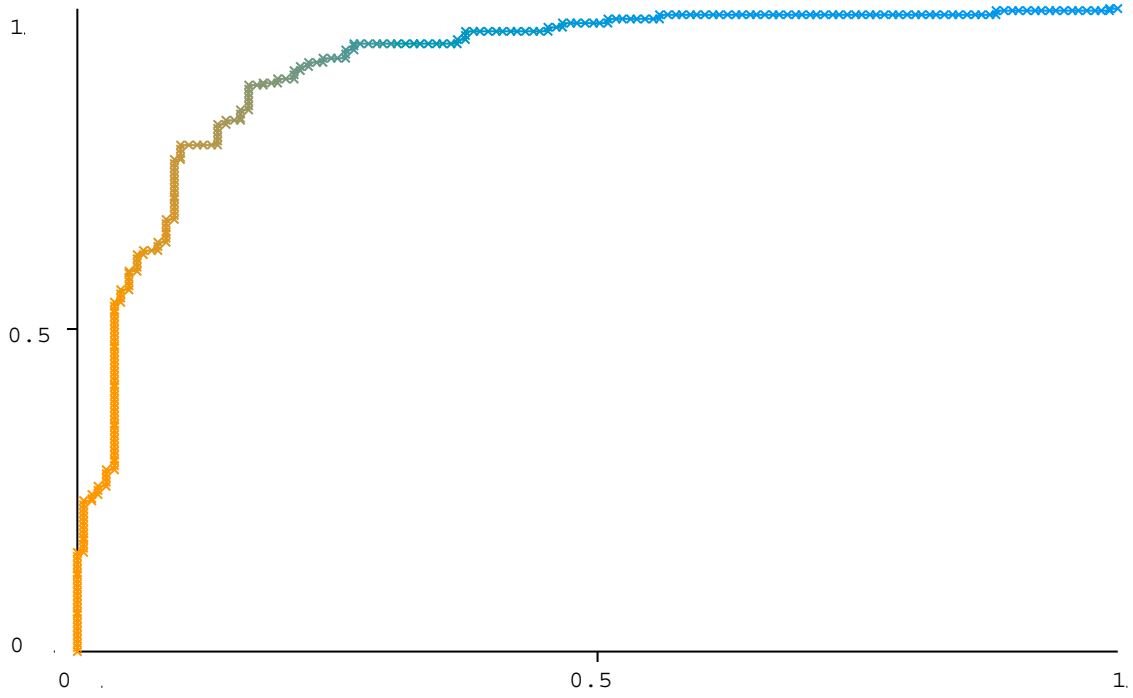


Figure 9. ROC curve of the bagging with LR meta classifier using the ReliefF attribute evaluation set.

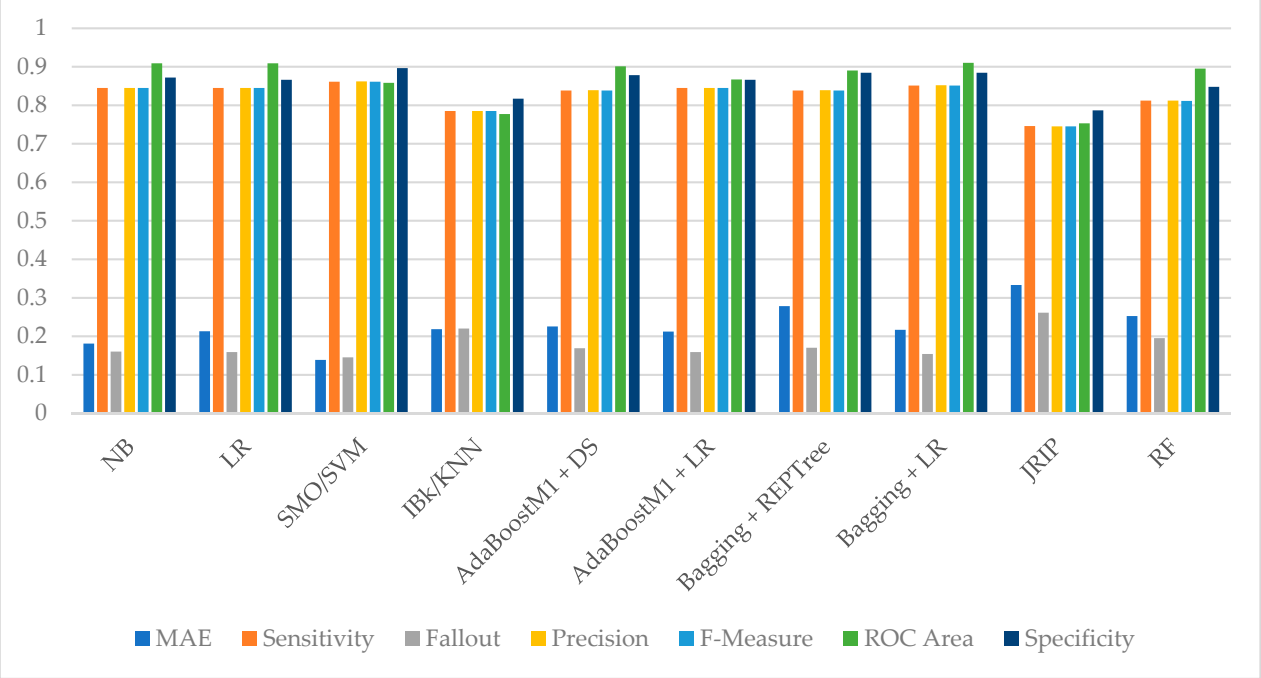


Figure 10. Performance metrics based on the optimal set obtained using the ReliefF attribute evaluation.

A comparison of the accuracy values using the full set of attributes of the Cleveland dataset and optimal attribute sets obtained using various attribute selection techniques performed in this work is shown in Table 11. The correlation-based feature selection method was not able to provide an accuracy value greater than that of the full attribute space. However, there was an improvement in accuracy of about 2% and 1% from the IBk and AdaBoostM1 + DS classifiers, respectively. Besides providing the highest accuracy of 86.468% from SMO, the chi-squared attribute evaluation technique improved most of the classifiers' performances, except for the NB, LR, and AdaBoostM1 + LR classifiers. An increase in the accuracy of about 2% was attained using the JRip algorithm and 1% using the IBk and Bagging + LR algorithms. The ReliefF attribute evaluation method offered the highest improvement in accuracy of about 3% when using the bagging + REPTree classifier, followed by 1.65% from the IBk classifier and about 1% from the SMO and AdaBoostM1 + DS classifiers.

Table 11. Accuracy comparison of the attribute selection techniques.

Classifier	Full Attributes	CfsSubset	Diff.	Chi-Squared	Diff.	ReliefF	Dif.
NB	83.828	84.158	0.330	83.498	−0.330	84.488	0.660
LR	84.818	83.168	−1.650	84.488	−0.330	84.488	−0.330
SMO	85.148	83.828	−1.320	86.468	1.320	86.138	0.990
IBk/KNN	76.897	78.877	1.980	77.887	0.990	78.547	1.650
AdaBoostM1 + DS	82.838	83.828	0.990	83.498	0.660	83.828	0.990
AdaBoostM1 + LR	84.818	83.168	−1.650	84.488	−0.330	84.488	−0.330
Bagging + REPTree	80.858	81.188	0.330	82.178	1.320	83.828	2.970
Bagging + LR	84.488	83.828	−0.660	85.478	0.990	85.148	0.660
JRip	74.917	74.257	−0.660	76.897	1.980	74.587	−0.330
RF	81.848	79.538	−2.310	83.168	1.320	81.188	−0.660

Diff.—difference.

Besides the training of the ML classifier on the full and optimal attribute sets obtained from the attribute evaluators, the hyperparameter 'number of nearest neighbors k' tuning was performed for various values of k = 3, 5, 7, 9, 11, 13, 15, 17, 19, and 21 in the IBk classifier. The best accuracy, accuracy improvement, and other performance metrics for specific 'k' values that were attained from the parameter tuning are presented in Table 12. Though the accuracy delivered by the IBk classifier was slightly less than that of the SMO classifier that was obtained from the chi-squared attribute set, i.e., 86.468%, there was a significant improvement in accuracy by tuning the hyperparameter 'k' value in all the cases. We observed that the greatest accuracy improvement of about 8.25% came from the chi-squared attribute evaluation with k = 9 compared to that of default parameter k = 1. The performance comparison of this research work with the related works is presented in Table 13.

Table 12. Performance comparison of the KNN algorithm by tuning the parameter 'k.'

Attribute set	Acc. (k = 1)	Acc ('k')	Acc. Impr. (%)	MAE	Sen.	Fallout	Pre.	F-Mea.	ROC Area	Spe.
Full attributes	76.897	-	-	0.184	0.769	0.235	0.769	0.769	0.764	0.790
Full attributes	76.897	84.158 (k = 5)	7.260	0.228	0.842	0.166	0.842	0.841	0.893	0.880
CfsSubset	78.877	83.498 (k = 11)	4.620	0.237	0.835	0.178	0.839	0.833	0.889	0.910
Chi-Squared	77.887	86.138 (k = 9)	8.250	0.224	0.861	0.146	0.862	0.861	0.905	0.900
ReliefF	76.897	84.488 (k = 9)	7.590	0.224	0.845	0.165	0.847	0.844	0.904	0.900

Acc.—accuracy, Impr.—improvement, Sen.—sensitivity, Pre.—precision, F-Mea.—F-measure, Spe.—specificity.

Table 13. Performance comparison of related works.

Research Author	Method	# Attr.	Acc. (%)	Pre.	Sen.	AUC
R. Perumal et al. [18]	LR with PCA	7	87.0	-	0.85	-
C.B.C Latha et al. [19]	Majority vote with NB, BN, RF, and MP	9	85.48	-	-	-
D. Ananey-Obiri et al. [20]	LR and GNB with Single value decomposition	4	82.75	-	-	0.87
N. K. Kumar et al. [21]	Random Forest	10	85.71	-	-	0.8675
A. Gupta et al. [22]	FAMD + RF	28	93.44	-	0.8928	-
M. Sultana et al. [23]	SMO	14	84.0741	-	-	0.8392
S. Kodati et al. [25]	SMO	14	-	0.84	0.8365	-
I. Tougui et al. [27]	ANN	14	85.86	-	0.8394	-
V. Pavithra et al. [28]	HRFLC (RF + AdaBoost + Pearson Coefficient)	11	79.0	0.78	0.79	-
C. Gazeloglu et al. [29]	Correlation-based feature selection with NB	6	84.818	-	-	0.905
C. Gazeloglu et al. [29]	Fuzzy Rough Set and Chi-square FS with Radial bias function (RBF) Network	7	81.188	-	-	0.261
B. A. Tama et al. [32]	Two-tier ensemble PSO	7	85.6	-	-	0.8586
S. M. Saqlain et al. [43]	Forward feature selection with Radial Basis Function SVM	7	81.19	-	72.92	-
Proposed method	**Chi-Squared + SMO**	**11**	**86.468**	**0.865**	**0.865**	**0.861**

Attr.—attributes, Acc.—accuracy, Pre.—precision, Sen.—sensitivity, AUC—area under the ROC curve.

This research work utilized the Cleveland heart dataset to achieve the highest accuracy of 85.148% with the SMO model based on the full set of attributes and an accuracy of 84.158% with the NB model based on an optimal set of seven attributes obtained from correlation-based feature selection. The SMO classifier further achieved the best prediction accuracies of 86.468% and 86.138% from the optimal sets obtained from chi-squared (11 attributes) and ReliefF (10 attributes) techniques, respectively. The best values of other performance metrics, namely, MAE (0.135), sensitivity (0.865), specificity (0.90), fallout (0.142), precision (0.865), and F-measure (0.864), was obtained from SMO with the chi-squared method. The bagging + LR classifier provided an ROC area of 0.91 on both the full attributes and optimal sets obtained from the ReliefF method. Nevertheless, the ensemble classifiers AdaBoost and bagging fell short in their predictions compared to the SMO, while the bagging + REPTree classifier achieved the highest improvement in accuracy of about 3% with the ReliefF method. Tuning of the hyperparameter 'k' in IBk reached an improvement in accuracy of 8.25% with the chi-squared evaluator for k = 9. Overall, the SMO classifier showed better performance on the full attributes and optimal sets obtained from the chi-squared and ReliefF attribute evaluators, whereas the NB classifier showed a better performance with the correlation-based feature selection technique.

5. Conclusions

In this study, three attribute evaluator techniques were utilized to select significant attributes from the Cleveland heart dataset to improve the performance of machine learning classifiers when predicting heart disease risk. A remarkable performance was achieved by the SMO classifier using the chi-squared attribute evaluation method. Eventually, we noticed that there was a significant improvement in the prediction performance with appropriate attribute selection and tuning the hyperparameters of the classifiers. Although the performance of the classifiers looks satisfactory, a smaller dataset of 303 instances, 10 machine learning classifiers, and 3 feature selection methods were used in this research. There is a huge scope to explore various machine learning algorithms and feature selection techniques. In the future, we intend to combine multiple datasets to obtain a higher number of observations and conduct more experiments by selecting appropriate attributes to improve the classifier's predictive performance.

Author Contributions: Conceptualization, K.V.V.R. and I.E.; methodology, K.V.V.R., I.E., A.A.A. and H.N.C.; software, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); validation, K.V.V.R., I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); formal analysis, K.V.V.R. and I.E.; investigation, K.V.V.R.; resources, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); data curation, K.V.V.R.; writing–original draft preparation, K.V.V.R.; writing–review and editing, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); visualization, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); supervision, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); project administration, I.E., A.A.A. and H.N.C.; funding acquisition, I.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Teknologi PETRONAS, grant number 0153AB-M66 and the APC was funded by 0153AB-M66.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [https://archive.ics.uci.edu/ml/datasets/heart+disease] accessed on 15 August 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO. Available online: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 (accessed on 9 February 2021).
2. Healthline. Available online: https://www.healthline.com/health/stroke-vs-heart-attack#treatment (accessed on 20 February 2021).
3. Chicco, D.; Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 1–16. [CrossRef] [PubMed]
4. Karthick, D.; Priyadharshini, B. Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using classification techniques within fifty years of age. In Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018, Coimbatore, India, 19–20 January 2018; pp. 1182–1186. [CrossRef]
5. Obasi, T.; Shafiq, M.O. Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases. In Proceedings of the 2019 IEEE International Conference on Big Data, Big Data 2019, Los Angeles, CA, USA, 9–12 December 2019; pp. 2393–2402. [CrossRef]
6. Sharma, H.; Rizvi, M.A. Prediction of Heart Disease using Machine Learning Algorithms: A Survey. *Int. J. Recent Innov. Trends Comput. Commun.* **2017**, *5*, 99–104.
7. Ramalingam, V.V.; Dandapath, A.; Raja, M.K. Heart disease prediction using machine learning techniques: A survey. *Int. J. Eng. Technol.* **2018**, *7*, 684–687. [CrossRef]
8. Alaa, A.M.; Bolton, T.; Di Angelantonio, E.; Rudd, J.H.F.; Van Der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS ONE* **2019**, *14*, e0213653. [CrossRef]
9. Uddin, S.; Khan, A.; Hossain, E.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1–16. [CrossRef]
10. Song, Q.; Zheng, Y.-J.; Yang, J. Effects of Food Contamination on Gastrointestinal Morbidity: Comparison of Different Machine-Learning Methods. *Int. J. Environ. Res. Public Heal.* **2019**, *16*, 838. [CrossRef]
11. Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease Prediction by Machine Learning Over Big Data from Healthcare Communities. *IEEE Access* **2017**, *5*, 8869–8879. [CrossRef]
12. Aljanabi, M.; Qutqut, M.; Hijjawi, M. Machine Learning Classification Techniques for Heart Disease Prediction: A Review. *Int. J. Eng. Technol.* **2018**, *7*, 5373–5379. [CrossRef]
13. Pasha, S.J.; Mohamed, E.S. Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction. *IEEE Access* **2020**, *8*, 184087–184108. [CrossRef]
14. Swain, D.; Pani, S.K.; Swain, D. A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning. In Proceedings of the 2018 International Conference on Advanced Computation and Telecommunication, ICACAT, Bhopal, India, 28–29 December 2018; pp. 1–6. [CrossRef]
15. Weng, S.F.; Reps, J.M.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* **2017**, *12*, e0174944. [CrossRef]
16. Khan, Y.; Qamar, U.; Yousaf, N.; Khan, A. Machine Learning Techniques for Heart Disease Datasets: A Survey. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing, Zhuhai, China, 22–24 February 2019; pp. 27–35. [CrossRef]
17. Goel, S.; Deep, A.; Srivastava, S.; Tripathi, A. Comparative Analysis of various Techniques for Heart Disease Prediction. In Proceedings of the 2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019, Mathura, India, 21–22 November 2019; pp. 88–94. [CrossRef]
18. Perumal, R. Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 4225–4234.
19. Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inform. Med. Unlocked* **2019**, *16*, 100203. [CrossRef]
20. Ananey-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms. *Int. J. Comput. Appl.* **2020**, *176*, 17–21. [CrossRef]
21. Kumar, N.K.; Sindhu, G.; Prashanthi, D.; Sulthana, A. Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 15–21. [CrossRef]
22. Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. *IEEE Access* **2019**, *8*, 14659–14674. [CrossRef]
23. Sultana, M.; Haider, A.; Uddin, M.S. Analysis of data mining techniques for heart disease prediction. In Proceedings of the 2016 3rd International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2016, Dhaka, Bangladesh, 22–24 September 2016; pp. 1–5. [CrossRef]
24. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **2019**, *7*, 81542–81554. [CrossRef]
25. Kodati, S.; Vivekanandam, R. Analysis of Heart Disease using in Data Mining Tools Orange and Weka Sri Satya Sai University Analysis of Heart Disease using in Data Mining Tools Orange and Weka. *Glob. J. Comput. Sci. Technol.* **2018**, *18*.

26. Ed-Daoudy, A.; Maalmi, K. Performance evaluation of machine learning based big data processing framework for prediction of heart disease. In Proceedings of the International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco, 26–27 December 2019; pp. 1–5. [CrossRef]
27. Tougui, I.; Jilbab, A.; El Mhamdi, J. Heart disease classification using data mining tools and machine learning techniques. *Health Technol.* **2020**, *10*, 1137–1144. [CrossRef]
28. Pavithra, V.; Jayalakshmi, V. Hybrid feature selection technique for prediction of cardiovascular diseases. *Mater. Today Proc.* **2021**, *22*, 660–670. [CrossRef]
29. Gazeloğlu, C. Prediction of heart disease by classifying with feature selection and machine learning methods. *Prog. Nutr.* **2020**, *22*, 660–670. [CrossRef]
30. Louridi, N.; Amar, M.; El Ouahidi, B. Identification of Cardiovascular Diseases Using Machine Learning. In Proceedings of the 7th Mediterranean Congress of Telecommunications 2019, CMT 2019, Fez, Morocco, 24–25 October 2019; pp. 1–6. [CrossRef]
31. Kavitha, M.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S. Heart Disease Prediction using Hybrid machine Learning Model. In Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, Coimbatore, India, 20–22 January 2021; pp. 1329–1333. [CrossRef]
32. Tama, B.A.; Im, S.; Lee, S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. *BioMed Res. Int.* **2020**, *2020*. [CrossRef]
33. Heart Disease Dataset. Available online: https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed on 24 May 2021).
34. Haq, A.U.; Li, J.P.; Memon, M.H.; Nazir, S.; Sun, R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mob. Inf. Syst.* **2018**, *2018*. [CrossRef]
35. Maini, E.; Venkateswarlu, B.; Maini, B.; Marwaha, D. Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India. *Med. J. Armed Forces India* **2021**, *77*, 302–311. [CrossRef]
36. Zeng, Z.-Q.; Yu, H.-B.; Xu, H.-R.; Xie, Y.-Q.; Gao, J. Fast training Support Vector Machines using parallel sequential minimal optimization. In Proceedings of the 2008 3rd International Conference on Intelligent System and Knowledge Engineering, ISKE 2008, Xiamen, China, 17–19 November 2008; Volume 1, pp. 997–1001. [CrossRef]
37. Ghosh, P.; Azam, S.; Jonkman, M.; Karim, A.; Shamrat, F.M.J.M.; Ignatious, E.; Shultana, S.; Beeravolu, A.R.; De Boer, F. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. *IEEE Access* **2021**, *9*, 19304–19326. [CrossRef]
38. Kang, K.; Michalak, J. Enhanced Version of AdaBoostM1 with J48 Tree Learning Method. [1802.03522] Enhanced Version of AdaBoostM1 with J48 Tree Learning Method. Available online: arxiv.org (accessed on 27 June 2021).
39. Almustafa, K.M. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinform.* **2020**, *21*, 1–18. [CrossRef]
40. Muhammad, Y.; Tahir, M.; Hayat, M.; Chong, K.T. Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Sci. Rep.* **2020**, *10*, 1–17. [CrossRef]
41. Al Janabi, K.B.; Kadhim, R. ('Weka' Feature Selection-bad results) Data Reduction Techniques: A Comparative Study for Attribute Selection Methods. *Int. J. Adv. Comput. Sci. Technol.* **2018**, *8*, 1–13.
42. Spencer, R.; Thabtah, F.; Abdelhamid, N.; Thompson, M. Exploring feature selection and classification methods for predicting heart disease. *Digit. Health* **2020**, *6*, 1–10. [CrossRef] [PubMed]
43. Saqlain, S.M.; Sher, M.; Shah, F.A.; Khan, I.; Ashraf, M.U.; Awais, M.; Ghani, A. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowl. Inf. Syst.* **2018**, *58*, 139–167. [CrossRef]

Healthcare Analytics 2 (2022) 100060

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

Analyzing the impact of feature selection on the accuracy of heart disease prediction

Muhammad Salman Pathan [a,b], Avishek Nag [c], Muhammad Mohisn Pathan [d], Soumyabrata Dev [a,b,*]

[a] ADAPT SFI Research Centre, Dublin, Ireland
[b] School of Computer Science, University College Dublin, Ireland
[c] School of Electrical and Electronic Engineering, University College Dublin, Ireland
[d] Institute of Biomedical Engineering and Technology, Liaquat University of Medical and Health Sciences, Pakistan

ARTICLE INFO

Keywords:
Heart disease
Machine learning
Dimensionality reduction
Feature correlation
Feature selection

ABSTRACT

Heart Disease has become one of the most serious diseases that has a significant impact on human life. It has emerged as one of the leading causes of mortality among the people across the globe during the last decade. In order to prevent patients from further damage, an accurate diagnosis of heart disease on time is an essential factor. Recently we have seen the usage of non-invasive medical procedures, such as artificial intelligence-based techniques in the field of medical. Specially machine learning employs several algorithms and techniques that are widely used and are highly useful in accurately diagnosing the heart disease with less amount of time. However, the prediction of heart disease is not an easy task. The increasing size of medical datasets has made it a complicated task for practitioners to understand the complex feature relations and make disease predictions. Accordingly, the aim of this research is to identify the most important risk-factors from a highly dimensional dataset which helps in the accurate classification of heart disease with less complications. For a broader analysis, we have used two heart disease datasets with various medical features. Firstly, we performed the correlation and inter-dependence of different medical features in the context of heart disease. Secondly, we applied a filter-based feature selection technique on both datasets to select most relevant features (an optimal reduced feature subset) for detecting the heart disease. Finally, various machine learning classification models were investigated using complete and reduced features subset as inputs for experimentation analysis. The trained classifiers were evaluated based on Accuracy, Receiver Operating Characteristics (ROC) curve, and F1-Score. The classification results of the models proved that there is a high impact of relevant features on the classification accuracy. Even with a reduced number of features, the performance of the classification models improved significantly with a reduced training time as compared with models trained on full feature set.

1. Introduction

Heart disease is rapidly increasing across the globe. As per a research report published by the World Health Organization (WHO), in 2016 approximately 17.90 million people died from heart disease [1]. This much number accounts for approximately 30% of all deaths worldwide. Nearly 55% of the heart patient die during the first 3 years, and the treatment costs for heart disease are around 4% of the annual healthcare expenditure. [2]. Observing the increasing stats, accurate and timely detection and treatment of this serious illness is very essential for disease prevention and effective utilization of medical resources.

Due to the recent technological advancements, the field of medical sciences has seen a remarkable improvement over time [3,4]. Specially, machine learning (ML) has been widely used in the field of cardiovascular medicine and has established a potential space [5]. The basic framework of ML is built on models that take input data (such as text or images) and through the usage of some statistical analysis and mathematical optimizations provides the desired prediction results (*e.g.,* disease, no disease, neutral) [6]. ML models can be trained on tons of raw electronic medical data gathered from low-cost wearable devices to allow efficient heart disease diagnosis with less resources and improved accuracy [7].

During the training process, ML models require a large number of data samples to avoid overfitting [8]. However, the inclusion of the large number of data features is not required for reasons related to the curse of dimensionality [9,10]. Mostly, medical datasets cover related as well as redundant features. Unnecessary features do not contribute any meaningful information to the prediction task, and also creates

* Corresponding author at: School of Computer Science, University College Dublin, Ireland.
E-mail address: soumyabrata.dev@ucd.ie (S. Dev).

https://doi.org/10.1016/j.health.2022.100060
Received 13 February 2022; Received in revised form 14 April 2022; Accepted 28 April 2022
Available online xxxx

noise in the description of target (output class) which leads to prediction errors [11]. Furthermore, such features increase the complexity of ML models and make the system runs slowly due to increased training time. To overcome the curse of dimensionality only those features which are closely related with the target should be selected/identified from datasets and provided as inputs to ML models [12]. Relevant feature selection can aid in performance improvement by decreasing the model complexity and increasing prediction accuracy which is very important in medical diagnosis [13]

Because of the benefits outlined previously, feature selection techniques are being actively used in the area of heart diseases and strokes [14–16].

The contributions of this research are listed as follows:

- The study uses two datasets of heart disease patients from different sources to cover a broader study of medical features.
- To perform the correlation and interdependence study between different features in datasets with respect to heart disease.
- The identification of the most relevant medical features which aids in the prediction of heart disease using a filter-based feature selection technique.
- Different ML classification models such as Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), Multi Layer Perceptron (MLP) *etc.*, are used on the datasets to identify the suitable models for the problem.
- The classification models were tested on full as well as the reduced feature subset to observe the impact of feature selection on the performance of models.
- With the spirit of reproducible research, the code of this article is shared in GitHub.[1]

2. Related work

ML has appeared to be an effective technique for assisting in the heart disease diagnosis, however the high dimensionality of datasets is a fundamental issue for ML prediction models. Feature selection is one of the techniques which is used to select only the most relevant features from datasets features that influence the disease outcome most. The identification of the most important features from the high dimensional datasets is an important aspect that can improve the accuracy of prediction models hence reduce the number of medical injuries.

In [17], Zhang et al. developed an efficient feature selection technique called weighting-and ranking-based hybrid feature selection (WRHFS) to determine the risk of heart stroke. For the weighing and ranking of features, WHRFS used a variety of filter-based feature selection techniques such as fisher score, information gain and standard deviation. The proposed technique selected 9 important input features out of 28 based on the knowledge provided for heart stroke prediction. In another research [18], the authors worked on the extraction of relevant risk factors form a large feature space for an efficient heart disease prediction. The features were selected based on their individual ranks. The authors used Latent Feature Selection (ILFS) method to rank the features which is a probabilistic latent graph-based feature selection technique. The results of the model were competitive using only half of the features from the set of 50. In [19], a feature selection model for detecting the risk of heart disease is proposed. The proposed model combined the glow-worm swarm optimization algorithm based on the standard deviation of the features to extract the quality features from a electronic healthcare record (EHR) of a community hospital in Beijing. 6 features including high blood pressure, Alkaline Phosphatase (ALP), age and Lactate Dehydrogenase (LDH) were indicated as important features to detect stroke excluding the family hereditary factors. The authors of [20] focused on finding the most relevant features from EHR to predict the early-stage risk of death from heart disease. The authors used minimum redundancy maximum (mRmR) relevance and recursive feature elimination (RFE) feature selection approaches based on NB for the selection of features. Two medical features *i.e.,* Serum Creatinine and Ejection Fraction were ranked higher by both feature selection technique as compared to other. When provided to a prediction model as input, the selected features proved out to be most important as an overall accuracy of 80% was achieved. Singh et al. [21], proposed an efficient approach for stroke prediction using the Cardiovascular Health Study (CHS) dataset. They used DT algorithm for feature selection and then principal component analysis (PCA) technique for reducing the dimensionality of feature space. Finally, the MLP network was used to construct the classification model. The model trained on the optimal feature set achieved 97.7% accuracy in detecting the stroke and outperformed other techniques in comparison. A wrapper based Genetic Algorithm (GA) is used in [22] to select the most significant features to detect heart disease. The proposed feature selection algorithm identifies 7 features out of 16 to detect heart disease from Cleveland heart disease dataset. The resultant features were supplied to support vector machine (SVM) for the accuracy evaluation. The classifier acquired 88.34% using the reduced feature set whereas only 83.34% was achieved when using whole dataset features. In terms of ROC curve, the GA-SVM performed well also when compared with the various existing feature selection algorithms also. This study [23] proposes a new heart disease prediction model by combining ML with deep learning techniques. The least absolute shrinkage and selection operator (LASSO) penalty method based on LinearSVC was applied as the feature selection module to generate a feature subset closely related to target. 12 most relevant features were chosen from dataset obtained from Kaggle and inputted to the MLP network. As per the experimental results, the proposed model obtained an accuracy of 98.56% with 99.35% recall and 97.84% precision. In [5], a ML based heart disease diagnosis system is proposed. Seven popular classifiers LR, k-Nearest Neighbor (K-NN), MLP, SVM, NB, DT, and RF were used for the classification of heart disease patients. Three feature selection algorithms RelieF, mRMR, and LASSO were used to select highly correlated features with target class. It was observed that the classification performance of models increased in terms of accuracy and computation time using the feature selection techniques. The LR model showed best accuracy of 89% when used with RelieF. The main objective of this research [24] was to predict the heart disease using minimal subset of features and adequate accuracy. To achieve this objective, the authors employed a two-stage feature subset retrieving technique. Three popular feature selection techniques *i.e.,* (embedded, filter, wrapper) were used to extract a feature subset based on a boolean process-based common "True" condition. To select the suitable prediction model, RF, SVM, K-NN, NB, XGBoost and MLP models were trained on the data. The experimental results showed that XGBoost classifier integrated with wrapper technique provided the best prediction results for heart disease. A comparative analysis of different classifiers was performed in [25] for the classification of the heart disease with minimal attributes. ML classifiers such as NB, LR, sequential minimal optimization (SMO), RF *etc.,* were trained for the accurate detection of heart disease. To obtain the optimal feature subset, RelieF, chi-squared and correlation-based feature subset evaluator were utilized. 10 features were selected from the set of 13 to train the classifiers. The SMO classifier achieved the highest accuracy of 86.468% when inputted with the optimal feature set obtained by chi-squared feature selection technique.

Despite their relevance, one major drawback of existing works on heart disease prediction is the lack of systematic guidance when selecting the input features for the development of prediction models which is an important aspect in terms of predictive performance. Previous research proposals chose features mostly in an impromptu manner without incorporating latest medical research findings. Mostly the focus is on the prediction models and their final prediction performance. However, a very less attention is paid on the correlation

[1] https://github.com/Sammyy092/analyzing-the-impact-of-feature-selection-on-heart-disease-prediction.

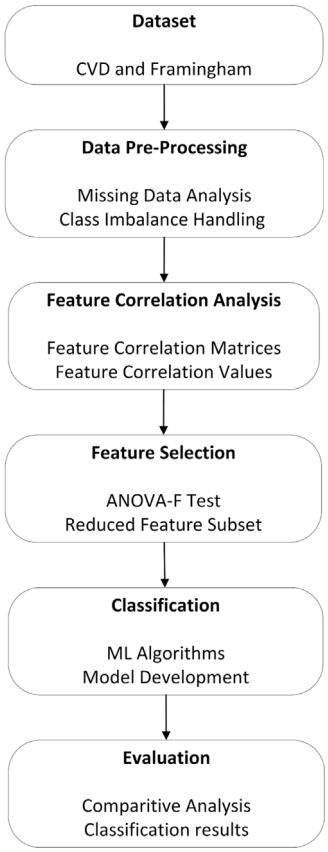


Fig. 1. Flowchart of the proposed methodology describing each step for heart disease prediction.

between different medical features and their individual importance in the prediction of heart disease. A few works present analysis of medical features but for the purpose of heart disease detection only. This research aims at addressing the ineffective feature selection in previous studies on heart disease prediction. Two heart disease patient datasets collected from different sources were utilized in this research to cover a broader study of features related to heart disease and to identify various medical procedures. To further analyze the role of each parameter in the prediction task, we obtain the interdependence and importance of the collected set of medical features. A detailed analysis of ML models trained on both full and selected feature set is provided to analyze the impact of feature selection techniques on the prediction performance as well as the identification of suitable classifiers for the specified problem.

3. Proposed methodology

This research paper highlights the importance feature selection in the accurate classification of heart disease. Fig. 1 demonstrates the workflow of the proposed methodology for heart disease prediction.

3.1. Datasets

In this research, two datasets named as cardiovascular disease (CVD) and Framingham were utilized to study the impact of different features on the occurrence of heart disease and to develop ML-based system for heart disease detection. The study uses two datasets to cover a broader study of medical features and various clinical pathways used for the detection of heart stroke. The datasets were collected from different sources. The datasets contained some main medical features like 'age', 'hypertension', 'glucose levels', 'blood pressure', 'cholesterol' *etc.* which are closely related to the occurrence of disease and provides a great flexibility for heart disease analysis. The datasets were chosen based on two criteria. The first criterion was the variance in the medical procedures, so to study the different medical procedures and the role of each feature in the context of heart disease. Secondly, the datasets were chosen based on the data availability. Datasets from different sources possess different amount of data and collection of features. So, we have chosen datasets which were offering a good volume of data and having a level of similarity in terms of features.

Table 1
Description of features CVD dataset.

Attribute	Description
i.d	patient's i.d
gender	includes ("male": 0, "female": 1, "other": 2)
age	patient's age (continuous)
hypertension	suffering from hypertension ("yes":1, "no":0)
heart _disease	suffering heart disease ("yes":1, "no":0)
ever_married	marital status of patient ("yes":1, "no":0)
work_type	job status ("children":0, "govt_job":1, "never_worked":2, "private":3, "self_employed":4)
residence_type	("rural:0, "urban":1)
avg_glucose_level	average glucose level of blood (continuous)
bmi	body mass index (decimal value)
smoking_status	("never smoked":0, "formerly smoked":1, "smokes":2)
stroke	("yes":1, "no":0)

3.1.1. CVD

The CVD dataset is controlled by McKinsey & Company which was a part of their healthcare hackathon.[2] The dataset can be accessible from a free dataset repository.[3] The collected dataset included 29072 patient observation with 12 data features. 11 of them are the common clinical symptoms and are considered as input features whereas the 12*th* feature 'stroke' is the target feature indicating whether a patient has had stroke or not. The complete description of data features for CVD dataset is given in Table 1.

3.1.2. Framingham

The Framingham dataset was created during an ongoing cardiovascular study involving the residents of Framingham, Massachusetts, and is available at the Kaggle website.[4] The dataset is mostly used in classification tasks to identify whether a patient has a chance to develop coronary heart disease (CHD) in 10 years. The dataset contains 4,240 patient records and 15 features, where each feature indicates a risk factor. 14 input features were used to detect the decisional feature *i.e.,* 10-year risk of CHD. Table 2 shows the description about the data features in Framingham dataset.

3.2. Pre-processing

Data pre-processing is one of the important part of ML life cycle as it makes data analysis easy and increases the accuracy and speed of the ML algorithms [26] . We applied some pre-processing steps as the collected dataset were having missing values and class imbalance problems. Referring the CVD dataset, the dataset contained a total of 43400 patient records out of which 14754 values were missing or null. Whereas, 4240 patient records were available for framingham dataset of which 645 values were null. A null value does not necessarily mean that the value does not exist, but it is unknown. In medical datasets, mostly the null or missing value is usually due to a lack of collection or the practitioner may not consider the observation

[2] https://datahack.analyticsvidhya.com/contest/mckinsey-analyticsonline-hackathon/.
[3] https://inclass.kaggle.com/asaumya/healthcare-dataset-stroke-data.
[4] https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset.

Table 2
Description of features Framingham dataset.

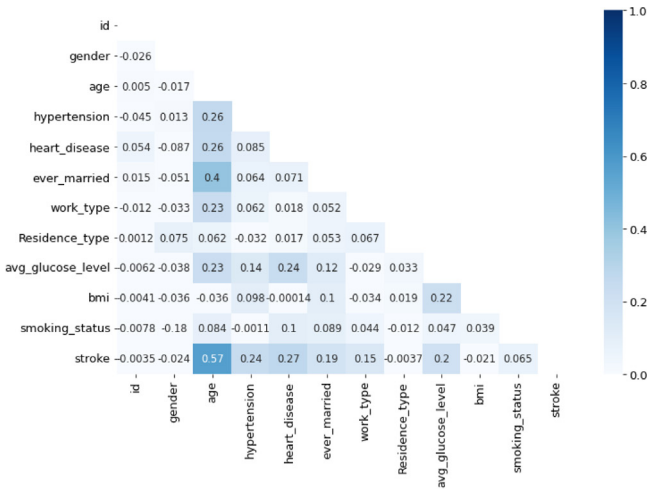
Attribute	Description
age	patient's age (continuous)
male	("male":0, "female":1)
education	level of education (1 to 4)
currentSmoker	("smoker":1, "non smoke":0)
CigsPerDay	average number of cigarettes consumed per day (continuous)
BPMeds	on blood pressure medication ("yes":1, "no":0)
prevalentStroke	previous stroke history ("yes": 1, "no":0)
prevalenHyp	hypertensive ("yes":1, "no":0)
diabetes	previous diabetes history("yes":1, "no":0)
totCHol	cholesterol level (continuous)
sysBP	systolic blood pressure (decimal)
diaBP	diastolic blood pressure (decimal)
BMI	body mass index (decimal)
HeartRate	heart rate measure (continuous)
glucose	glucose level (continuous)
TenYearCHD	target ("yes": 1, "no": 0)

since the medical test is considered to be low yield for the patient. Data imputation methods are useful in handling the missing data, however their usage in medical field is limited and specific efficacy for disease detection is not clear [27]. Most of the times, researchers do not consider the observations with missing values and drop the incomplete cases intentionally, since the traditional data imputation methods are not sufficient to capture the missing data complexities in health care applications [28,29]. However, only a deep knowledge of specific disease will likely aid in the selection of the suitable data imputation methods. As per the mentioned analysis, we dropped all the observations with null value from both the datasets to avoid any accuracy biases.

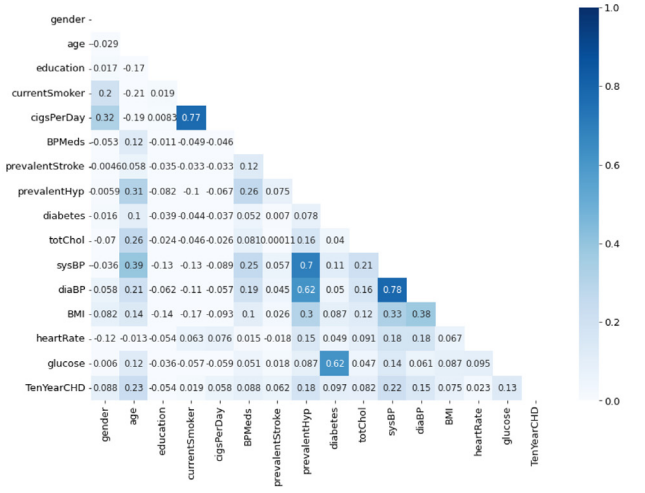
Furthermore, looking at the class distribution, both datasets were highly unbalanced in nature. Only 548 patients out of 29,072 in CVD dataset had stroke conditions, whereas 28,524 patients had no occurrence of stroke. In framingham dataset, only 557 patient records showed the risk of CHD out of 3101. The unbalanced nature of the datasets leads to classification errors during the training of ML models [30]. As a result, we adopted a 'Random Down-Sampling' technique to mitigate the adverse effects caused by unbalanced data. We made two classes referred as 'minority' and 'majority' classes. The patients with heart disease were included in minority class, whereas the patients having no symptoms were included in majority class. In the case of CVD dataset, 548 observations were included into the minority class and the remaining 28,524 were considered as majority class. We created a balanced dataset of 1096 observations by selecting all 548 observations from minority class and 548 random observations from a total of 28,524 majority cases. Same process was performed for framingham dataset where 557 random observations from 3101 majority cases were derived making a total of 1114 observations in a balanced dataset shape. In this way, two balanced datasets were made to study the features importance and disease classification in an efficient manner.

3.3. Feature correlation analysis

Feature correlation is a method which helps in understanding the underlying relationships between various data features present in a dataset. Feature correlation can be useful in many ways such as determining the inter-dependencies between the data features and how each feature effects the output feature [31]. We obtained the correlation values between the data features by calculating the correlation coefficients of the feature matrix M having dimension $p \times q$, denoted as: $M = [v_1, v_2, \ldots, v_q]$, where $v_1, v_2, \ldots, v_q$ are the vectors having q number of features. p indicates the length of the vector, where each vector is a complete medical procedure at a specific time. The computed correlation values between different medical features and the target disease for each dataset are shown Fig. 2.



(a) CVD Dataset



(b) Framingham Dataset

Fig. 2. The correlation values for each medical features and the target heart disease for both datasets.

As we can see from Fig. 2, of the 11 features of CVD dataset 4 features are having a positive correlation with the decision feature *i.e.*, 'stroke'. Features 'age', 'hypertension', 'heart_disease' and 'avg_glucose_lvl' are having values 0.57, 0.24, 0.27 and 0.2 when correlated with 'stroke' showing a significant correlation. Similarly, for framingham dataset, features 'age', 'sysBP', 'prevalentHyp', 'diabBP' and 'glucose' showed positive values of 0.23, 0.22, 0.18, 0.15 and reflect the motif of the desired output feature 'TenYearCHD'. For both the datasets, features like 'gender', 'bmi', 'heart rate' and other non-medical features like smoking habits, education, social status and living standards showed very less correlation with the output feature having no or very less effect on the output. Overall, the common medical features like 'age', 'hypertension' and 'glucose' in both datasets are closely related with the outcome and can be considered as the important risk factors.

As per medical research findings, with aging, major changes can be observed in the heart and blood vessels. For example, the heartbeat rate is not as fast during any physical activity as it could when you are younger. The age-related changes may raise a person's risk of heart disease according to National Heart, Lung, and Blood Institute Trusted Source [32]. Hypertension is an established risk factor for stroke, ischemic heart disease and renal dysfunction [33]. Hypertension

causes the blood pressure over the normal range. The higher blood pressure levels make the arteries less elastic and decreases the oxygen and blood flow towards the heart which potentially leads to a heart disease. The diabetic patients are more likely to develop heart disease at an earlier stage. High blood glucose from diabetes causes stronger contraction of blood vessels that control your heart and blood vessels which leads to heart disease [34]. Over time, this process can lead to a heart stroke.

3.4. Feature selection

The main motivation of this research is to select the medical features that can improve the accuracy of heart disease prediction. Feature selection is the process of selecting a subset of most relevant features from a larger collection of original features, that influence the outcome most. The advantages of feature selection includes: data quality improvement, less computational time by prediction model, predictive performance improvement and efficient data collection process.

In this work we have used a filter-based feature selection technique namely, ANOVA-F test to identify most important features from both datasets. Filter-based feature selection techniques employ the use of statistical methods such as similarity, dependence, information, distance to point out the important dependencies or correlation between the input and the target features [35]. Analysis of Variance (ANOVA) is a collection of parametric statistical models and their estimation procedures that determines if the means of two or more samples of data originate from the same distribution. F-test also known as F-statistic, is a set of statistical tests that uses some statistical techniques to calculate the ratio of variance values such variance of two separate samples *etc.* The ANOVA method is a type of F-statistic referred here as an ANOVA f-test. It is a univariate statistical test where each feature is compared to the target feature, to see whether there is any statistically significant relationship between them [36]. Mostly, ANOVA is used in such classification tasks where the type of input features is numerical the target feature is categorical.

The ANOVA-F test can be implemented in python language using the f_classif() function provided by scikit-learn library. The f_classif() function is used in selecting the most important features (features with largest values) via the SelectKBest class. SelectKBest is a method made available in the scikit-learn which takes a scoring function and ranks the features by these scores. Here The scoring function is f_classif() *i.e.,* ANOVA-F test and we have defined SelectKBest class to identify most important features from datasets. The equation to obtain ANOVA-F values is given below:

$$variance\_between\_groups = \frac{\sum_{i=1}^{j} j_i(\bar{K}_i - \bar{K})^2}{(S-1)}$$

$$variance\_within\_groups = \frac{\sum_{i=1}^{S}\sum_{p=1}^{j_i}(K_{ip} - \bar{K}_i)^2}{(N-S)}$$

$$F\_value = \frac{variance\_between\_groups}{variance\_within\_groups}$$

where N is the overall sample size, S is the number of groups, j_i is the number of observations in the jth group, $\bar{K}_i$ is the ith group sample mean, $\bar{K}$ is the overall mean of the data, K_{ip} is the pth observation in the ith out of S groups

The feature importance scores obtained using the ANOVA-F test are shown in Fig. 3(a) and (b) for both datasets. According to the statistics in Fig. 3(a), the most important features for predicting 'stroke' are 'age', 'hypertension', 'heart_disease' and 'avg_glucose_lvl' possessing suitable scores when related with the outcome. However, features 'gender', 'bmi', 'residence_type' and 'smoking_status' showed less or 0 significance for the feature 'stroke'. Looking at 3 (b), we can observe that features 'age', 'prevalentHyp', 'diabetes', 'sysBP', 'diaBP' and 'glucose' obtains highest scores as compared to the other features of the dataset when related to 'TenYearCHD'. Looking at the importance values of the features for each dataset, we can observe a similarity with the correlation results listed in *i.e.*, in most cases the features related with age, hypertension, glucose, blood pressure has a significant influence in the prediction of the heart disease. Similarly, the features identified using ANOVA-F test are also listed as the potential risk factors for heart disease as cited by the American Heart Association [37].

4. Evaluation matrices

We have used three popular performance evaluation metrices *i.e.,* Accuracy, F1-score and ROC to evaluate the performance of ML classification models [38]. Confusion matrix is a table that helps ML practitioners to describe the performance of a classification model. Confusion matrix consists of four used to determine the performance matrices of a classifier and can be described as (1) True Positive (TP) test result that correctly classify the presence of heart disease in patient, (2) True Negative (TN) test result that correctly classify the absence of heart disease in patient, (3) False Negative (FN) test result that wrongly classify that a particular patient does not have heart disease and (4) False Positive (FP) test result which wrongly classify that a particular patient has heart disease. In medical field, FN are considered as most harmful predictions. For a given dataset of size n, accuracy is measured as

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

F1-Score is the harmonic mean of Precision and Recall.

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 - Score = 2(Precision \times Recall)/(Precision \times Recall)$$

The Receiver Optimistic Curves (ROC) examine the classification capability of a classification model. It evaluates the "true positive rate" and "false positive rate" in a ML model output.

$$TPR = TP/(TP + FN)$$

$$FPR = FP/(FP + TN)$$

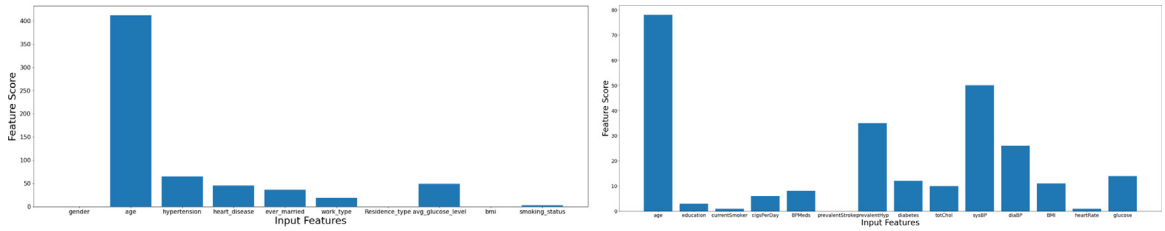
5. Results and discussions

In this section, we will discuss the performance of the selected classification models from different perspectives. First, we checked the performance of model individually for both datasets with full features to examine which models work well for each dataset. Secondly, we evaluated the performance of the models on the selected set of feature to analyze the effect of feature selection technique on the accuracy of the classifiers. The classifiers performance was checked using the Accuracy, F1-score and ROC evaluation matrices.

5.1. Classification results using full feature set

In this section, all the ML models were tested on both datasets using full set of features to predict the binary disease outcome. We trained all the prediction models on entire data with 80% training and 20% testing subsets. The overall computational time consumed during the training of prediction models was 10.98 iterations per second (it/s) for CVD dataset and 24.20 iterations per second (it/s) using framingham dataset. Table 3 and 4 shows the binary classification results of the ML model in predicting the heart disease for both datasets.

Looking at the classification results listed In Table 3, the highest accuracy reported was 0.73 achieved by MLP for CVD dataset with ROC of 0.74 and F1-Score of 0.73. Along with MLP other classifiers like LR, SVC and RF worked well and provided reasonable prediction accuracy with full feature set. The reason behind the improved accuracy achieved by MLP is that it is good at discovering patterns from



(a) CVD Dataset (b) Framingham Dataset

Fig. 3. Feature importance scores for each feature in both datasets.

Table 3
Classification results of various ML models for CVD dataset using full feature set.

Model	Accuracy	Balanced accuracy	ROC AUC	F1-Score
Perceptron	0.73	0.74	0.74	0.73
SGD Classifier	0.72	0.73	0.73	0.71
Logistic Regression	0.73	0.73	0.73	0.73
Quadratic Discriminant Analysis	0.72	0.73	0.73	0.72
Linear SVC	0.72	0.72	0.72	0.72
SVC	0.71	0.72	0.72	0.71
Nu SVC	0.71	0.72	0.72	0.71
Nearest Centroid	0.71	0.72	0.72	0.71
Calibrated Classifier CV	0.71	0.72	0.72	0.71
Bernoulli NB	0.71	0.72	0.72	0.71
Gaussian NB	0.71	0.71	0.71	0.71
Passive Agressive Classifier	0.71	0.71	0.71	0.71
Ridge Classifier CV	0.70	0.71	0.71	0.70
Ridge Classifier	0.70	0.71	0.71	0.70
Linear Discriminant Analysis	0.70	0.71	0.71	0.70
Random Forest Classifier	0.70	0.70	0.70	0.70
AdaBoost Classifier	0.70	0,70	0.70	0.70
KNeighbors Classifier	0.69	0.69	0.69	0.69
Bagging Classifier	0.68	0.68	0.68	0.68
Extra Tree Classifier	0.66	0.66	0.66	0.66
LGBM Classifier	0.65	0.66	0.66	0.65
XGB Classifier	0.65	0.65	0.65	0.65
Decision Tree Classifier	0.61	0.61	0.61	0.61
Label Spreading	0.60	0.60	0.60	0.60
Label Propagation	0.60	0.59	0.59	0.60
Dummy Classifier	0.46	0.46	0.46	0.46

Table 4
Classification results of various ML models for framingham dataset using full feature set.

Model	Accuracy	Balanced accuracy	ROC AUC	F1-Score
Linear SVC	0.66	0.67	0.67	0.66
Linear Discriminant Analysis	0.66	0.67	0.67	0.66
Calibrated Classifier CV	0.66	0.67	0.67	0.66
Ridge Classifier CV	0.66	0.67	0.67	0.66
Ridge Classifier	0.66	0.67	0.67	0.66
Logistic Regression	0.65	0.66	0.66	0.65
Nearest Centroid	0.64	0.66	0.66	0.64
KNeighbors Classifier	0.64	0.65	0.65	0.64
Random Forest Classifier	0.64	0.65	0.65	0.64
Bernoulli NB	0.63	0.64	0.64	0.63
LGBM Classifier	0.63	0.64	0.64	0.63
AdaBoost Classifier	0.63	0.64	0.64	0.63
Extra Tree Classifier	0.62	0.63	0.63	0.62
XGB Classifier	0.62	0.63	0.63	0.61
Decision Tree Classifier	0.61	0.62	0.62	0.61
Bagging Classifier	0.60	0.61	0.61	0.59
SGD Classifier	0.60	0.60	0.60	0.60
Gaussian NB	0.57	0.60	0.60	0.53
Passive Agressive Classifier	0.58	0.59	0.59	0.57
Nu SVC	0.59	0.59	0.59	0.59
Extra Tree Classifier	0.59	0.59	0.59	0.59
SVC	0.58	0.58	0.58	0.58
Quadratic Discriminant Analysis	0.55	0.58	0.58	0.51
Perceptron	0.57	0.55	0.55	0.55
Label Propagation	0.54	0.54	0.54	0.53
Label Spreading	0.53	0.53	0.53	0.53
Dummy Classifier	0.52	0.52	0.52	0.52

Table 5
Classification results of various ML models for CVD dataset using reduced feature set.

Model	Accuracy	Balanced accuracy	ROC AUC	F1-Score
SVC	0.74	0.75	0.74	0.74
Nearest Centroid	0.74	0.75	0.74	0.74
Logistic Regression	0.73	0.74	0.73	0.73
SGD Classifier	0.73	0.73	0.73	0.73
Linear SVC	0.72	0.73	0.72	0.73
Linear Discriminant Analysis	0.72	0.73	0.72	0.73
Ridge Classifier CV	0.72	0.73	0.72	0.73
Ridge Classifier	0.72	0.73	0.72	0.73
Quadratic Discriminant Analysis	0.72	0.73	0.72	0.72
Calibrated Classifier CV	0.72	0.73	0.72	0.72
Label Spreading	0.71	0.71	0.71	0.71
Bagging Classifier	0.70	0.70	0.70	0.70
AdaBoost Classifier	0.70	0.71	0.70	0.71
Label Propagation	0.70	0.70	0.70	0.70
Bernoulli NB	0.70	0.70	0.70	0.70
Nu SVC	0.70	0.70	0.70	0.70
LGBM Classifier	0.70	0.70	0.70	0.70
Extra Trees Classifier	0.69	0.70	0.69	0.69
XGB Classifier	0.69	0.70	0.69	0.69
Random Forest Classifier	0.69	0.70	0.69	0.69
Gaussian NB	0.69	0.69	0.69	0.69
Decision Tree Classifier	0.68	0.69	0.68	0.69
Extra Tree Classifier	0.68	0.69	0.68	0.69
KNeighbors Classifier	0.67	0.68	0.67	0.68
Perceptron	0.64	0.64	0.64	0.64
Passive Agressive Classifier	0.63	0.63	0.63	0.63
Dummy Classifier	0.50	0.50	0.50	0.50

complex medical datasets. Furthermore, this network model is good at generalizing data without having the prior domain knowledge. The worst results were obtained by the dummy classifier *i.e.*, only a 0.46 accuracy when predicting heart stroke. Possible reasons behind poor classification result is that the dummy classifier makes predictions using simple rules which is not useful when dealing with real world problems. The classification results with the same techniques for the framingham dataset are shown in Table 4. The accuracy results were not very good as the highest accuracy achieved was 0.66 with 0.67 ROC and 0.66 F1-score. Other algorithms like Linear Discriminant Analysis (LDA), LR and ridge classifier performed the similar. The reason behind weak results might be the range of values between the data features. Feature scaling helps in normalizing the data within a particular range, which can improve the results of the models in general [39,40]. However, any data manipulation strategy in medical studies may introduce significant biases, that is why we have kept all the feature values unchanged.

5.2. Classification results using reduced feature set

Given the goal of identifying the potential bio-markers and to analyze the impact of feature selection technique on the classification accuracy, we selected the most prominent features from the full feature space based on individual feature scores. The features impacting the outcome most for each dataset were identified by ANOVA-F test as shown in Fig. 3(a) and (b). As per the feature scores, 4 features *i.e.,* {age, hypertension, heart_disease, avg_glucose_lvl} were selected for CVD dataset out of 11. Only 5 features out of 15 from framingham dataset *i.e.,* {age, prevalentHyp, sysBp, diaBp, glucose} were chosen considering the feature weights obtained using ANOVA-F test. We evaluated the performance of each classification model using only the selected features as inputs. Table 5 shows the classification performance of each model using the reduced feature subset from CVD dataset. The analysis showed that even after limiting the number of features, ML models showed better performance as compared to the models using full feature set. The highest accuracy achieved was 0.74 by SVC model with 0.74 F1-Score and 0.74 ROC with only 4 input features. Considering Table 6 results, the highest accuracy achieved is 0.71, which is higher than all the accuracy results using full feature set for framingham dataset. Furthermore, the models trained on reduced feature set also consumed less computational time *i.e* only 3.86 iterations per second(it/s) using CVD and 15.52 iterations per second(it/s) using framingham dataset. We have also validated our findings by comparing our work with other published proposals [41,42] where same datasets were used with full feature set and the obtained accuracy results were less or equal to the results that we obtained using reduced feature set. Overall, the experimental results proved that the performance of the ML models increased significantly by using only the relevant features. Furthermore, during the training of classification models using the reduced feature set, a less computational iterations per second (it/s) were observed. These experimental results clear the concepts about the impact of feature selection techniques, that it not only reduces the size feature space, but it also improves performance of ML models also in various aspects.

6. Conclusion and future works

Heart disease is the most fatal disease which is rapidly increasing and became one of the causes of death around the world. The damage caused by this disease can be reduced significantly, if adequate treatment procedures are applied at the early stages. This paper studies the prediction of heart disease and the selection of the important features. The main goal of this research study is to observe the impact of feature selection techniques on the performance of ML models. This analysis was performed for CVD and Framingham heart disease datasets which are available online. In this research, first, we performed a data preprocessing step in which data transformation, cleansing and balancing were involved. Secondly, we used a filter-based feature selection technique namely the ANOVA-F test to identify the most important features from the datasets for an effective heart disease prediction. Using the ANOVA-F test most relevant features with outcomes from both datasets were identified using the individual feature scores. We observed that features like age, hypertension, glucose, previous heart disease, and blood pressure were found to reflect the most important risk factors for heart disease except the traditional factors using both datasets. Furthermore, the classification experiments were performed with full as well as the reduced feature sets to analyze the effect of selected features on the prediction accuracy of various ML prediction models. Using the full feature set the highest accuracy achieved was 0.73 for CVD and 0.66 for the Framingham heart disease dataset. After using the reduced feature set the accuracy increased to 0.75 and 0.71 for

Table 6
Classification results of various ML models for framingham dataset using reduced feature set.

Model	Accuracy	Balanced accuracy	ROC AUC	F1-Score
Perceptron	0.71	0.72	0.72	0.71
AdaBoost Classifier	0.71	0.71	0.71	0.71
SGD Classifier	0.69	0.69	0.69	0.69
Logistic Regression	0.69	0.69	0.69	0.69
Bernoulli NB	0.68	0.69	0.69	0.68
Linear Discriminant Analysis	0.68	0.68	0.68	0.68
Ridge Classifier CV	0.68	0.68	0.68	0.68
Ridge Classifier	0.68	0.68	0.68	0.68
Linear SVC	0.68	0.68	0.68	0.68
Calibrated Classifier CV	0.68	0.68	0.68	0.68
Gaussian NB	0.66	0.68	0.68	0.65
SVC	0.67	0.67	0.67	0.67
Nearest Centroid	0.66	0.67	0.67	0.66
KNeighbors Classifier	0.65	0.66	0.66	0.65
Bagging Classifier	0.63	0.64	0.64	0.63
Quadrant Discriminant Analysis	0.62	0.64	0.64	0.58
Decision Tree Classifier	0.62	0.62	0.62	0.61
Extra Tree Classifier	0.62	0.62	0.62	0.62
Random Forest Classifier	0.62	0.62	0.62	0.62
Label Spreading	0.59	0.59	0.59	0.59
Label Propagation	0.58	0.58	0.58	0.58
LGBM Classifier	0.57	0.58	0.58	0.57
Nu SVC	0.57	0.58	0.58	0.57
XGB Classifier	0.57	0.57	0.57	0.57
Passive Agressive Classifier	0.57	0.56	0.56	0.57
Dummy Classifier	0.55	0.56	0.56	0.55
Extra Tree Classifier	0.51	0.51	0.51	0.51

both datasets. The analysis showed that even after limiting the number of features, ML models showed better performance as compared to the models using a full feature set. The experimental results reveal that by employing a feature selection technique, we may accurately classify the heart disease even with a small number of features and less time. We can conclude that using the feature selection only the most important features related to heart disease are selected which reduces the computational complexities and improve the accuracy of prediction models. In the intended future work, we will try to work on enhancing the prediction accuracy by using a vast combination of ML and deep learning models [43] to obtain the best feasible model for the heart disease diagnosis. We will benchmark our analysis on additional datasets as a part of our future work. We will also try to use more than one feature selection technique to obtain more feasible feature subsets which are more direct with medical studies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology grant number 13/RC/2106_P2.

References

[1] S. Nalluri, R.V. Saraswathi, S. Ramasubbareddy, K. Govinda, E. Swetha, Chronic heart disease prediction using data mining techniques, in: Data Engineering and Communication Technology, Springer, 2020, pp. 903–912.
[2] R.A. Manji, J. Witt, P.S. Tappia, Y. Jung, A.H. Menkis, B. Ramjiawan, Cost–effectiveness analysis of rheumatic heart disease prevention strategies, Expert Rev. Pharmacoecon. Outcomes Res. 13 (6) (2013) 715–724.
[3] P. Saranya, P. Asha, Survey on big data analytics in health care, in: 2019 International Conference on Smart Systems and Inventive Technology, ICSSIT, IEEE, 2019, pp. 46–51.
[4] G. Sivapalan, K.K. Nundy, S. Dev, B. Cardiff, D. John, ANNet: a lightweight neural network for ECG anomaly detection in IoT edge sensors, IEEE Transactions on Biomedical Circuits and Systems 16 (1) (2022) 24–35.
[5] A.U. Haq, J.P. Li, M.H. Memon, S. Nazir, R. Sun, A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms, Mob. Inf. Syst. 2018 (2018).
[6] A. Gavhane, G. Kokkula, I. Pandya, K. Devadkar, Prediction of heart disease using machine learning, in: 2018 S International Conference on Electronics, Communication and Aerospace Technology, ICECA, IEEE, 2018, pp. 1275–1278.
[7] N.K. Kumar, G.S. Sindhu, D.K. Prashanthi, A.S. Sulthana, Analysis and prediction of cardio vascular disease using machine learning classifiers, in: 2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS, IEEE, 2020, pp. 15–21.
[8] S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha, Privacy risk in machine learning: Analyzing the connection to overfitting, in: 2018 IEEE 31st Computer Security Foundations Symposium, CSF, IEEE, 2018, pp. 268–282.
[9] O.O. Aremu, D. Hyland-Wood, P.R. McAree, A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data, Reliab. Eng. Syst. Saf. 195 (2020) 106706.
[10] S. Manandhar, S. Dev, Y.H. Lee, S. Winkler, Y.S. Meng, Systematic study of weather variables for rainfall detection, in: Proc. IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2018, pp. 3027–3030.
[11] V. Pavithra, V. Jayalakshmi, Review of feature selection techniques for predicting diseases, in: Proc. 5th International Conference on Communication and Electronics Systems, ICCES, IEEE, 2020, pp. 1213–1217.
[12] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79.
[13] G. Wang, F. Lauri, A.H. El Hassani, A study of dimensionality reduction's influence on heart disease prediction, in: 2021 12th International Conference on Information, Intelligence, Systems & Applications, IISA, IEEE, 2021, pp. 1–6.
[14] B. Remeseiro, V. Bolon-Canedo, A review of feature selection methods in medical applications, Comput. Biol. Med. 112 (2019) 103375.
[15] M.S. Pathan, Z. Jianbiao, D. John, A. Nag, S. Dev, Identifying stroke indicators using rough sets, IEEE Access 8 (2020) 210318–210327.
[16] C.S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, D. John, Predicting stroke from electronic health records, in: Proc. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 5704–5707.
[17] Y. Zhang, Y. Zhou, D. Zhang, W. Song, A stroke risk detection: improving hybrid feature selection method, J. Med. Internet Res. 21 (4) (2019) e12437.
[18] H.M. Le, T.D. Tran, L. Van Tran, Automatic heart disease prediction using feature selection and data mining technique, J. Comput. Sci. Cybern. 34 (1) (2018) 33–48.
[19] Y. Zhang, W. Song, S. Li, L. Fu, S. Li, Risk detection of stroke using a feature selection and classification method, IEEE Access 6 (2018) 31899–31907.
[20] M. Al Mehedi Hasan, J. Shin, U. Das, A. Yakin Srizon, Identifying prognostic features for predicting heart failure by using machine learning algorithm, in: 2021 11th International Conference on Biomedical Engineering and Technology, 2021, pp. 40–46.

[21] M.S. Singh, P. Choudhary, Stroke prediction using artificial intelligence, in: 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference, IEMECON, IEEE, 2017, pp. 158–161.
[22] C.B. Gokulnath, S. Shantharajah, An optimized feature selection based on genetic approach and support vector machine for heart disease, Cluster Comput. 22 (6) (2019) 14777–14787.
[23] D. Zhang, Y. Chen, Y. Chen, S. Ye, W. Cai, J. Jiang, Y. Xu, G. Zheng, M. Chen, Heart disease prediction based on the embedded feature selection method and deep neural network, J. Healthcare Eng. 2021 (2021).
[24] N. Hasan, Y. Bao, Comparing different feature selection algorithms for cardiovascular disease prediction, Health Technol. 11 (1) (2021) 49–62.
[25] K.V.V. Reddy, I. Elamvazuthi, A.A. Aziz, S. Paramasivam, H.N. Chua, S. Pranavanand, Heart disease risk prediction using machine learning classifiers with attribute evaluators, Appl. Sci. 11 (18) (2021) 8352.
[26] J. Huang, Y.-F. Li, M. Xie, An empirical analysis of data preprocessing for machine learning-based software cost estimation, Inf. Softw. Technol. 67 (2015) 108–127.
[27] S. Sachan, F. Almaghrabi, J.-B. Yang, D.-L. Xu, Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance, Expert Syst. Appl. 185 (2021) 115597.
[28] H. Wang, J. Tang, M. Wu, X. Wang, T. Zhang, Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example, BMC Med. Inf. Decis. Making 22 (1) (2022) 1–14.
[29] M.R. Stavseth, T. Clausen, J. Røislien, How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data, SAGE Open Med. 7 (2019) 2050312118822912.
[30] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw. 106 (2018) 249–259.
[31] N. Gopika, A.M.K. ME, Correlation based feature selection algorithm for machine learning, in: 2018 3rd International Conference on Communication and Electronics Systems, ICCES, IEEE, 2018, pp. 692–695.
[32] R.G. Williams, G.D. Pearson, R.J. Barst, J.S. Child, P. Del Nido, W.M. Gersony, K.S. Kuehl, M.J. Landzberg, M. Myerson, S.R. Neish, et al., Report of the national heart, lung, and blood institute working group on research in adult congenital heart disease, J. Am. College Cardiol. 47 (4) (2006) 701–707.
[33] E. Escobar, Hypertension and coronary heart disease, J. Hum. Hypertens. 16 (1) (2002) S61–S63.
[34] R. Huxley, F. Barzi, M. Woodward, Excess risk of fatal coronary heart disease associated with diabetes in men and women: meta-analysis of 37 prospective cohort studies, Bmj 332 (7533) (2006) 73–78.
[35] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, M. Lang, Benchmark for filter methods for feature selection in high-dimensional classification data, Comput. Statist. Data Anal. 143 (2020) 106839.
[36] P. Mishra, U. Singh, C.M. Pandey, P. Mishra, G. Pandey, Application of student's t-test, analysis of variance, and covariance, Ann. Card. Anaesth. 22 (4) (2019) 407.
[37] E.J. Benjamin, P. Muntner, A. Alonso, M.S. Bittencourt, C.W. Callaway, A.P. Carson, A.M. Chamberlain, A.R. Chang, S. Cheng, S.R. Das, et al., Heart disease and stroke statistics—2019 update: a report from the american heart association, Circulation 139 (10) (2019) e56–e528.
[38] S. Dev, F.M. Savoy, Y.H. Lee, S. Winkler, Nighttime sky/cloud image segmentation, in: Proc. IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 345–349.
[39] D. Thara, B. PremaSudha, F. Xiong, Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques, Pattern Recognit. Lett. 128 (2019) 544–550.
[40] M. Jain, T. AlSkaif, S. Dev, Validating clustering frameworks for electric load demand profiles, IEEE Transactions on Industrial Informatics 17 (12) (2021) 8057–8065.
[41] S. Dev, H. Wang, C.S. Nwosu, N. Jain, B. Veeravalli, D. John, A predictive analytics approach for stroke prediction using machine learning and neural networks, Healthc. Anal. 2 (2022) 100032.
[42] J.-J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Koleva, C. Hurtado, M.F. Landecho, Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease), J. Biomed. Inform. 97 (2019) 103257.
[43] B.P. Das, M.S. Pathan, Y.H. Lee, S. Dev, Estimating ground-level nitrogen dioxide concentration from satellite data, in: Proc. Photonics & Electromagnetics Research Symposium (PIERS), IEEE, 2021, pp. 1176–1182.

Biomedical Engineering Advances 3 (2022) 100035

Contents lists available at ScienceDirect

Biomedical Engineering Advances

journal homepage: www.journals.elsevier.com/biomedical-engineering-advances

Classifier identification using deep learning and machine learning algorithms for the detection of valvular heart diseases

Tanmay Sinha Roy [a,*], Joyanta Kumar Roy [b], Nirupama Mandal [c]

[a] Department of Electrical Engineering, Haldia Institute of Technology, Haldia, WB, India
[b] Chairman, Eureka Scientech Research Foundation, Kolkata, India
[c] Department of Electronics Engineering, Indian Institute of Technology (ISM), Dhanbad, India

ARTICLE INFO

Keywords:
PCG Signal Analysis
Valvular Heart Disease
Artificial Intelligence
Deep Learning Neural Network
Machine Learning Algorithms
Feature Extraction
Classification
Acoustic Stethoscope

ABSTRACT

Heart-related disorders are rapidly growing throughout the world. Artificial Intelligence with computational methods plays a significant role in early detection and diagnosis. This study has been devoted to finding the best classifiers for different valvular heart problems using popular CNN-based deep learning models and machine learning algorithms written in Python 3.8. In this research, the CNN-based Xception network model for the first time has been proposed for valvular heart sound analysis, which achieved an accuracy of 99.45% on the test dataset with a sensitivity of 98.5% and specificity of 98.7%. Compared with other deep learning models like LeNet-5, AlexNet, VGG16, VGG19, DenseNet121, Inception Net, and Residual Net, it is observed that accuracy for predicting the prediction of valvular heart disease is the highest, and testing time is the lowest in the proposed modified Xception network model. The features are Root Mean Square, Energy, Power, Zero Crossing Rate, Total Harmonic Distortion, Skewness, and Kurtosis in the time domain. The analysis has been made on heart sounds of normal and diseased patients available from the standard heart sound data repository. Finally, all the evaluated results were compared, and found SVM and Random Forest algorithms are the most effective among machine learning methods. The proposed modified CNN-based Xception model works the best among all deep learning methods.

1. Introduction

The human heart makes sounds. The sounds come from the various atrioventricular valves. As these valves open and close, allowing blood flow to and from the heart, in this course, it produces the heartbeat sound. Analysis of heart sound is fundamental to detect any heart-related disorders. Under the investigation of heart sounds, the classification of heart-related disease is also crucial for quickly taking the correct preventive action. In practice, the background noise signals need to be removed in the valvular heart disease analysis after auscultation is received through an electronic acoustic stethoscope. Then the noise-free heart sound has been digitized for computational needs. The convenient feature extracting algorithms are applied in the computational intelligence to extract the essential features to classify the heart sound for diseases. As heart sounds are generated from the opening & closing of valves, they are repetitive and mechanical vibrations that occur at certain fixed time intervals and are analogous to an electrical signal. The heart sound can be analyzed in the Time and Frequency domain using different developed algorithms and tools applied in various computational intelligence techniques. In order to identify the suitable classifiers, detailed studies based on various Machine learning algorithms are required on normal and abnormal heart sounds. The study aims to search for the best cost-effective, simplified classifier tool for early screening of heart diseases.

2. Related work

In this research, the signal classification of valvular heart disease is essential. According to the literature survey, many researchers have been working in this area. Cota Navin Gupta et al. [1] in 2005 did Segmentation & Classification of heart sound for PCG signal analysis. Talha J. Ahmad et al. [2] in 2009 made PCG Signal analysis & its classification using an Adaptive Fuzzy Inference System based Mamdani type Fuzzy inference classifier. The experiment was done on a standard heart soundbank. It was an offline method and not validated with the human subject. Mandeep Singh et al. [3] in 2013 made work on Heart

* Corresponding author.
E-mail addresses: tanmoysinha.roy@gmail.com (T.S. Roy), jkroy.cal51@gmail.com (J.K. Roy), nirupama_cal@rediffmail.com (N. Mandal).

https://doi.org/10.1016/j.bea.2022.100035
Received 25 November 2021; Received in revised form 28 April 2022; Accepted 9 May 2022
Available online 20 May 2022

Sounds Classification using Feature Extraction of Phonocardiography Signal. However, no data science was involved in that.

Ajay Kumar Roy et al. [4] in 2014 has surveyed the Classification of PCG signals. S.Barma et al. [5] in 2015 has done work on the analysis of second heart sound, where it deals with calculating the duration and the energy of normalized IFs of A2s and P2s, but they cannot classify heart sounds. Siddique Latif et al. [6] in 2018 made work on Phonocardiographic Sensing Using Deep Learning for Abnormal Heartbeat Detection. However, it was used to distinguish between normal and abnormal sounds. Dr. Naveen Kumar Dewangan et al. [7] in 2018 and Gyanaprava Mishra et al. [8] in 2013 have done work on PCG Signal analysis using the DWT method. It has certain limitations in real-time analysis. Joyanta Kumar Roy et al. [18, 26] in 2017 & 2018 did work on a Simple technique for heart sound detection and identification using the Kalman filter in real-time analysis, where various feature extraction techniques were discussed. S.Rubin Bose et al. [39] 2021 did research on the recognition of hand gestures using the xception model and achieved good accuracy. Md. Zabirul Islam et al. [40] 2020 studied covid-19 disease using the deep CNN-LSTM method. Thus for PCG signal analysis, involvement of data science and artificial intelligence is necessary. Table 1 shows some of the recent works that have been done on PCG signal analysis and classification methods.

3. Methods & materials

3.1. Heart sound dataset description

The normal and diseased heart sound samples have been taken from the repository of Yaseen et al. [14] available from https://github.com/yaseen21khan/Classification-of-Heart-Sound-Signal-Using-Multiple-Features-. These sound samples may be considered error-free and usable because it was tested and implemented in SVM and a deep neural network (DNN) to extract Mel Frequency Cepstral Coefficients (MFCCs) and Discrete Wavelets Transform (DWT) features [11,12]. The heart sound dataset comprises 1000 PCG recordings, and it is composed of five categories with different age groups and gender. These PCGs recorded the heart sounds of adults using clinical and non-clinical methods. Their sampling frequency is 2000 Hz. The length of the recordings varies from 2 s to over 5 s. The dataset includes 200 normal sounds, 200 Mitral Regurgitation (MR) samples, 200 Mitral stenosis (MS) samples, 200 Mitral Valve Prolapse (MVP) Samples & 200 Atrial Stenosis samples and is available in ".wav" format. The normal heart sound recordings are taken from healthy subjects, and the abnormal recordings are from subjects with a past medical history of heart disease, as given in Heart Sound Dataset 1 shown in Fig. 1

Heart Sound Dataset 2, as shown in Fig. 2, has been taken from a clinical trial in hospitals using the digital stethoscope, which comes under the Classification of Heart Sound Recordings-Pascal Challenge Dataset B [15, 37]. The length of the recordings varies from 1 s to over 30 s, and their sampling frequency is 2000 Hz. Mainly, three categories of heart sounds have been considered, namely Normal, Murmur and Extra–Systole.

Fig. 3 depicts a Physio Net Challenge [16, 38] training set composed of five training databases (A through E) containing a total of 3126 heart sound samples which varies from 5 s to over 120 s. All samples have been sampled at a sampling rate of 2000 Hz per sec. Mainly, two categories of heart sounds have been considered for analysis, namely Normal and Abnormal.

Heart sounds are stored in the memory in ".wav format." The developed python program fetches the heart sound data stored in the memory and consequently processes them through deep learning models and various machine learning methods for predicting the type of heart disease.

Pre-processing: Heart sound is fed to pre-processing block, where it uses a bandpass filter having a bandwidth of 30 to 500 Hz to get rid of the noise. A window frame of 5 s is selected and kept as fixed for every heart sound sample undergoing preprocessing. Different features have been extracted from the preprocessed signal, and finally, classification [13] of the signal takes place for evaluation of the proposed model.

$$x'(t) = f(x(t)) \tag{1}$$

Where x'(t) is the filtered heart sound signal.

Each dataset has been divided into training data (85%) and test data (15%) as per standard practice. Further, training data is decomposed

Category	No. of Samples
Normal	200
Mitral Regurgitation (MR)	200
Mitral Stenosis (MS)	200
Mitral Valve Prolapse (MVP)	200
Atrial Stenosis (AS)	200

Fig. 1. Heart Sound Dataset1.

Table 1
RECENT WORK DONE ON PCG SIGNAL ANALYSIS AND CLASSIFICATION METHODS.

S. No	Reference	Method	Features Used	Segmentation	Optimizer	Types of Heart Sound	Accuracy on Test Dataset
1	Baghel et al., 2020 [21]	1D-CNN	1D time-series signals	No	SGD	Normal, MR, MVP, MS & AS	98.60%
2	Humayun et al., 2020 [23]	1D-CNN	1D time-series signals	Yes	SGD	N, A	83.29%
3	Xu et al., 2018 [17]	1D-CNN	1D time-series signals	No	SGD	N, A	93.25%
4	Shu Lih Oh et al., 2020 [28]	1D-CNN Wavelet	1D time-series signals	No	Adam	Normal, MR, MVP, MS & AS	97.00%
5	Noman et al., 2019 [27]	Ensemble CNN	1D time-series signals+ MFCC	Yes	Adam	N, A	89.22%
6	Dominguez et al., 2018 [22]	2D-CNN	Spectrograms	No	Adam	N, A	97.05%
7	Xiao et al., 2020 [32]	1D-CNN	1D time-series signals	No	SGD	N, A	93.00%
8	Khan et al., 2020 [33]	LSTM	MFCC	No	Adam	N, A	91.39%
9	Li et al., 2019 [34]	1D -CNN	Spectrograms	No	Adam	N, A	96.48%
10	Wu et al., 2019 [35]	Ensemble CNN	Spectrograms + MFSC+ MFCC	No	Adam	N, A	87.91%
11	Yang et al., 2016 [36]	RNN	1D time-series signals	No	SGD	N, A	82.87%
12	Yaseen et al., 2018 [14]	SVM + DNN	MFCC + DWT	No	Adam	Normal, MR, MVP, MS & AS	87.08%

*Abbreviations—N: normal heart sounds, M: murmur heart sounds, EXT: extra systole heart sounds, AS: aortic stenosis, MS: mitral stenosis, MR: mitral regurgitation, MVP: mitral valve prolapse, MS: mitral stenosis.

Category	No. of Samples
Normal	320
Murmur	95
Extra-Systole	46

Fig. 2. Heart Sound Dataset2.

Category	No. of Samples	
	Normal	Abnormal
Training-A	117	292
Training-B	386	104
Training-C	07	24
Training-D	27	28
Training-E	1958	183
Training-F	80	34

Fig. 3. Heart Sound Dataset3.

into validation data (15%) and the rest for training the model.

Features of the heart sound considered for the entire study has been limited to:

1 Root Mean Square (RMS)
2 Signal Energy and Power
3 Zero-Crossing Rate (ZCR)
4 Total Harmonic distortion(THD)
5 Skewness and Kurtosis

Following Deep Learning Methods are used to classify the above features:

1 LeNet-5
2 AlexNet
3 VGG16
4 VGG19
5 DenseNet121
6 Inception Net
7 Residual Network
8 **Proposed Modified Xception Network**

Following Machine Learning Algorithms are used to classify the above features:

1 K-Nearest Neighborhood (KNN)
2 Support Vector Machine (SVM)
3 Random Forest
4 Naive Bayes
5 Artificial Neural Network (ANN)

All classifier tools are written in Python ver. 3.8 for the ease of convenience as it is open-sourced and popular for having various machine learning libraries. The detailed method of analysis with the classifier as mentioned above tools are written para wise given below.

3.2. CNN based deep learning neural network

3.2.1. Proposed modified xception model

A Deep Neural Network (DNN) [16, 7] has multiple hidden layers between the input and output layers. It can also be able to model complex real-time scenarios from sample data. Neural network [17, 8, 29] is widely used in supervised learning and Unsupervised learning-related problems. A neural network responds to a set of inputs, performs complex operations on them, and finally gives output to solve real-world problems like classification. **Adam (adaptive moment estimation)** is the main optimization algorithm used in training Deep Learning models. A Python-based Keras sequential model [17] has been taken for implementation and the block diagram of PCG signal classification [9,10] is shown in Fig. 4

The proposed block diagram of the modified CNN-based Deep Learning Model is shown in Fig. 6. The fundamental structure of the inception module (Fig. 5) contains entry flow, middle flow, and exit flow. The entry flow uses the concept of depth-wise separable convolution that produces less time complexity than conventional convolution. The middle flow is repeated two times, and the exit flow contains a global average pooling layer followed by two fully connected layers and a softmax function.

In this proposed modified Xception network, the first hidden layer contains 256 filters, each having kernel size 3 using the ReLU activation function, followed by the second hidden layer containing another 256 filters of kernel size 3. The third and fourth hidden layers include the xception module, followed by global average pooling and fully connected layers of 1200 & 150 nodes. The output layer contains five nodes using the Softmax activation function to classify five different types of heart sounds [18].

Learning Curves is obtained for this Xception model developed with normal and abnormal heart sounds, and they have been plotted accordingly, as shown in Fig. 7 and Fig. 8. The proposed model has been trained with 100 epochs as it can be observed that with the increase in epoch, the loss, i.e., Cross-Entropy [19, 25], decreases close to Zero, as shown in Fig. 7.

From the Model Accuracy Curve in Fig. 8, it can be seen that with the increase in epoch, the accuracy [19, 30] also increases close to one. After training the model, an accuracy of approximately 99.51% is obtained, as shown in Table 3.

Evaluation Metrics are different types of measures to evaluate the performance of a deep learning model. They are mainly Accuracy, Precision, Recall, and F-Measure. To measure these parameters, the number of true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) values are required, as mentioned below.

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{FP} + \text{FN} + \text{TN} \tag{2}$$

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{3}$$

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{4}$$

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}) \tag{5}$$

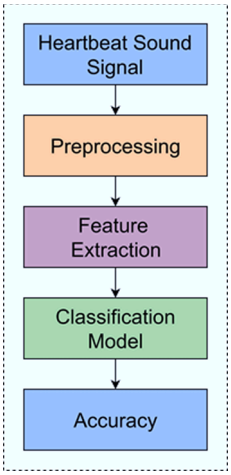


Fig. 4. Block Diagram of PCG Signal classification.

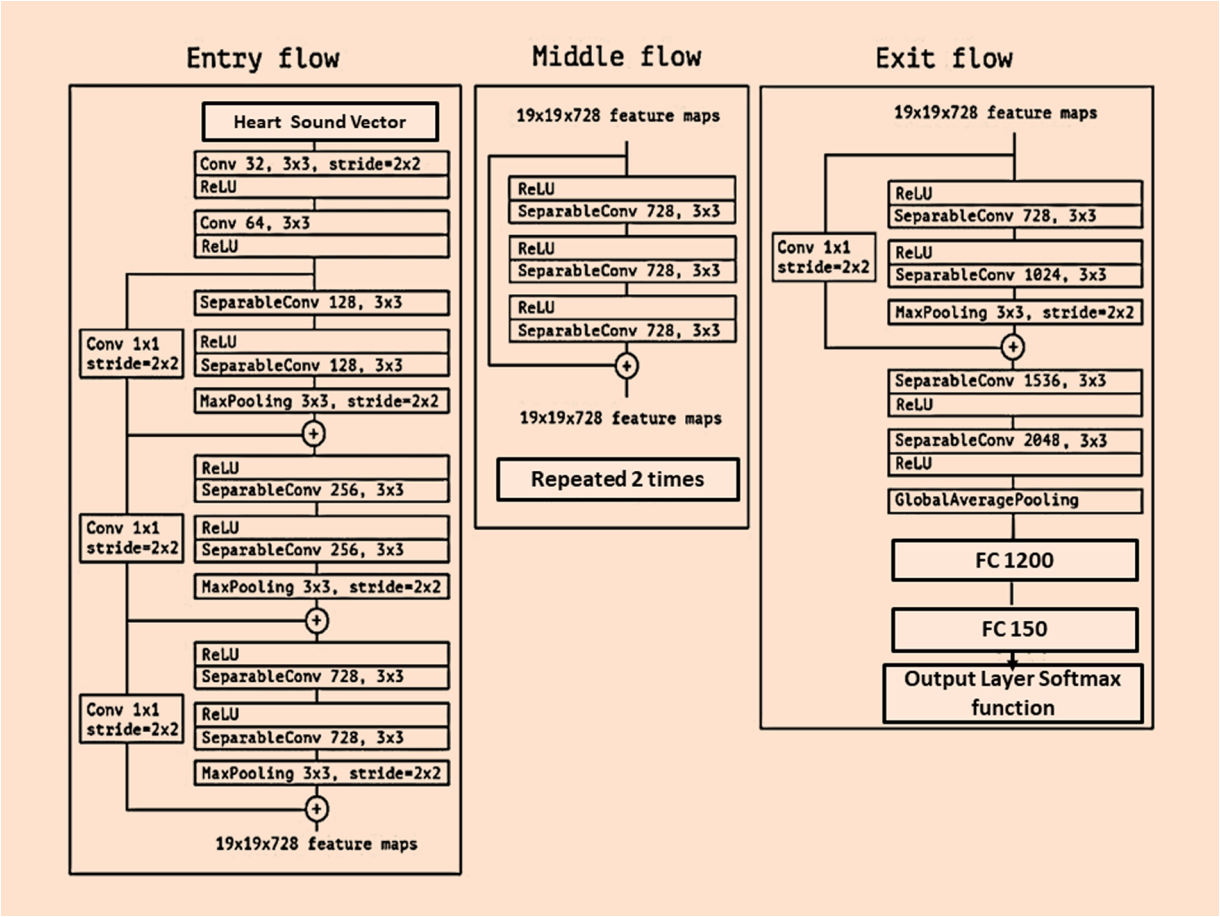


Fig. 5. Proposed modified CNN Based Xception Net Model.

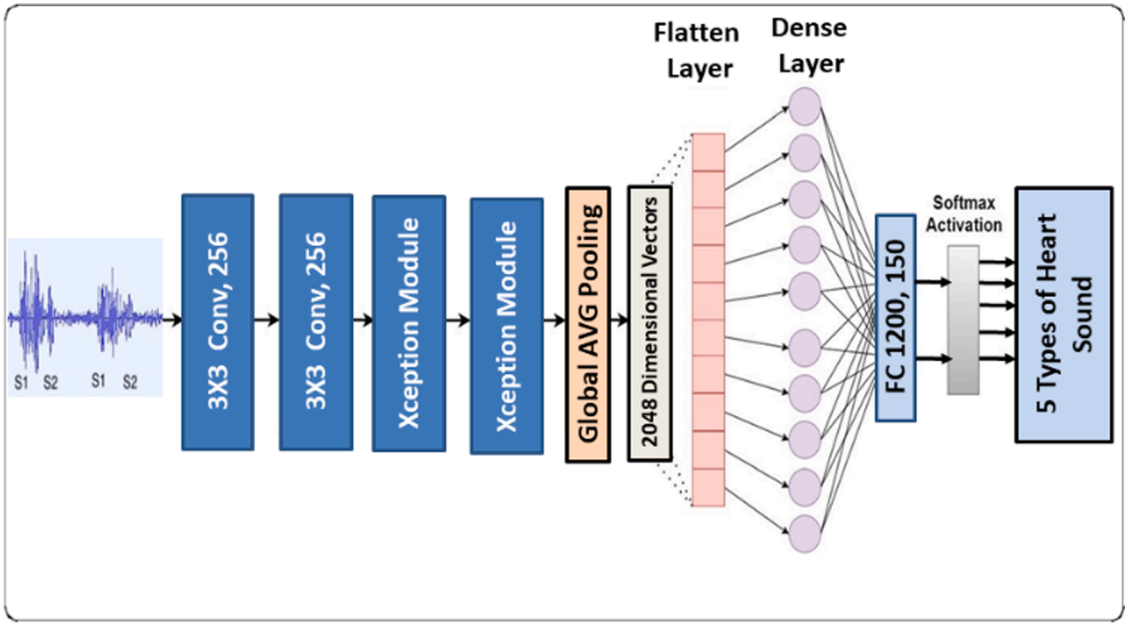


Fig. 6. Block Diagram of the Proposed modified CNN Based Xception Net Model.

The performance evaluation of the proposed CNN-based xception network is carried on 15% of test data from the same heart sound bank given in Table 2.

Different CNN-based deep learning models are considered starting from LeNet-5 to the proposed Xception network model. All of them have been trained using the same datasets, and the results have been obtained, as shown in Table 3.

All the CNN-based deep learning models are compared by evaluating

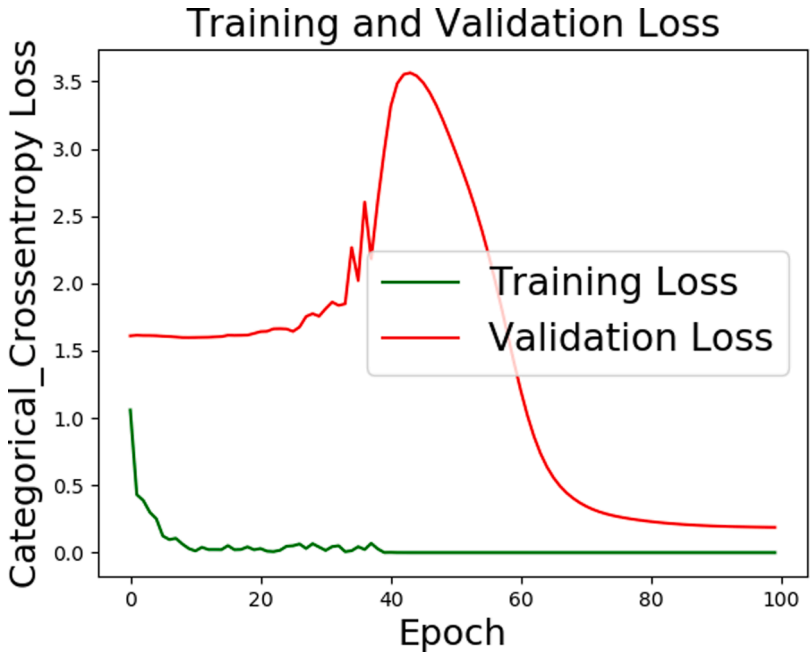


Fig. 7. Plot of Cross-Entropy Loss vs. Epoch.

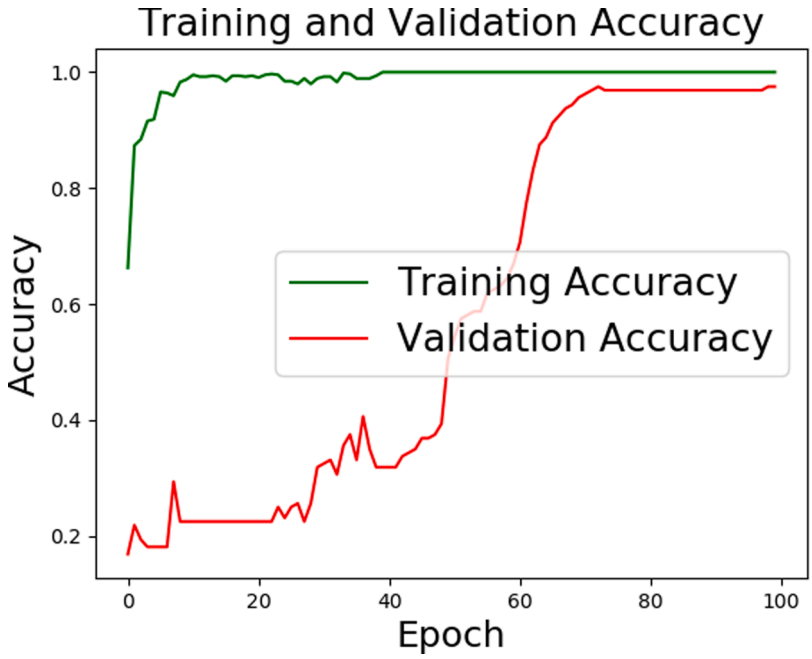


Fig. 8. Plot of Accuracy vs. Epoch in CNN.

Table 2
Summary of Test-Data result in Xception Network.

Model	Accuracy	Precision	Recall	F1-Score
Xception	0.994	0.985	0.987	0.993

different performance metrics, as shown in Table 4.

3.3. Comparison of machine learning methods

3.3.1. K –Nearest neighborhood method

A supervised machine learning [20, 31] algorithm depends on

Table 3
Comparison of different deep learning models applied on Dataset1.

Model	Training loss	Training accuracy	Validation loss	Validation accuracy
LeNet-5	0.4455	0.7098	0.4675	0.6898
Alex Net	0.3567	0.7456	0.3923	0.7256
VGG16	0.2987	0.8090	0.3242	0.7876
VGG19	0.2712	0.8694	0.2987	0.8570
DenseNet121	0.2476	0.9087	0.2656	0.8776
Inception Net	0.0472	0.9843	0.0654	0.9863
Residual Net	0.0432	0.9856	0.0533	0.9892
Xception Net	0.0320	0.9951	0.0325	0.9926

Table 4
Comparison of performance metrics of different deep learning models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Training time (s)	Testing time (s)
LeNet-5	68.98	69.76	67.45	66.67	1087	1134
Alex Net	72.35	70.73	73.87	71.23	1232	1336
VGG16	74.09	75.28	75.09	75.14	1353	1012
VGG19	82.18	83.34	82.18	84.22	1344	946
DenseNet121	92.48	93.54	93.48	94.49	1200	936
Inception Net	96.96	98.17	98.96	97.02	710	974
Residual Net	97.32	98.42	98.32	98.35	783	1056
Xception Net	99.43	98.58	98.74	99.39	750	865

labeled input data to learn a function that produces an appropriate output when given new unlabeled data. The KNN algorithm [20, 23] considers that similar things exist close. In other words, similar things are near each other, staying in the same group.

Fig. 9 represents the flowchart of KNN used in the research, which is available in the machine learning library of Python ver.3.8. The heart sound data is stored in the working computer, which was used during analysis by the KNN module during runtime. In this experimentation, the optimized number of nearest neighbors was considered as $K = 3$. Further KNN classifier is defined for the model's training Process for predicting the type of valvular heart disease.

Algorithm Used in the Heart Sound analysis

1 Consider K number of neighbors in training dataset of heart sound samples.
2 Compute the Euclidean distance of K number of neighbors.
3 Compute the K nearest neighbors as per the computed Euclidean distance.
4 Among these k neighbors, find the number of the data points in each category.
5 Assign the new data points to that category for which the number of the neighbor is maximum.
6 **Training Time Complexity** = O(k*n*d)

Where,
n= number of training examples,
d=number of dimensions of the data,
k= number of neighbours

Fig. 10 shows the accuracy of the proposed KNN model under the training and validation phase. The plot of accuracy with training set size indicates that accuracy increases as the training set size increases in the training and validation phase for the proposed KNN model. Cross-validation has been done by choosing the value of the number of folds as 10.

This work has considered one thousand PCG recordings of five different subsets (N, AS, MR, MS, and MVP). Each subset has 200 heart

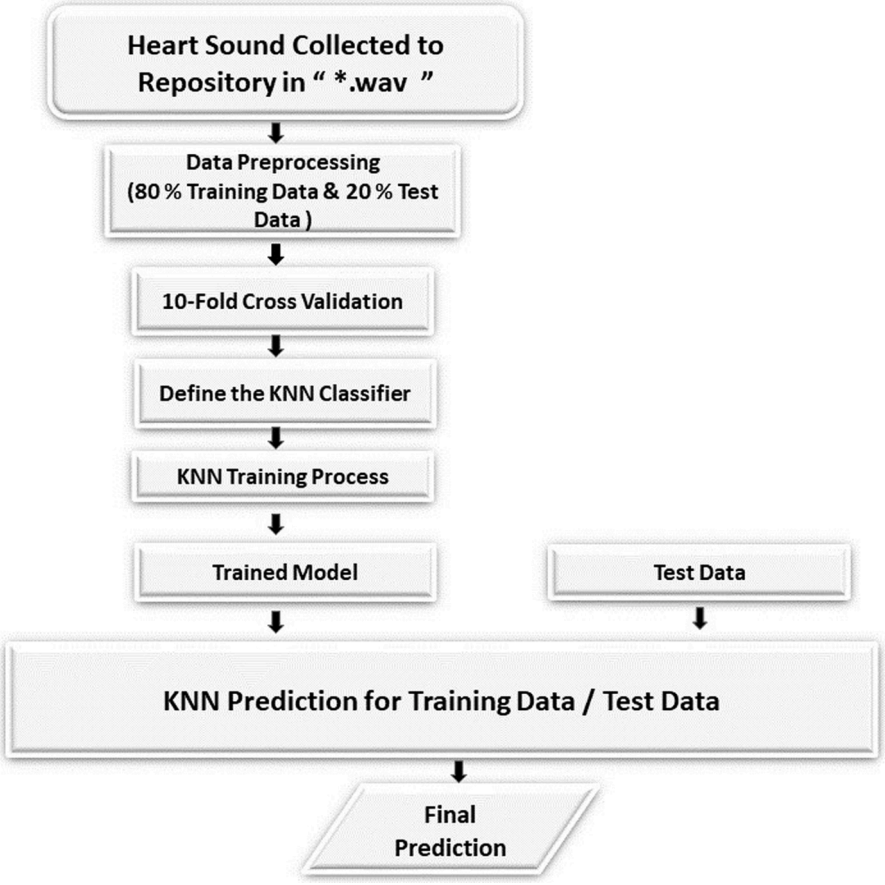


Fig. 9. Flowchart of K-Nearest Neighborhood.

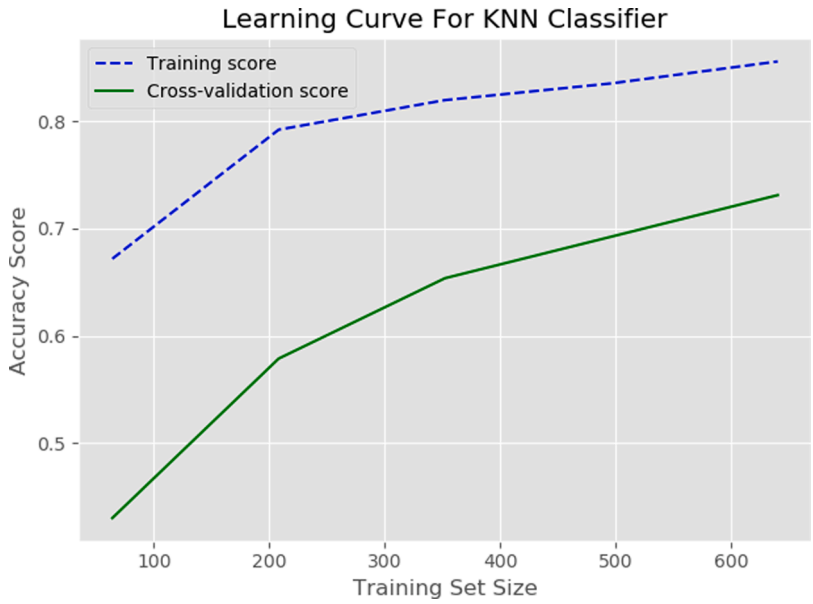


Fig. 10. Plot of Accuracy vs. Training Set Size in Training and validation in KNN.

sound samples. Eighty percent of the data of each subset has been used to train the model. The rest, 20% of the data of each subset, are used to test the model for validation. Fig. 11.(a) is the visualization of the KNN normalized classification plot, has been obtained after the training of the model with 80% dataFig. 11.(b) is the visualization of the KNN normalized classification plot, which was obtained to validate the model with 20% of the rest data of each subset of heart samples. Comparing the two results of Fig. 11(a) and Fig. 11(b), it is observed that the model is able to classify the heart sounds properly.

The performance of the KNN model has been computed, and different metrics have been considered for the performance evaluation of the KNN model in Table 5.

3.3.2. *Support vector machine algorithm*

Like the KNN model, the Support vector model (SVM) has been written in Python ver. 3.8 and the same heart sound data has been fed to the model for the evaluation of the SVM. The flowchart of the SVM algorithm developed is shown in Fig. 12. Different optimized hyperparameters like C = 1, Gamma = 0.1, and Kernel 'Radial Basis Function' have been used in this study.

Algorithm Used in the Heart Sound analysis

1 Initially, load the important libraries for SVM.
2 Divide the imported dataset into training data and test data.
3 Define the SVM classifier.
4 Fit the SVM classifier model.
5 Finally predicts the output result.
6 **Training Time Complexity** = $O(n^2)$

Where,
n= number of training examples.
d=number of dimensions of the data.
k= number of support vectors.
Run Time Complexity = O(k*d)

Fig. 13 shows the accuracy of the proposed SVM model under the training and validation phase. The plot of accuracy with training set size indicates that accuracy increases as the training set size increases in the training and validation phase for the proposed SVM model. This SVM algorithm has been taken from the Python library also. The data preprocessing is applied by considering 80% training data & 20% test data. Cross-validation has been done by choosing the value of $k = 10$.

Fig. 14.(a) is the visualization of the SVM normalized classification plot, obtained after the model's training with 80% data. Fig. 14.(b) is the SVM normalized classification plot visualization, which was obtained to validate the model with 20% of the rest data of each subset of heart samples. Comparing the two results of Fig. 14.(a) and Fig. 14.(b), the model can classify the heart sounds [21, 28] properly.

The performance of the SVM model is summarized in Table 6. It indicates that the SVM model can be deployed in the valvular heart disease analysis system. The accuracy of the SVM Classifier is 0.99, obtained with high precision, recall, and F1 Score.

3.3.3. *Random forest algorithm*

Random Forest is well known supervised machine learning algorithm used primarily for classification-related applications [22, 25]. Like others, the Random Forest algorithm has been used to test its performance to classify normal and abnormal heart sounds.

Algorithm Used in the Heart Sound analysis

1 Initially, random samples have been selected from the given training dataset of a standard heart sound bank.
2 Data Pre-processing has been applied to the standard heart sound bank.
3 After that, this algorithm will frame out a decision tree for every selected random sample. Further, it will receive the prediction outcome from every created decision tree.
4 In this step, voting will be done for every predicted result.
5 Finally, the most voted prediction result is selected as the final predicted output.
6 **Training Time Complexity** = O(n*log(n)*d*k) Where,

k=number of Decision Trees
n=number of data's in the training set
d=number of features in the data
Run Time Complexity = O(depth of tree* k)

It is the fastest among all other machine learning classifiers in terms of run- time complexity.

In Fig. 15, a heart sound bank has been decomposed into the training and test sets. Training set again sub decomposed into 800 datasets termed as estimators for making decisions. The final prediction for valvular heart disease is made through the average of all the decisions made through estimators.

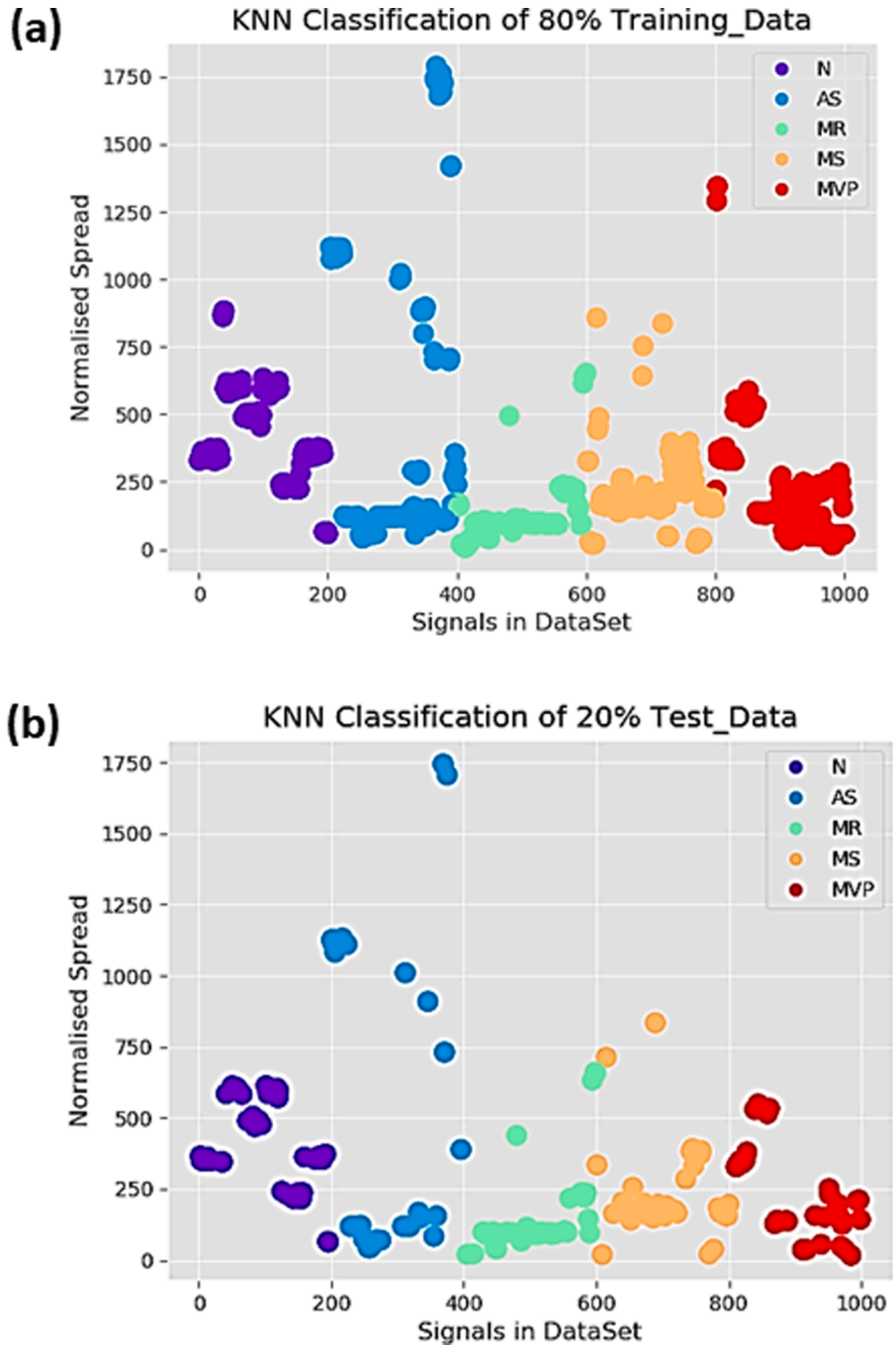


Fig. 11. (a): KNN Classification of 80% training data
(b): KNN Classification of 20% Test data.

Table 5
Summary of KNN Model performance.

Model	Accuracy	Precision	Recall	F1-Score
KNN	0.76	0.75	0.77	0.74

Fig. 16 shows the accuracy of the proposed random forest model under the training and validation phase. The plot of accuracy with training set size indicates that accuracy of training score and validation score converges close to 1 as the training set size increases in the training and validation phase for the proposed RF model.

Fig. 17.(a) is the visualization of the Random Forest normalized classification plot, which has been obtained after the training of the model with 80% data

Fig. 17.(b) is the Random Forest normalized classification plot visualization, which was obtained to validate the model with 20% of the rest data of each subset of heart samples.

The performance of the Random Forest model has been computed, and different metrics have been considered for the performance evaluation of the Random Forest model in Table 7. Comparing the two results of Fig. 17(a) and Fig. 17(b), the model can classify the heart sounds more efficiently with higher accuracy than the SVM classifier.

3.3.4. *Naïve bayes algorithm*

Naive Bayes [23, 31] is a probability-based method used for producing a classification. Naive Bayes classifiers have good results in

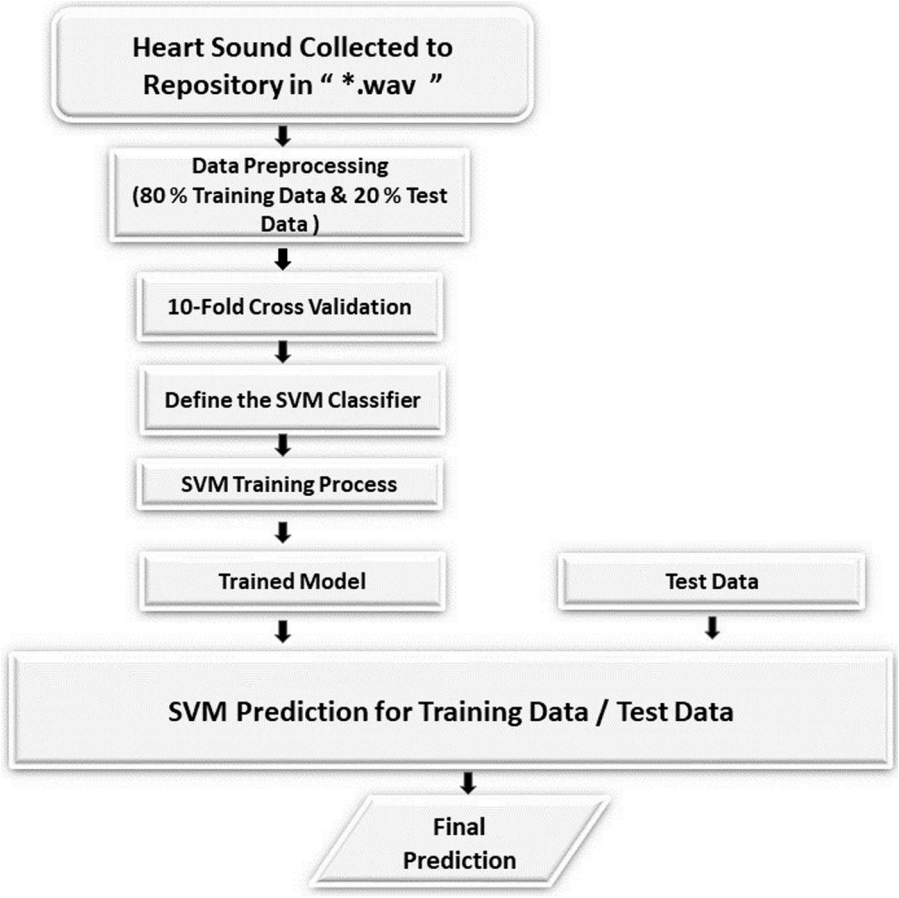


Fig. 12. Flowchart of Support Vector Machine.

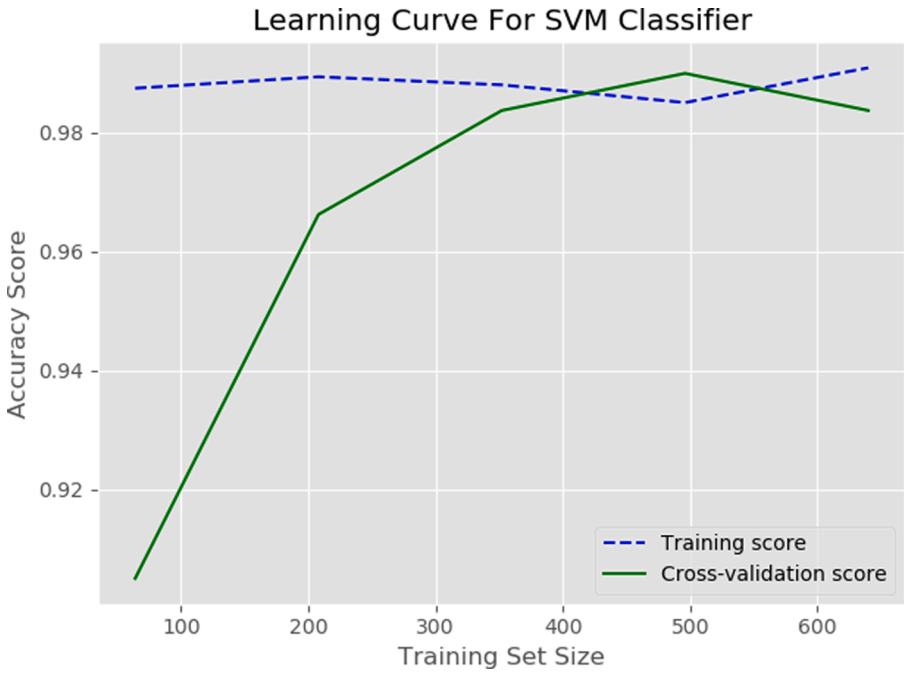
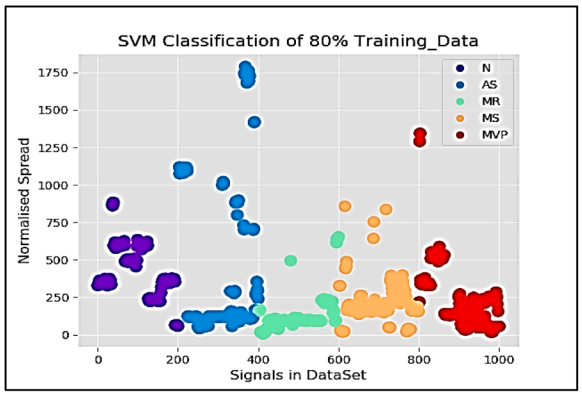
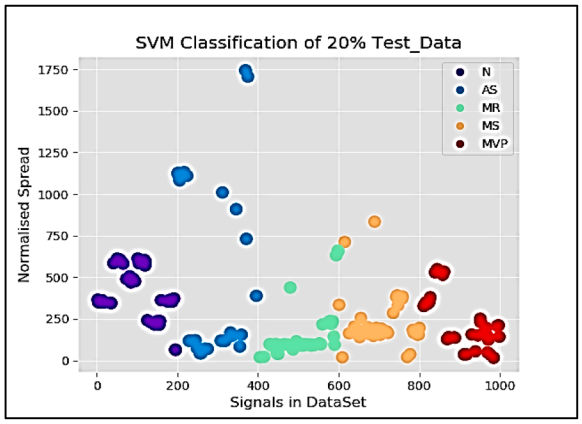


Fig. 13. Plot of Accuracy vs. Epoch in Training and validation in SVM.

complex real-world problems. An advantage of Naive Bayes is that it only requires a small set of training data to estimate the parameters needed for classification. The same data architecture and standard Python Library-based Naïve Bayes classifier have been deployed to classify normal and abnormal heart sounds. Fig. 18 is the flowchart of the Naïve Bayes Algorithm used.



(a) SVM Classification of 80% Training data



(b) SVM Classification of 20% Test Data

Fig. 14. (a): SVM Classification of 80% Training data
(b): SVM Classification of 20% Test Data.

Table 6
Summary of SVM Model performance.

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.99	0.98	0.97	0.98

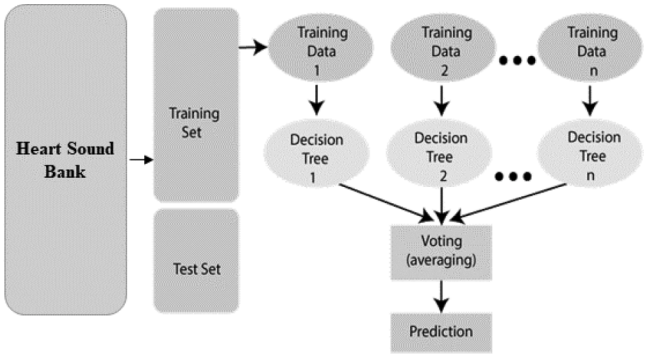


Fig. 15. Flowchart of Random Forest.

Algorithm Used in the Heart Sound analysis

1 Initially, samples have been selected from the given training dataset of a standard heart sound bank.
2 Data Pre-processing has been applied on the standard heart sound bank.

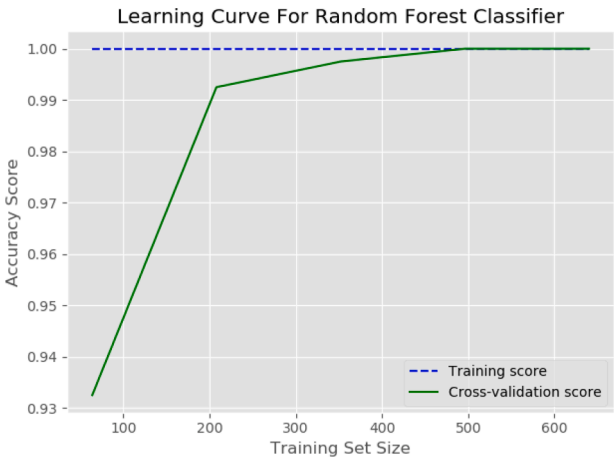
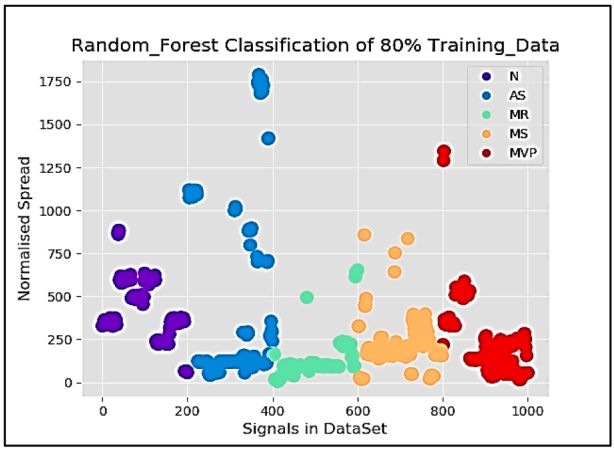
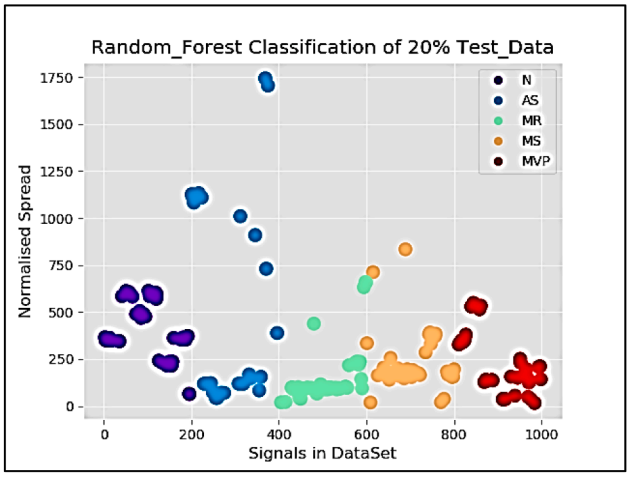


Fig. 16. Plot of Accuracy vs. Epoch in Training and Validation in Random Forest.



(a) Random Forest Classification of 80% training data



(b) Random Forest Classification of 20% Test data

Fig. 17. (a): Random Forest Classification of 80% training data
(b): Random Forest Classification of 20% Test data.

Table 7
Summary of Random Forest Model performance.

Model	Accuracy	Precision	Recall	F1-Score
RF	0.99	0.98	0.97	0.99

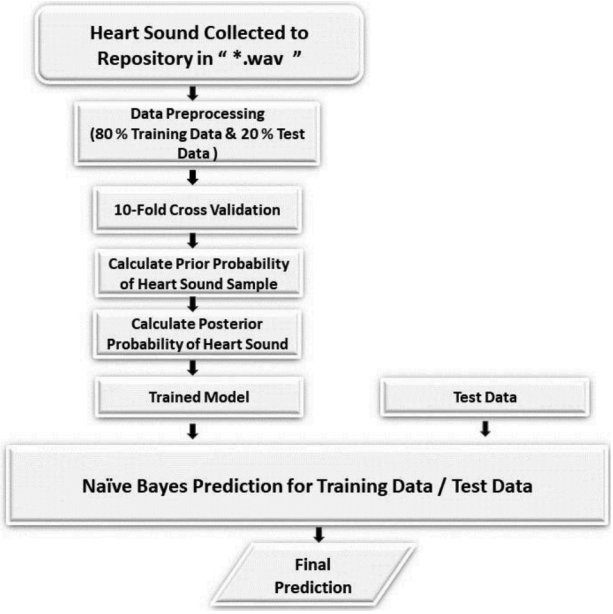


Fig 18. Flowchart of Naïve Bayes.

3 Calculate the predictor prior probability, class prior probability and probability of predictor given class.
4 Calculate the posterior probability.

$$P(c|x) = \frac{P(x|c\ )\ P(C)}{P(x)}$$

Where,
$P(c|x)$ = Posterior probability.
$P(x)$ = Predictor prior probability.
$P(C)$ = Class prior probability.
$P(x|c\ )$ = Probability of predictor given class.

1 **Training Time Complexity** = O(n*d)

Run Time Complexity = O(c*d)
Where,
c=number of classes
n=number of data's in the training set
d=number of features in the data

Fig. 19 shows the accuracy of the proposed NB model under the training and validation phase. The plot of accuracy with training set size indicates that the accuracy of training score and validation score converges as the training set size increases in the training and validation phase for the proposed NB model.

Fig. 20.(a) is the visualization of the Naïve Bayes normalized classification plot, has been obtained after the training of the model with 80% data.

Fig. 20.(b) is the Naïve Bayes normalized classification plot visualization, which was obtained to validate the model with 20% of the rest data of each subset of heart samples.

The performance of the Naïve Bayes model is summarized in Table 8. It indicates that the Naïve Bayes model can also be deployed in the valvular heart disease analysis system.

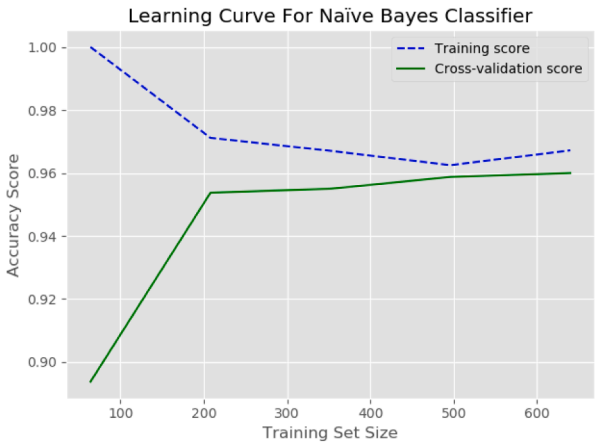
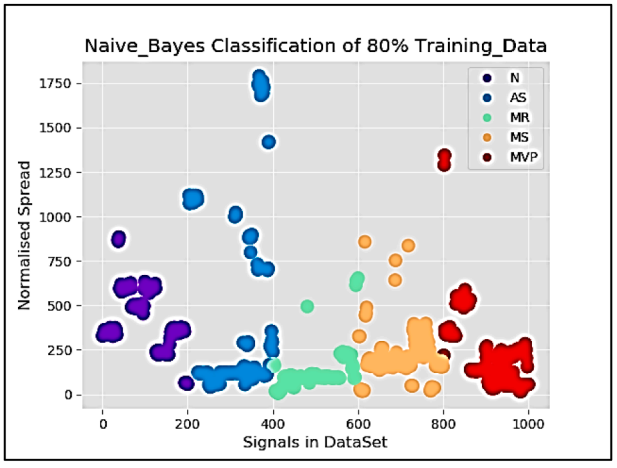
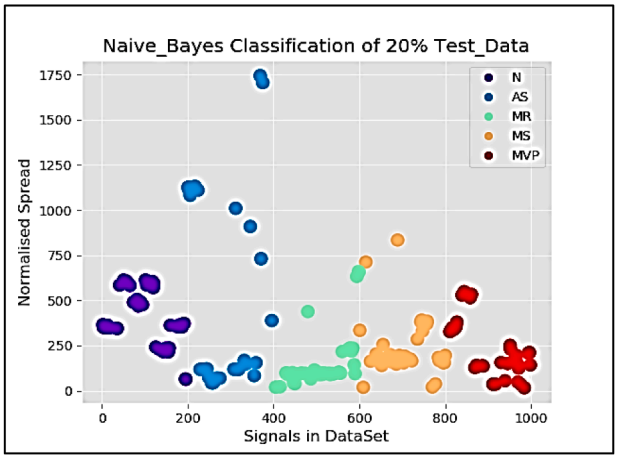


Fig. 19. Plot of Accuracy vs. Epoch in Training and validation in Naïve Bayes.



(a) Naive Bayes Classification of 80% training data



(b) Naïve Bayes Classification of 20% test data

Fig. 20. (a): Naive Bayes Classification of 80% training data
(b): Naïve Bayes Classification of 20% test data.

Table 8
Summary of Naïve Bayes Model performance.

Model	Accuracy	Precision	Recall	F1-Score
NB	0.96	0.95	0.94	0.95

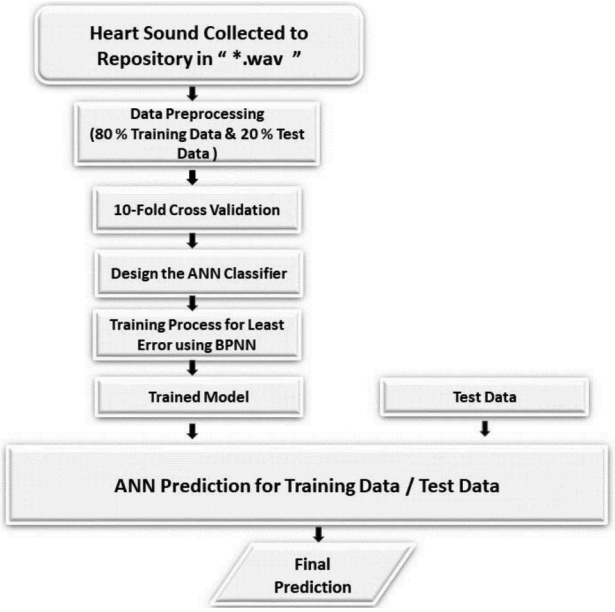


Fig. 21. Flowchart of the Artificial Neural Network.

3.3.5. *Artificial neural network (ANN)*

ANN algorithm used for heart sound analysis is given in Fig. 21. From the general model of an artificial neural network [24, 27], the final output can be computed by applying the activation function over the obtained net input. Heart sound is collected from a heart sound bank with the same architecture used in the entire experimentation described. Then data preprocessing is used by considering 80% training data & 20% test data. ANN classifier is defined for the model's training Process for predicting the type of valvular heart disease. The training process is done using Back Propagation Neural Network (BPNN) until the least error is not obtained.

Algorithm Used in the Heart Sound analysis

1. Initially, load the important libraries for ANN.
2. Divide the imported dataset into training data and test data.
3. Define the ANN classifier.
4. Fit the ANN classifier model.
5. Finally, it predicts the output result.
6. **Training Time Complexity** = O(n*m*d)

Where,
n= input dimension.

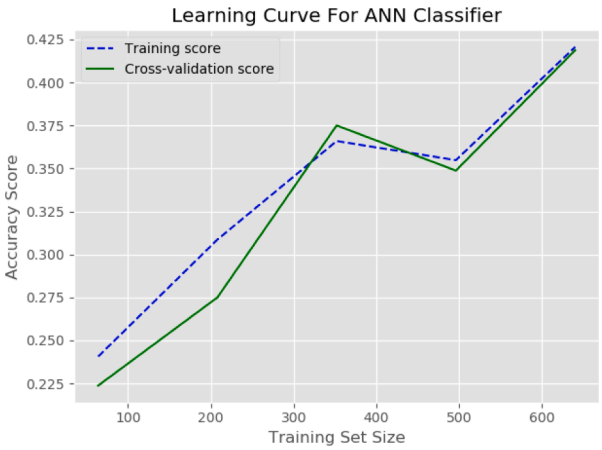


Fig. 22. Plot of Accuracy vs. Epoch in Training and validation in ANN.

d= batch size.
m= output dimension.
Run Time Complexity = O(n*m*d)

Fig. 22 shows the accuracy of the proposed ANN model under the training and validation phase. The plot of accuracy with training set size indicates that accuracy increases as the training set size increases in the training and validation phase for the proposed ANN model. The used ANN model comprises an input layer, one hidden layer, and an output layer that runs for 100 epochs with training data during the training phase of the model. Cross-validation has been done by choosing the value of $k = 10$.

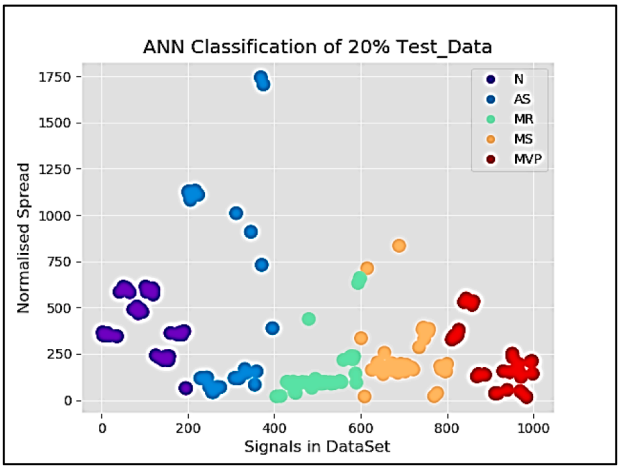
Fig. 23.(a) is the visualization of the ANN classification plot. It has been obtained after the training of the model with 80% data

Fig. 23.(b) is the ANN normalized classification plot visualization, which was obtained to validate the model with 20% of the rest data of each subset of heart samples.

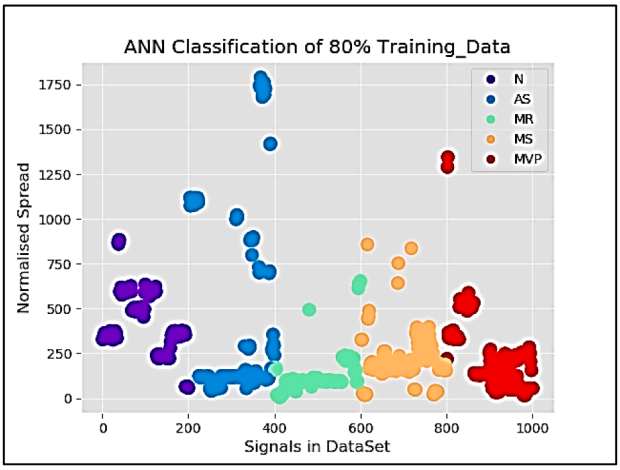
The performance of the ANN model has been computed, and different metrics have been considered for the performance evaluation of the ANN model in Table 9.

4. Results and discussions

As discussed in this research work, a comparison of all different classifiers in terms of Accuracy, Precision, Recall, and FI-Score has been made and given in Table 10.



(a) ANN Classification of 20% test data



(b) ANN Classification of 80% training data

Fig. 23. (a): ANN Classification of 20% test data
(b): ANN Classification of 80% training data.

Table 9
Summary of ANN Model performance.

Model	Accuracy	Precision	Recall	F1-Score
ANN	0.65	0.56	0.62	0.59

Table 10
COMPARISON OF DIFFERENT CLASSIFIERS.

Classifiers	Accuracy	Precision	Recall	F1-Score
SVM	0.99	0.98	0.97	0.98
KNN	0.76	0.75	0.77	0.74
Random Forest	0.99	0.98	0.97	0.99
Naïve Bayes	0.96	0.95	0.94	0.95
ANN	0.65	0.56	0.62	0.59

The table shows that the metrics obtained by Random Forest Algorithm are the highest compared to the other classifiers studied. However, the Random Forest algorithm is a supervised machine learning mechanism with limitations in real-time application and large data. Deep learning models like LeNet-5 scored a classification accuracy of 68.98%, AlexNet achieved an accuracy of 72.35%, VGG16 obtained an accuracy of 74.09%, VGG19 attained an accuracy of 82.18, DenseNet121 obtained 92.48%, Inception Net achieved 96.96%, and Residual Net attained 97.32%. The proposed modified Xception net attained the highest accuracy of 99.43%. Therefore it is suitable for designing AI-based equipment for the early screening of valvular heart diseases. The novelty of the research work lies in the application of xception network model in valvular heart sound analysis, where training and testing time takes very less amount of time. However, earlier, it found quite a few applications in research works that are related to image analysis.

5. Conclusions

The purpose of the experimental studies was to find suitable Classifiers using Python-based Deep Learning Neural Networks and a few selected Machine Learning Algorithms using verified normal and abnormal heart sounds of five classes as described in the text. Classifier Accuracy, Precision, Recall, and F1-Scores are evaluated with the same heart sound database in this entire experiment. It has been observed that the Random Forest Classifier can be chosen as the best Machine Learning Classifier though it has some complexity when the size of training data becomes very large. CNN-based **modified deep learning xception network** is most suitable as an AI classifier of Heart sound though it takes much time in the learning phase compared to machine learning. However, once learned, it gives a speedy result. In the subsequent research, the objective will be to use CNN based deep learning model and random forest classifier to develop a low-cost IoT-enabled Valvular Heart disease screening system for rural use.

Declaration of Competing Interest

The authors declare no conflict of interest.

References

[1] Cota Navin Gupta, Ramaswamy Palaniappan, Sreeraman Rajan, Sundaram Swaminathan, S.M. Krishnan, Segmentation and Classification of heart sounds, in: International Conference: Canadian Conference on Electrical and Computer Engineering, IEEE Xplore, June 2005, https://doi.org/10.1109/CCECE.2005.1557305.

[2] Talha J. Ahmad, Hussnain Ali, Shoab A. Khan, Classification of phonocardiogram using an adaptive fuzzy inference system, in: Conference: Proceedings of the 2009 International Conference on Image Processing, Las Vegas, Nevada, USA, Computer Vision, & Pattern Recognition, IPCV, 2009. July 13-16, 2009Vol-II.

[3] Mandeep Singh, Amandeep Cheema, Heart sounds classification using feature extraction of phonocardiography signal, Int. J. Comput. Appl. Volume 77 (4) (September 2013). ISSN NO:0975 –8887.

[4] Ajay Kumar Roy, Abhishek Misal, G.R. Sinha, Classification of PCG signals: a survey, Int. J. Comp. Appl. Recent Adv. Inf. Technol. (2014). ISSN No: 0975-8887.

[5] S. Barma, B.-.W. Chen, W. Ji, F. Jiang, J.-.F. Wang, Measurement of duration, energy of instantaneous-frequencies, and splits of subcomponents of the second heart sound, IEEE Trans. Instrum. Meas. 64 (7) (Jul. 2015) 1958–1967.

[6] Siddique Latif, Muhammad Usman, Rajib Rana, Junaid Qadir, Phonocardiographic sensing using deep learning for abnormal heartbeat detection, Sensors J. IEEE 18 (22) (2018) 9393–9400.

[7] Dr.Naveen Kumar Dewangan, Dr.S.P. Shukla, Mrs.Kiran Dewangan, PCG signal analysis using discrete wavelet transform, Int. J. Adv. Manag. Technol. Eng. Sci. 8 (Issue III) (2018). MARCH/ISSN NO: 2249-7455.

[8] Gyanaprava Mishra, Kumar Biswal, Asit Kumar Mishra, denoising of heart sound signal using wavelet transform, Int. J. Res. Eng. Technol. 2 (04) (2013). Apr-ISSN: 2319-1163.

[9] Simarjot Kaur Randhawa, Mandeep Singh, Classification of heart sound signals using multimodal features, in: Second International Symposium on Computer Vision and the internet 58, Elsevier, Procedeia Computer science, 2015, pp. 165–171.

[10] Hong Tang, Ziyin Dai, Yuanlin Jiang, Ting Li, Chengyu Liu, PCG Classification using multidomain features and SVM classifier, Hindawi BioMed Res. Int. Vol. (2018), https://doi.org/10.1155/2018/4205027. Article ID 4205027.

[11] Fan Li, Hong Tang, Shang Shang, Klaus Mathiak, Fengyu Cong, Classification of heart sounds using convolutional neural network, Appl. Sci. 10 (2020) 3956, https://doi.org/10.3390/app10113956.

[12] Matin Z. Othman, Asmaa N Khaleel, Phono cardiogram signal analysis for murmur diagnosing using shannon energy envelop and sequenced DWT decomposition, J. Eng. Sci. Technol. 12 (9) (2017). ISSN:2393 –2402.

[13] G. Venkata Hari Prasad, Dr.P.Rajesh Kumar, Analysis of various DWT methods for feature extracted PCG signals, Int. J. Eng. Res. Technol. (IJERT) 4 (04) (2015). April-ISSN: 2278-0181.

[14] Yaseen, Gui-Young Son, Soonil Kwon, Classification of heart sound signal using multiple features, Appl. Sci. 8 (2018) 2344, https://doi.org/10.3390/app8122344.

[15] R. Amarnath, Methods for Classification of Phonocardiogram. TENCON 2003, in: Conference on Convergent Technologies for the Asia-Pacific Region 4, 2003, pp. 1514–1515.

[16] Your first deep learning project in python with Keras Step.... https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/.

[17] Y. Xu, B. Xiao, X. Bi, W. Li, J. Zhang, X. Ma, Pay more attention with fewer parameters: a novel 1-D convolutional neural network for heart sounds classification, in: Proceedings of the Computing in Cardiology Conference (CinC), Maastricht, The Netherlands 45, 2018, pp. 1–4, 23–26 September.

[18] Joyanta Kumar Roy, Tanmay Sinha Roy, A Simple technique for heart sound detection and real-time analysis, in: Proceedings of ICST 2017 held at Macquarie University Sidney, Sensing Technology (ICST), 2017 Eleventh International Conference, 2017, https://doi.org/10.1109/ICSensT.2017.8304502, 4-6 Dec.

[19] M. El-Segaier, O. Lilja, S. Lukkarinen, L. Sörnmo, R. Seppanen, E. Pesonen, Computer-based detection and analysis of heart sound and murmur, Ann. Biomed. Eng. 33 (7) (2005 Jul) 937–942. http://www.ncbi.nlm.nih.gov/pubmed/16060534.

[20] H. Nygaard, et al., Assessing the severity of aortic valve stenosis by spectral analysis of cardiac murmurs (spectral vibrocardiography). Part I: technical aspects, J. Heart Valve Dis. 2 (4) (1993) 454–467.

[21] N. Baghel, M.K. Dutta, R. Burget, Automatic diagnosis of multiple cardiac diseases from PCG signals using convolutional neural network, Comput. Methods Programs Biomed. 197 (2020), 105750.

[22] J.P. Dominguez-Morales, A.F. Jimenez-Fernandez, M.J. Dominguez-Morales, G. Jimenez-Moreno, Deep Neural Networks for the Recognition and Classification of Heart Murmurs Using Neuromorphic Auditory Sensors, IEEE Trans. Biomed. Circuits Syst. 12 (2018) 24–34.

[23] A.I. Humayun, S. Ghaffarzadegan, I. Ansari, Z. Feng, T. Hasan, Towards domain invariant heart sound abnormality detection using learnable Filterbanks, IEEE J. Biomed. Health Inform. 24 (2020) 2189–2198.

[24] Anju, Sanjay Kumar, Detection of cardiac murmur, Anjuet al, Int. J. Comput. Sci. Mobile Comput. 3 (7) (2014) 81–87. July-ISSN 2320–088X.

[25] Joyanta Kumar Roy, Tanmay Sinha Roy, Subhas Chandra Mukhopadhyay, Heart sound: detection and analytical approach towards diseases, Modern Sens. Technol. (2019) 103–145, https://doi.org/10.1007/978-3-319-99540-3_7. Edited by Subhas Chandra Mukhopadhyay, Published by Springer Nature Switzerland AG.

[26] Joyanta Kumar Roy, Tanmay Sinha Roy, Nirupama Mandal, Octavian Adrian Postolache, A Simple technique for heart sound detection and identification using Kalman filter in real-time analysis, in: Proceedings of ISSI 2018 held at Shanghai, China, International Symposium Sensing And Instrumentation IOT Era (ISSI), 2018 First International Conference, IEEE, 2018, 6-7 Sept.978-1-5386-5638-9/18/$31.00 ©2018.

[27] F. Noman, C.-.M. Ting, S.-.H. Salleh, H. Ombao, Short-segment heart sound classification Using an ensemble of deep convolutional neural networks, in: Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 1318–1322, 12–17 May.

[28] S.L. Oh, V. Jahmunah, C.P. Ooi, R.-.S. Tan, E.J. Ciaccio, T. Yamakawa, M. Tanabe, M. Kobayashi, U.R. Acharya, Classification of heart sound signals using a novel deep WaveNet model, Comput. Methods Programs Biomed 196 (2020), 105604.

[29] A. Cheema, M. Singh, Steps involved in heart sound analysis- a review of existing trends, Int. J. Eng. Trends. Technol. 4 (7) (2013) 2921–2925.
[30] J.B. Wu, S. Zhou, Z. Wu, X.M. Wu, Research on the method of characteristic extraction and Classification of Phonocardiogram, in: Systems and Informatics (ICSAI), 2012 International Conference on, 2012, pp. 1732–1735.
[31] Lubaib. P., Ahammed Muneer K.V., "The Heart Defect Analysis Based on PCG Signals using Pattern Recognition Techniques", Elsevier, International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST- 2015).
[32] B. Xiao, Y. Xu, X. Bi, W. Li, Z. Ma, J. Zhang, X. Ma, Follow the sound of Children's heart: a deep-learning-based computer-aided pediatric CHDs diagnosis system, IEEE Internet Things J 7 (2020) 1994–2004.
[33] F.A. Khan, A. Abid, M.S. Khan, Automatic heart sound classification from segmented/unsegmented phonocardiogram signals using time and frequency features, Physiol. Meas. 41 (2020), 055006.
[34] F. Li, M. Liu, Y. Zhao, L. Kong, L. Dong, X. Liu, M. Hui, Feature extraction and classification of heart sound using 1D convolutional neural networks, EURASIP J. Adv. Signal Process. 2019 (2019) 1–11.
[35] J.M.-T. Wu, M.-.H. Tsai, Y.Z. Huang, S.H. Islam, M.M. Hassan, A. Alelaiwi, G. Fortino, Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model, Appl. Soft Comput. 78 (2019) 29–40.
[36] T.-.C. Yang, H. Hsieh, Classification of acoustic physiological signals based on deep learning neural networks with augmented features, in: Proceedings of the 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 569–572, 11–14 September.
[37] E.F. Gomes, P.J. Bentley, M. Coimbra, E. Pereira, Y. Deng, Classifying heart sounds: approaches to the PASCAL challenge, in: Proceedings of the HEALTHINF 2013-Proceedings of the International Conference on Health Informatics, Barcelona, Spain, 2013, pp. 337–340, 11–14 February.
[38] C. Liu, D. Springer, Q. Li, B. Moody, R.A. Juan, F.J. Chorro, F. Castells, J.M. Roig, I. Silva, A.E. Johnson, Z. Syed, S.E. Schmidt, C.D. Papadaniil, L. Hadjileontiadis, H. Naseri, A. Moukadem, A. Dieterlen, C. Brandt, H. Tang, M. Samieinasab, M. R. Samieinasab, R. Sameni, R.G. Mark, G.D. Clifford, An open-access database for the evaluation of heart sound algorithms, Physiol. Meas. 37 (12) (2016 Dec) 2181–2213.
[39] S. Rubin Bose, Sathiesh Kumar, In-situ recognition of hand Gesture via enhanced Xception based single-stage deep convolutional neural network, Elsevier, Expert Syste. Appl. 193 (2021), 116427, https://doi.org/10.1016/j.eswa.2021.116427.
[40] Md. Zabirul Islam, Md. Milon Islam, Amanullah Asraf. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Elsevier, Informatics in Medicine Unlocked. 10.1016/j.imu.2020.100412.

Tanmay Sinha Roy was born in West Bengal, India, in 1988. He received his B.Tech. Degree in Instrumentation and Control engineering from West Bengal University of Technology, West Bengal, India, in 2009, and the M.Tech. Degree in Applied Electronics and Instrumentation Engineering from West Bengal University of Technology, India, in 2011. He is currently an Assistant Professor in the Electrical Engineering Department, Haldia Institute of Technology, West Bengal University of Technology, and pursuing his Ph.D. from IIT(ISM), Dhanbad.

His-research interests include PCG signal analysis, development of systems for heart sound acquisition, instrumentation, and control, design of low-cost acoustic stethoscopes for diseased patients.

Joyanta Kumar Roy received a Ph.D. degree from Calcutta University, West Bengal. He has been an Electronics and Automation Engineer for the last 40 years as a Company Director, Consulting, Engineering, Developer, Researcher, and Educationist. He is currently the Chairman of Eureka Scientech Research Foundation (ESRF), working as a Visiting Professor with the Narula Institute of Technology, a Company Director with System Advance Technologies, and a Freelance Consultant with a number of industries to give design support toward smart technology in the water sector. His-research interests include the development of smart measurement and control systems, multifunction sensors, IoT-based health and technology-assisted living, and smart homes and cities. He is a Senior Member of IET, a Fellow Member of IETE and IWWA, and a regular reviewer of IEEE and Springer Journals. He is an Associate Editor of S2IS journal. He published a significant number of scientific and technical publications in the form of books, book chapters, design documents, and research papers.

Nirupama Mandal received a Ph.D. degree from Calcutta University, West Bengal, in 2012. She is currently an Associate Professor with the Department of Electronics Engineering, Indian Institute of Technology (ISM), Dhanbad. She was a HOD with the EIE department, Asansol Engineering College, West Bengal, in 2013. She Received a National Scholarship award from the Government of India in 2001 and was then elevated to IEEE senior member in February 2019. Her research interests include Transducer Development, Controller Design, Process Plant Instrumentation, Process Modeling, Smart sensing system, and Smart instrumentation.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 177 (2020) 432–437

Procedia Computer Science

www.elsevier.com/locate/procedia

The 6th International Workshop on Ambient Assisted Technologies for HealthCare (AATCARE 2020)
November 2-5, 2020, Madeira, Portugal

Machine learning for the evaluation of the presence of heart disease

Ivan Miguel Pires[a,b,*], Gonçalo Marques[a], Nuno M. Garcia[a] and Vasco Ponciano[c,d]

[a] *Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã, Portugal*
[b] *Department of Computer Science, Polytechnic Institute of Viseu, Viseu, Portugal*
[c] *R&D Unit in Digital Services, Applications and Content, Polytechnic Institute of Castelo Branco, Castelo Branco, Portugal*
[d] *Altranportugal, Lisbon, Portugal*

Abstract

Currently, heart diseases are prevalent in the population. Machine learning methods may help in the identification of heart diseases in the different people with the analysis of various features of heart rate, such as PPE, spread, spread2, MDVP:Fo(Hz), MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA, DFA, RPDE, D2, MDVP:Fhi(Hz), MDVP:Flo(Hz), NHR, HNR, MDVP:Jitter(Abs), MDVP:Jitter(%), MDVP:RAP, MDVP:PPQ, and Jitter:DDP. The analysis was performed with the dataset from the UCI Machine learning repository from the Center for Machine Learning and Intelligent Systems. This paper proposes the use of different methods, such as Neural Network, Decision Tree, k-Nearest Neighbor (kNN), Combined nomenclature (CN2) rule inducer, Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD). The best results on the 20-fold Cross-validation and the 10-fold Cross-validation are reported by DT and SVM methods (87.69%). Also, the best results on the 5-fold Cross-validation are reported by SGD (87.69%).

Keywords: heart diseases; feature extraction; artificial intelligence.

* Corresponding author. Tel.: +351-966-379-785.
E-mail address: impires@it.ubi.pt

10.1016/j.procs.2020.10.058

1. Introduction

Every year, 3.9 million deaths are related to the presence of heart diseases in Europe, which is related to 45% of all deaths in Europe [1]. However, 85-millions of persons in Europe are living with heart diseases [2]. There are several heart diseases over the world, such as Congenital heart disease, Arrhythmia, Coronary artery disease, Dilated cardiomyopathy, Myocardial infarction, Heart failure, Hypertrophic cardiomyopathy, Mitral regurgitation, Mitral valve prolapse, and Pulmonary stenosis [3]. Different factors may influence the presence of heart diseases over the world, such as genetics, age, gender, tobacco, physical inactivity, diet, celiac disease, sleep, socioeconomic disadvantage, air pollution, cardiovascular risk assessment, depression and traumatic stress, occupational exposure, and somatic mutations [4]. The difficulty in accessing health services is a great challenge to detect and treat this type of disease. There are several reports and cases of people who seem to be healthy and end up dying due to heart problems [5]. Even highly competitive athletes and healthy young people end up suffering heart attacks [6]. The acceleration of day-to-day the possibility of the existence of an automatic method of detecting this type of diseases based on machine learning can be a relevant contribution to avoid these types of events.

This paper's motivation is related to the high prevalence of cardiovascular diseases over the world, which is one of the causes of mortality in the world [7]. The automatic identification of these diseases with electrocardiography sensors is relevant to prevent the evolution of these problems. The use of artificial intelligence methods automates its recognition. It is one subject related to the development of Ambient Assisted Living solutions [8]–[12]. This paper proposed the use of artificial intelligence methods, including Neural Network [13], Decision Tree (DT) [14], k-Nearest Neighbor (kNN) [15], Combined nomenclature (CN2) rule inducer, Support Vector Machine (SVM) [5], and Stochastic Gradient Descent (SGD) [17], for the identification of the presence of heart diseases in the dataset available at Wisconsin Breast Cancer data from the UCI Machine Learning Depository [18]. As a result of this paper, it is verified that the DT and SVM reported the best results (87.69%) in recognition of the presence of heart diseases with 20-fold Cross-validation and 100-fold Cross-validation. The same results were reported by SGD for 5-fold Cross-validation.

The paper is organized as follows: Section 2 presents the parameters of the methods implemented for the recognition of the presence of heart diseases. The results of the application of the machine learning methods are presented in Section 3. We are finalizing with the discussion of this paper and conclusions in Section 4.

2. Methods

Different methods were implemented to identify the presence of heart disease in the dataset available at [18]. This dataset is composed of 270 observations related to the presence or absence of heart disease. The evaluation of the methods has been conducted using Stratified 20-fold, 10-fold, and 5-fold Cross validations. This dataset includes 22 features, as presented in Table 1. These features were used to implement different methods for automatic identification.

Table 1. Description of the parameters of artificial intelligence methods implemented.

Features:	Description:
PPE; spread; spread2	Nonlinear measures of fundamental frequency variation
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Shimmer; MDVP:Shimmer(dB); Shimmer:APQ3; Shimmer:APQ5; MDVP:APQ; Shimmer:DDA	Measures of variation in amplitude
DFA	Signal fractal scaling exponent
RPDE; D2	Nonlinear dynamical complexity measures
MDVP:Fhi(Hz)	Maximum fundamental frequency
MDVP:Flo(Hz)	Minimum fundamental frequency
NHR; HNR	Measures of the ratio of noise to tonal components in the voice status
MDVP:Jitter(Abs); MDVP:Jitter(%); MDVP:RAP; MDVP:PPQ; Jitter:DDP	Measures of variation in fundamental frequency

Regarding the automatic identification of observations that revealed the presence or absence of health disease, the authors implemented several machine learning methods, such as Neural Network [13], DT [14], kNN [15], CN2 rule inducer, SVM [5], and SGD [17], which the details on the implementation are presented in Table 2.

The dataset is composed of data related to individuals between 29 and 77 years old, which are 183 are men, and 87 are women. The resting blood pressure of the sample is between 94 and 200 bpm. The population reported a serum cholesterol between 126 and 564 mg/dl, where 40 records have fasting blood sugar. It also reported a maximum heart rate between 71 and 202 bpm, where 183 individuals reported angina. The individuals reported an ST depression induced by exercise relative to rest between 0 and 6.2. Also, the slope of the peak exercise ST segment is between 1 and 3. In total, 120 individuals reported the presence of heart disease.

Table 2. Description of the parameters of artificial intelligence methods implemented.

Methods:	**Parameters:**
Neural Network (NN)	Hidden layers: 100
	Activation: Logistic
	Solver: Stochastic Gradient Descent (SGD)
	Alpha: 0.0001
	Maximum iterations: 200
	Replicable training: Yes
Decision Tree (DT)	Pruning: at least two instances in leaves, at least five instances in internal nodes, maximum depth 100
	Splitting: Stop splitting when the majority reaches 95% (classification only)
	Binary trees: Yes
k-Nearest Neighbor (kNN)	Number of neighbors: 5
	Metric: Euclidean
	Weight: Uniform
Combined nomenclature (CN2) rule inducer	Rule ordering: ordered
	Covering algorithm: exclusive
	Gamma: 0.7
	Evaluation measure: entropy
	Beamwidth: 5
	Minimum rule coverage: 1
	Maximum rule length: 5
	Default alpha: 1.0
	Parent alpha: 1.0
Support Vector Machine (SVM)	SVM type: SVM, C=1.0, ε=0.1
	Kernel: RBF, exp(-auto|x-y|^2)
	Numerical tolerance: 0.001
	Iteration limit: 100
Stochastic Gradient Descent (SGD)	Classification loss function: Hinge
	Regression loss function: Squared Loss
	Regularization: Ridge (L2)
	Regularization strength (α): 1e-05
	Learning rate: Constant
	Initial learning rate (η0): 0.01
	Shuffle data after each iteration: Yes

3. Results and Discussion

Three types of validation were implemented to validate the results obtained by the different machine learning methods applied. These are:

- Stratified 20-fold Cross-validation (Section 3.1);
- Stratified 10-fold Cross-validation (Section 3.2);
- Stratified 5-fold Cross-validation (Section 3.3).

3.1. Stratified 20-fold Cross-validation

According to highlighted rows in Table 3, the use of stratified 20-fold Cross-validation reports best accuracy on the classification of the presence of heart disease are reported by DT and SVM methods with 87.69%. At the same point, the recall values are the same 87.69%, but the SVM method reported the best precision (88.79%).

Table 3. Results obtained with stratified 20-fold cross-validation.

Methods	Accuracy	F1-score	Precision	Recall
NN	75.38%	64.80%	56.83%	75.38%
DT	87.69%	87.78%	87.88%	87.69%
kNN	87.18%	86.45%	86.83%	87.18%
CN2 rule inducer	82.56%	82.00%	81.79%	82.56%
SVM	87.69%	86.29%	88.79%	87.69%
SGD	85.64%	84.90%	85.05%	85.64%

3.2. Stratified 10-fold Cross-validation

Following the highlighted results presented in Table 4, the use of stratified 10-fold Cross-validation shows that the best accuracy for the classification of the presence of heart disease is reported by DT and SVM methods with 87.69%. At the same point, the recall values are the same 87.69%. However, the SVM method reported the best precision (88.79%). Table 4 prove that the results of the use of stratified 20-fold Cross-validation suggest the same methods as stratified 10-fold Cross-validation.

Table 4. Results obtained with stratified 10-fold cross-validation.

Methods	Accuracy	F1-score	Precision	Recall
NN	75.38%	64.80%	56.83%	75.38%
DT	87.69%	87.60%	87.53%	87.69%
kNN	85.12%	84.27%	84.47%	85.13%
CN2 rule inducer	81.03%	81.09%	81.16%	81.02%
SVM	87.69%	86.29%	88.79%	87.69%
SGD	87.17%	86.30%	86.96%	87.18%

3.3. Stratified 5-fold Cross-validation

Related to the results highlighted in Table 5, the use of stratified 5-fold Cross-validation shows the results with the best accuracy on the classification of the presence of heart disease are also reported by SGD method with 87.69%. It is the same value of recall, but the precision obtained was 87.67%. This method reported the same accuracy of stratified 20-fold Cross-validation and stratified 10-fold Cross-validation.

Table 5. Results obtained with stratified 5-fold cross-validation.

Methods	Accuracy	F1-score	Precision	Recall
NN	75.38%	64.80%	56.83%	75.38%
DT	84.10%	84.45%	85.03%	84.10%
kNN	82.05%	80.82%	80.91%	82.05%
CN2 rule inducer	82.05%	80.83%	83.10%	82.56%
SVM	87.17%	85.62%	88.37%	87.18%
SGD	87.69%	86.78%	87.67%	87.69%

4. Conclusion

Heart diseases are highly present in the world. Thus, this study proposed the implementation of machine learning methods for the automatic identification of the presence of heart diseases over the world. Based on the dataset used, the implementation of 20-fold Cross-validation and 10-fold Cross-validation with DT and SVM reported the same accuracy of the implementation of 5-fold Cross-validation with SGD, which is 87.69%.

The early detection of health problems may reduce the significant cause of death related to Ambient Assisted Living. Moreover, the existence of an automatic method of detecting heart problems guarantees the development of a low-cost method for the detection of heart disease. Next, the use of this type of sensors with machine learning methods makes more critical the research in this area.

The accuracy values give high strength to this study and increase the speed and capacity of the sensor electrocardiogram. With the use of automatic detection methods, it speeds up the detection and ability to generalize its use by different fringes of society. It also supports health professionals in the early detection of symptoms. With the combination of these types of methods, we can change and help in the development of systems. These systems may be diverse, including a pacemaker with even more capacity, and devices to assist people with identified problems.

Acknowledgements

This work is funded by FCT/MEC through national funds and when applicable co-funded by FEDER-PT2020 partnership agreement under the project **UIDB/EEA/50008/2020**. (*Este trabalho é financiado pela FCT/MEC através de fundos nacionais e cofinanciado pelo FEDER, no âmbito do Acordo de Parceria PT2020 no âmbito do projeto UIDB/EEA/50008/2020*).

This article is based upon work from COST Action IC1303-AAPELE—Architectures, Algorithms and Protocols for Enhanced Living Environments and COST Action CA16226–SHELD-ON—Indoor living space improvement: Smart Habitat for the Elderly, supported by COST (European Cooperation in Science and Technology). COST is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career, and innovation. More information in www.cost.eu.

References

[1] N. Townsend, L. Wilson, P. Bhatnagar, K. Wickramasinghe, M. Rayner, and M. Nichols, “Cardiovascular disease in Europe: epidemiological update 2016,” *Eur Heart J*, vol. 37, no. 42, pp. 3232–3245, Nov. 2016, doi: 10.1093/eurheartj/ehw334.

[2] “CVD Statistics 2017.” http://www.ehnheart.org/cvd-statistics/cvd-statistics-2017.html (accessed May 07, 2020).

[3] Broderick J P, Phillips S J, O’Fallon W M, Frye R L, and Whisnant J P, “Relationship of cardiac disease to stroke occurrence, recurrence, and mortality.,” *Stroke*, vol. 23, no. 9, pp. 1250–1256, Sep. 1992, doi: 10.1161/01.STR.23.9.1250.

[4] A. M. A. Alkinain, “Application of genetic polymorphism & gene experience of RSPONDIN3 as biomarker of cardiometabolic traits associated with or without obesity in sample of Sudanese patients in Khartoum state.,” Thesis, Kamal Eldin Ahmed Abdelsalam, 2019.

[5] N. Ghorayeb, C. S. S. de S. Colombo, R. C. Francisco, and T. G. Garcia, “Sudden Cardiac Death in Sports: Not a Fatality!,” *International Journal of Cardiovascular Sciences*, vol. 32, no. 1, pp. 84–86, 2019.

[6] F. van Buuren and K. P. Mellwig, “Specific Cardiovascular Diseases and Competitive Sports Participation: Valvular Heart Disease,” in *Textbook of Sports and Exercise Cardiology*, Springer, 2020, pp. 291–302.

[7] G. A. Kaplan, J. T. Salonen, R. D. Cohen, R. J. Brand, S. Leonard Syme, and P. Puska, “SOCIAL CONNECTIONS AND MORTALITY FROM ALL CAUSES AND FROM CARDIOVASCULAR DISEASE: PROSPECTIVE EVIDENCE FROM EASTERN FINLAND,” *Am J Epidemiol*, vol. 128, no. 2, pp. 370–380, Aug. 1988, doi: 10.1093/oxfordjournals.aje.a114977.

[8] N. M. Garcia and J. J. P. C. Rodrigues, Eds., *Ambient Assisted Living*, 0 ed. CRC Press, 2015.
[9] V. Felizardo *et al.*, "E-Health: current status and future trends," in *Handbook of Research on Democratic Strategies and Citizen-Centered E-Government Services*, IGI Global, 2015, pp. 302–326.
[10] E. Zdravevski *et al.*, "Improving Activity Recognition Accuracy in Ambient-Assisted Living Systems by Automated Feature Engineering," *IEEE Access*, vol. 5, pp. 5262–5280, 2017, doi: 10.1109/ACCESS.2017.2684913.
[11] G. Marques, R. Pitarma, N. M. Garcia, and N. Pombo, "Internet of Things Architectures, Technologies, Applications, Challenges, and Future Directions for Enhanced Living Environments and Healthcare Systems: A Review," *Electronics*, vol. 8, no. 10, 2019, doi: 10.3390/electronics8101081.
[12] P. S. Sousa, D. Sabugueiro, V. Felizardo, R. Couto, I. Pires, and N. M. Garcia, "mHealth Sensors and Applications for Personal Aid," in *Mobile Health*, vol. 5, S. Adibi, Ed. Cham: Springer International Publishing, 2015, pp. 265–281.
[13] M. H. Hassoun, *Fundamentals of artificial neural networks*. MIT press, 1995.
[14] B. Kamiński, M. Jakubczyk, and P. Szufel, "A framework for sensitivity analysis of decision trees," *Cent Eur J Oper Res*, vol. 26, no. 1, pp. 135–159, Mar. 2018, doi: 10.1007/s10100-017-0479-6.
[15] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992, doi: 10.1080/00031305.1992.10475879.
[16] P. Clark and T. Niblett, "The CN2 induction algorithm," *Mach Learn*, vol. 3, no. 4, pp. 261–283, Mar. 1989, doi: 10.1007/BF00116835.
[17] F. Pfaff, B. Noack, and U. D. Hanebeck, "Data validation in the presence of stochastic and set-membership uncertainties," in *Information Fusion (FUSION), 2013 16th International Conference on*, 2013, pp. 2125–2132.
[18] "UCI Machine Learning Repository: Statlog (Heart) Data Set." https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29 (accessed May 07, 2020).

Machine Learning with Applications 7 (2022) 100206

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Heart sound classification using signal processing and machine learning algorithms

Yasser Zeinali [a], Seyed Taghi Akhavan Niaki [b,*]

[a] Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran
[b] Department of Industrial Engineering, Sharif University of Technology, PO Box 11155-9414 Azadi Ave., Tehran 1458889694, Iran

ARTICLE INFO

Keywords:
Heart sound
Signal processing algorithms
Machine learning algorithms
Dimensional reduction algorithms
Feature selection
Gradient boosting classifier

ABSTRACT

According to global statistics and the world health organization (WHO), about 17.5 million people die each year from cardiovascular disease. In this paper, the heart sounds gathered by a stethoscope are analyzed to diagnose several diseases caused by heart failure. This research's primary process is to identify and classify the data related to the heart sounds categorized in four general groups of S_1 to S_4. The sounds S_1 and S_2 are considered as the heart's normal sounds, and the sounds S_3 and S_4 are the abnormal sounds of the heart (heart murmurs), each expressing a specific type of heart disease. In this regard, the desired features are first extracted after retrieving the data by signal processing algorithms. In the next step, feature selection algorithms are used to select the compelling features to reduce the problem's dimensions and obtain the optimal answer faster. While the existing algorithms in the literature classify the sound into two groups of normal and abnormal, in the final section, some of the most popular classification algorithms are utilized to classify the type of sound into three classes of normal, S_3 and S_4 categories. The proposed methodology obtained an accuracy rate of 87.5% and 95% for multiclass data (3 classes) and 98% for binary classification (normal vs. abnormal) problems.

1. Introduction

This research aims at using artificial intelligence algorithms to diagnose heart failure by classifying heart sounds. According to studies conducted at several large hospitals, heart specialists utilize different medical tests to diagnose heart disease accurately. However, heart sound diagnosis using a stethoscope is very difficult due to the hospitals' noisy environment (Leng et al., 2015). As such, physicians do not hear the heart sound very well and need some relevant tests to analyze patients' conditions.

Each of the two sides of a human heart has two chambers called the ventricles and the atrium, connected by some valves. The heart cycle refers to all the heart events from the beginning of one beat to the beginning of the next beat. This cycle is divided into two parts, which have two modes of contraction and rest called systole and diastole. While normal sounds S_1 and S_2 are received by a healthy heart, any disturbance, including narrowing and widening of the valves, can cause the heart to no longer pump efficiently and generate abnormal sounds S_3 and S_4 (Bonow et al., 2011). The heart's first sound (S_1) comes from the opening and closing of the tricuspid and mitral valves. The second heart sound (S_2) is due to the tracheal valves' opening and closing. The other two sounds are abnormal sounds caused by various reasons, such as congenital cavities between the left and right ventricles, sagging and clogging, or the valves' calcification. An example of the heart sounds signals classified in the above four categories are shown in Fig. 1. This paper aims to classify heart sounds into normal, abnormal type 3, and abnormal type 4.

The problem of classifying heart sounds is directly related to signal processing algorithms. Due to the nature of the raw data collected in audio with a sampling frequency of 2000 kHz, the data are first reviewed and preprocessed in this study. This means that the noise in the sounds is first passed through special filters based on their frequency and amplitude so that the noise can be eliminated and the sound quality can be improved. The second and more important goal is extracting features, which is done by implementing signal processing algorithms. In the next step, a dataset is obtained to form the classification algorithms' input using the extracted features.

The significant difference between this research and previous works available in the literature is considering three heart sound classes instead of two. More specifically, while in previous studies, the data were classified into two modes of normal and abnormal sounds, in the current research, the heart sounds are classified into three classes of normal, abnormal due to the third heart sound, and abnormal due to the fourth heart sound. As each of the abnormal sounds can cause different heart problems, this approach can analyze the sounds better to

* Corresponding author.
E-mail addresses: yasser.zeinali@gmail.com (Y. Zeinali), Niaki@Sharif.edu (S.T.A. Niaki).

https://doi.org/10.1016/j.mlwa.2021.100206
Received 30 March 2021; Received in revised form 29 October 2021; Accepted 31 October 2021
Available online 12 November 2021

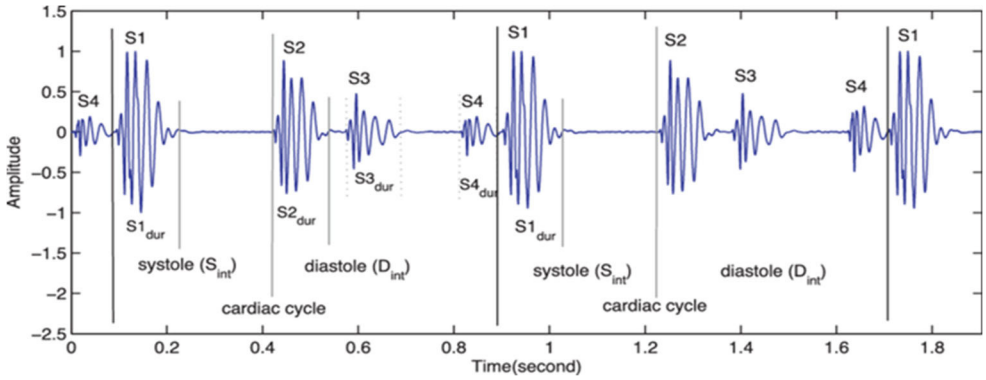


Fig. 1. Types of heart sounds.

make better decisions. Another contribution of this paper is to improve the accuracy of previous works on heart sound classifications.

The rest of this paper is organized as follows. Section 2 will review what has been done so far by machine learning algorithms to classify heart sounds and diseases. Section 3 explains the proposed methodology and its implementation in detail. Comparison and sensitivity analyses are performed in Section 4 to determine the strengths and weaknesses of each method. Finally, the best method is introduced, and future works are presented in Section 5.

2. Literature review

Long before the relatively new advent of artificial intelligence algorithms to classify heart sounds, various statistical studies have been performed in healthcare analytics. However, due to the great importance and the special view created by both engineers and physicians on this issue, significant progress has been made during a short period. In this section, the activities performed in heart sounds classification and heart disease diagnosis are explicitly examined. Then, data mining techniques and machine learning algorithms in various areas of health are briefly reviewed to explore the role of artificial intelligence in decision support systems.

Here, we gathered some of the research in heart sound classification in chronological order. Jia et al. (2012) tried to classify normal and abnormal sounds using fuzzy neural networks, which were initially extracted by features such as DWT and Shannon Entropy. The research used actual data collected by the researchers themselves. Deng and Han (2016) utilized the auto-correlation of sounds in the PASCAL dataset (one of the well-known datasets in the field of heart sound classification) and employed the discrete wavelet transform (DWT) technique for the feature extraction part. They also achieved a high accuracy using the SVM algorithm. However, the authors did not consider the use of signal filters to clean the dataset. Zabihi et al. (2016) conducted research that utilized an ensemble neural network algorithm to classify normal and abnormal heart sounds. In this study, the PhysioNet dataset was used, for which 91.5 percent accuracy was achieved. Besides, Potes et al. (2016) used deep learning to perform classification using AdaBoost and CNN algorithms, recording an accuracy rate of 94.24%. In this study, a sample consisting of 665 abnormal and 2575 normal sounds was used.

In 2017, research led to two articles' publication (Zhang et al., 2017a, 2017b). In the first article, the authors tried to classify heart sounds using spectrograms and a regression-based method, where the PASCAL dataset and the SVM algorithm were used. In the second article, the PhysioNet dataset was examined and analyzed, for which the feature extraction process was performed by the tensor decomposition method. They classified the dataset using the SVM algorithm. Another research conducted by Arabasadi et al. (2017) was about utilizing a hybrid neural network-Genetic algorithm. They proposed a method to improve the performance of the neural network by enhancing its initials weights. Using such methodology, they achieved accuracy, sensitivity, and specificity rates of 93.85%, 97%, and 92%, respectively. In another work, Dominguez-Morales et al. (2017) first classified the normal and abnormal sounds, and then, using electrical circuits and sensors, an automatic sound type detection system was invented. They used the deep convolutional neural network (DCNN) and recorded an accuracy of about 97%. In this study, the heart sounds were converted to sonogram shapes and placed as the input of the deep learning algorithms. Khateeb and Usman (2017) utilized kNN and Naïve Bayesian (NBs) algorithms on heart disease data, including 303 samples with 14 features. The researchers divided the test results into six categories, with the best results at 79.2% for kNN and 66.6% for NBs. Hamidi et al. (2018) employed curve fitting as a tool to extract features and a kNN algorithm to classify the normal and abnormal sounds of the heart, reaching an accuracy rate of 92%.

Recently, Noman et al. (2019) provided a framework for automatic heart sound detection based on neural network algorithms. The best performance was related to the combination of 1D-CNN and 2D-CNN with accuracy and sensitivity of 89.22% and 89.94%, respectively. Moreover, Zhang et al. (2019) used the long short-term memory (LSTM) network to classify sounds with an accuracy of 94.66 percent. They used the PhysioNet dataset as well. Kui et al. (2021) also utilized the dynamic frame length method to extract log Mel-frequency spectral coefficients (MFSC) features from the heart sound signal based on the heart cycle. Afterward, the convolution neural network (CNN) was used to classify the MFSC features. Finally, the majority voting algorithm was used to get the optimal classification results. Researchers know that patients do not know the device to understand the abnormal heart sounds; therefore, they proposed a system that enables remote patient monitoring by integrating advanced wireless communications with a customized low-cost stethoscope. A smartphone application also facilitates recording, processing, visualizing, listening to, and classifying heart sounds for patients and specialists. They built their classification model using the Mel-Frequency Cepstral Coefficient and Hidden Markov Model and tested the application in a hospital environment. The smartphone application correctly detected 92.68% of abnormal heart conditions in clinical trials at UT Southwestern Hospital (Thiyagaraja et al., 2018). Another research that utilized a new algorithm to classify the heart sounds was (Bilal, 2021). He proposed a model based on Local Binary Pattern (LBP) and Local Ternary (LTP) Pattern features and deep learning. He then used these methods to extract features from the heart sounds and feed them to a Dimensional Convolutional Neural Network (1D-CNN) to complete the classification process. Experiments were done with the help of two popular datasets that were used in the context to determine the efficiency of distinct techniques. These datasets are PASCAL and PhysioNet 2016. He obtained 91.66% and 91.78% accuracy rates for the datasets, respectively.

One of the most common diseases, especially among women, is breast cancer, which has been the subject of many articles and researches for a long time. Chen and Yang (2012) analyzed a breast

cancer dataset from 97 patients using support vector machine (SVM) algorithms and a combination of genetic and SVM algorithms. Joshi and Mehta (2018) used the well-known kth nearest neighbor (kNN) algorithm to investigate breast cancer data's function. These data included 569 samples with 32 features. They used principal component analysis (PCA) and linear discriminant analysis (LDA) to reduce the data dimension. The result showed that kNN with LDA- reduction technique was better than kNN without dimension reduction and kNN with PCA. Their results were accurate 97.06%, 95.29%, and 95.88% of the time, respectively. Septiani et al. (2017) employed the kNN algorithm and tested 670 patients with nine features to achieve an accuracy of about 98 percent. Researchers are also trying to diagnose cancer by examining medical photographs and image processing algorithms. Kaymak et al. (2017) used 176 breast cancer images of benign and malignant cancers. They utilized a recurrent neural network technique and a neural network with the radial kernel to achieve 59% and 70% accuracy.

In addition to breast cancer, many diseases, including lung cancer, have been studied and analyzed by researchers. Kaucha et al. (2017) performed image processing on lung images using the K-means clustering approach in the image segmentation stage and finally using the SVM algorithm to diagnose cancer. They reached a 95.16% accuracy rate. Recently, Mousavi et al. (2021) proposed an intelligent classification algorithm comprising a fuzzy rule-based approach, a harmony search (HS) algorithm, and a heuristic to classify medical datasets intelligently. They used nine well-known medical datasets to evaluate the efficiency of their proposed approach. Maleki et al. (2021) used the kNN algorithm with an optimized k on a dataset, in which a genetic algorithm extracted its essential features. In the end, their result illustrated that their implemented technique gave 100% accuracy in diagnosing the stage of disease in lung cancer patients.

The above brief review is just a few previous activities on implementing artificial intelligence algorithms to predict, diagnose, or classify diseases. Obviously, by the use of more accurate data, more critical and better results can be achieved. This requires more cooperation between the health care sector and the engineering departments. In the next section, the performances of these algorithms in the field of heart sounds classification are examined to gain a deeper understanding of the subject.

3. Methodology and implementation

The proposed methodology's general framework is shown in Fig. 2, for which each box will be explained in this section separately. Before doing this, the dataset is introduced first.

3.1. The dataset

The dataset used in this study is part of the PhysioNet and PASCAL challenge data collected carefully by a physician (https://physionet.org/physiobank/database – http://www.peterjbentley.com/heartchallenge). It contains 650 samples, of which 104 patients with abnormal and 52 patients with normal heart sounds are selected to create a dataset consisting of 156 heart sound samples (Bentley et al., 2011; Liu et al., 2016).

3.2. Data-preprocessing

As the heart sounds in the dataset are generally recorded in noisy environments with a relatively large number of patients and medical staff, a heart sound specialist preprocessed them to obtain better quality data before using them in the experiment. It should be noted that the heart works in the vicinity of the human lung and is naturally accompanied by the sound of the lungs or human respiration, which causes a high-frequency sub-sound to be present. Fig. 3 is an illustration of a noisy heart sound. In this regard, one needs to eliminate the noises using different signal filters, such as the Savitzky–Golay filter (Potes et al., 2016) used in the current research.

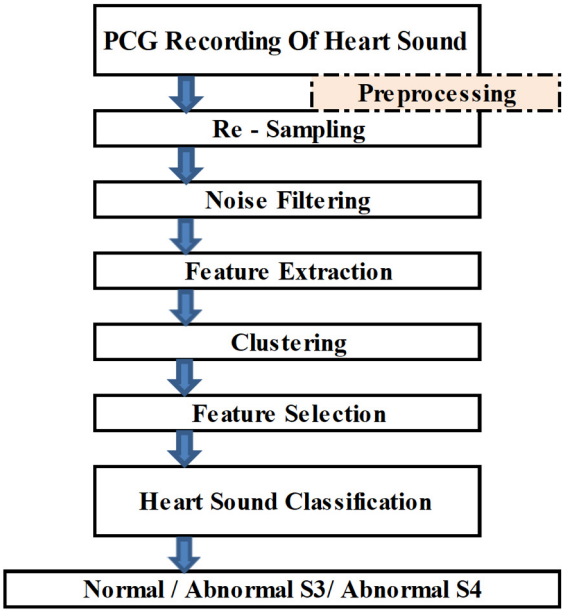


Fig. 2. The general framework of the proposed methodology.

3.3. Feature extraction

A feature is a measurable characteristic of a phenomenon that is observed. As the phenomenon possesses different features, when they are extracted together, a feature vector is obtained. For instance, if x_1 denotes the weight, x_2 is the height, ..., and x_n is the sex of a person, then the feature vector is denoted by $\boldsymbol{X} = [x_1, x_2, \ldots, x_n]$. To extract features from the heart sounds, since the average human heart cycle is 0.8 s, each of the four heart sounds described previously is divided into five separate ranges, each with an average duration of 0.16 s. Consequently, the dataset created consists of features obtained in 5 different ranges. The feature extraction procedure of the current work is explained using Fig. 4. Each of the boxes shown in this figure is explained in the following subsections. Note here that another dataset that contains only the Mel-Frequency Cepstral Coefficients (MFCC) (Hasan et al., 2004) is also created to compare the performance of the proposed approach to the ones available in the literature. This will be discussed later.

3.3.1. Statistical features

The statistical features include standard deviation, skewness, and kurtosis. They are used to examine how the data is distributed.

3.3.2. Signal features

The signal features include amplitude and dominant frequencies. Amplitude is the maximum displacement or distance made by a point on a wave measured from its equilibrium position. Besides, as the lowest frequency component is known as the fundamental frequency, the dominant frequency is the fundamental frequency with the highest amplitude (Giron-Sierra, 2016). For instance, the fundamental frequency of about 450 Hz in Fig. 5 is the dominant frequency.

3.3.3. Wavelet features

A wavelet series represents a real- or complex-valued function by a particular orthonormal series generated by a wavelet. Wavelet transformation is one of the most important mathematical transformations used in various fields of science. The main idea of the wavelet transformation is to overcome the weaknesses and limitations of the Fourier transformation. Unlike the Fourier transformation, this transformation can be

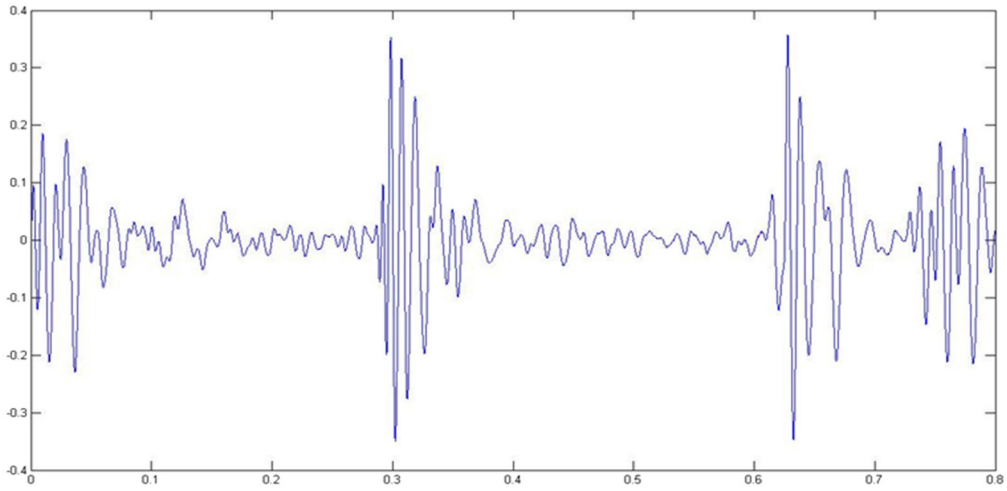


Fig. 3. Noisy heart sound.

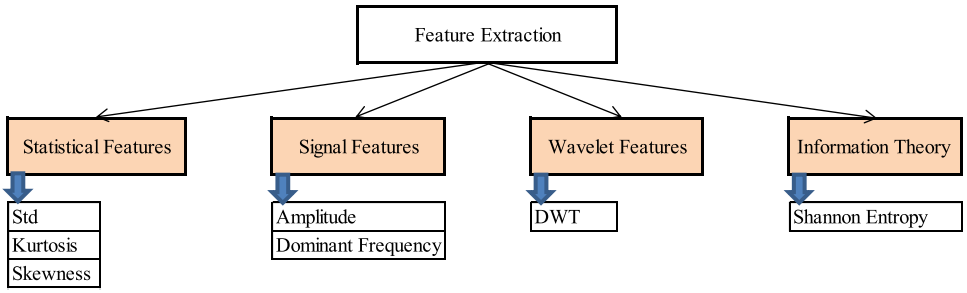


Fig. 4. Feature extraction process.

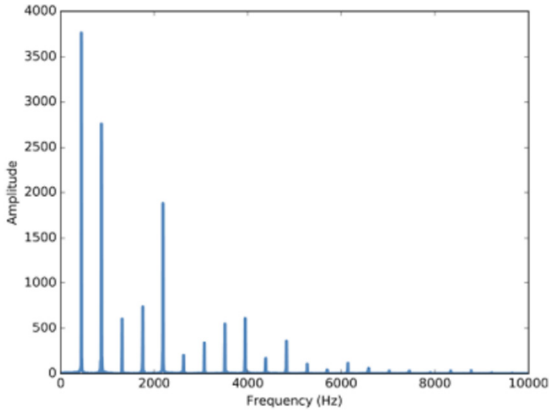


Fig. 5. Amplitude and frequency of a signal.

used for non-stationary signals (Akansu et al., 2010). Many of the real-world signals have non-stationary nature. In practical projects, when one tries to process ECG signals, stock market data, sensors data, and so on, she/he is more likely to encounter non-stationary signals of some dynamic systems. An excellent solution for processing non-stationary signals is the use of wavelet transformations instead of Fourier transformations. In this study, the discrete wavelet transform (DWT) approach is utilized for the heart sound data that is non-stationary.

The discrete wavelet transform (DWT) is a wavelet transform for which the wavelets are discretely sampled. The most commonly used set of DWTs was formulated by the Belgian mathematician Ingrid Daubechies in 1988. This formulation is based on utilizing recurrence relations to generate progressively finer discrete samplings of an implicit mother wavelet function; each resolution is twice that of the previous scale. This transformation is employed in the current work to extract wavelet features from the processed heart sounds.

3.3.4. Information theory

Information theory is the mathematical treatment of the concepts, parameters, and rules governing messages transmission through communication systems (Smelser & Baltes, 2001). Although there are many different concepts and techniques related to this area, the key measure is entropy. The entropy of a discrete random variable X with the mass probability function $p(x)$ is denoted by $H(x)$ and is defined as:

$$H(x) = E(I(x)) = -log_b(p(x)) \tag{1}$$

While the logarithmic basis b in Eq. (1) can be 2, e, or 10, which calculates the entropy unit in a bit, nat, and Hartley, respectively, the most common basis to measure information is shown in Eq. (2) that obtains the information in the bit or Shannon unit (Cover & Thomas, 2012).

The Mel Frequency Cepstral Coefficients (MFCCs) are widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980s and have been state-of-the-art ever since (Hasan et al., 2004). The formulae to convert the frequency to Mel scale are

$$M(f) = 1125\ln(1 + f/700) \tag{2}$$

$$\Delta f_{mel} = {}^{mel(f_{max})}/_{L} \tag{3}$$

$$c(l) = 700\left[10^{\frac{l\Delta f_{mel}}{2595}} - 1\right] \quad l = 1, 2, \ldots, L. \tag{4}$$

In the above equations, f is the frequency given in Hertz, Δf_{mel} is determined according to the upper limit frequency f_{max} and the number of filters L, and $c(l)$ is the center frequency of the filter.

3.4. The feature selection algorithms

Feature selection algorithms are used to deal with high-dimensional data. These algorithms can be defined as the process of identifying relevant features and eliminating unrelated and repetitive features to observe a subset of traits that best describe the problem. Selecting a subset of data can reduce the data size and storage space required, affecting the processing time.

Feature selection algorithms have a wide range, including statistical-based, artificial intelligence-based, and meta-heuristic algorithms. Statistical-based algorithms such as forward selection, backward selection, hybrid models, and even the use of P-value are among the widely-used algorithms. In the field of artificial intelligence-based algorithms, one can name genetic algorithms (GA), particle swarm optimization (PSO), simulated annealing (SA), etc. One of the most advanced algorithms for feature selection is GA. GA is a stochastic method for function optimization based on the mechanics of natural genetics and biological evolution. Some advantages of genetic algorithms are the following:

- They outperform traditional feature selection methods.
- GAs handle datasets with many features.
- They do not need specific knowledge about the problems being studied.

After creating a set of data obtained from the heart's sounds in 156 rows and 65 columns with a total of 10 140 data, in the current research, the desired features are selected using the well-known GA due to its performance and applicability. The loss function used in this algorithm to be minimized is the weighted sum of the miss-classification ratio (mcr) and the number of selected features (n_f) shown in Eq. (5).

$$Min\, Z = w_1 * mcr + w_2 * n_f. \tag{5}$$

Using this objective function, GA tries to find the best combination with the minimum number of features that minimize both the cost and the misclassification rate. Here, the stopping criterion to end the iterations is chosen to be a predefined number of iterations.

If we divide the equation by w_1, we will have:

$$Min\, Z = mcr + {}^{w_2}/_{w_1} * n_f. \tag{6}$$

Assuming ${}^{w_2}/_{w_1} = W$. Therefore, the objective function becomes:

$$Min\, Z = mcr + W * n_f. \tag{7}$$

Now, this W can be defined as:

$$W \propto mcr \rightarrow W = \beta * mcr \rightarrow Min\, Z = mcr + \beta * mcr * n_f \tag{8}$$

Finally, the objective function will be like this:

$$Min\, Z = mcr(1 + \beta * n_f). \tag{9}$$

β can be defined as a penalty for having an additional feature ($0 \leq \beta \leq 1$).

3.5. The dimension reduction algorithms

The performance of machine learning algorithms degrades when there are too many input variables. Having a large number of dimensions in the feature space implies a substantial space volume, and in turn, the points in space (rows of data) often represent a small and non-representative sample. This problem that can dramatically impact machine learning algorithms' performance is known as the "curse of dimensionality". In this study, two algorithms, namely the principal components analysis (PCA) and the linear discriminant analysis (LDA), reduce the data dimension.

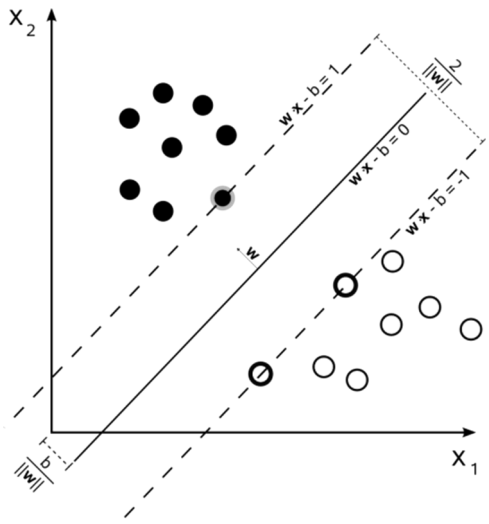


Fig. 6. Classifying by SVM.

PCA is a simple and efficient linear transformation method. The primary purpose of PCA analysis is to recognize patterns in data by identifying the variables' correlations. If there is a strong correlation, efforts to reduce the dimensions will be significant. In general, PCA finds the maximum direction of variance in high-dimensional data and plots it in sub-dimensions with fewer dimensions to retain most of the information.

LDA aims to project a dataset onto a lower-dimensional space with good class separability to avoid overfitting and reduce computational costs. The general LDA approach is very similar to a PCA, with the difference that in addition to finding the component axes that maximize the variables' variance, finding the axes that maximize the separation between multiple classes is also aimed.

3.6. Machine learning algorithms

The last step of this study is to cluster and classify the dataset. To this aim, while three well-known clustering algorithms, including K-means, DBscan, and hierarchical clustering algorithms, are used for clustering, support vector machines classifier (SVC), gradient boosting classifier (GBC), and random forest classifier (RFC) algorithms are utilized in the current research to classify the sounds. The classifiers are discussed in the following subsections.

3.6.1. Support vector machine algorithm

The support vector machine (SVM) is one of the most powerful machine-learning algorithms. It is a supervised machine learning algorithm that can be used for both classification and regression problems (Huang et al., 2006). This algorithm first takes the data to a higher spatial dimension so that it can create a distance between them by one or more hyperplanes. This can generally be used to draw multiple lines between data as decision lines; among them, the line that reduces the risk of categorization is found. Fig. 6 depicts this approach.

Suppose that the data used in the model to learn is: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where the y_i are either 1 or -1, each indicating the class to which the point x_i belongs. We want to find the "maximum-margin hyperplane" that divides the group of points x_i for which $y_i = 1$ from the group of points for which $y_i = -1$, which is defined so that the distance between the hyperplane and the nearest point from either group is maximized. A hyperplane can be written as:

$$W^T x - b = 0 \tag{10}$$

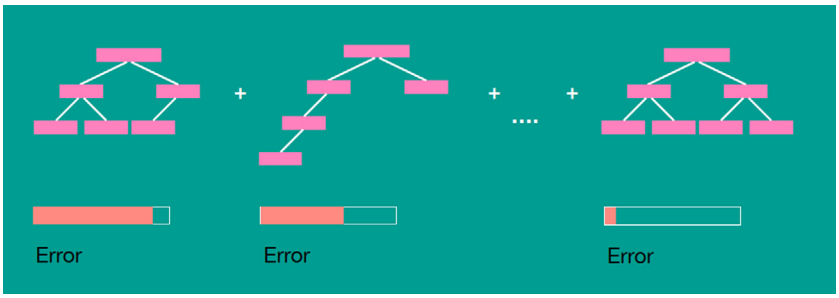


Fig. 7. Updating a model using a gradient boosting method.

where W is the vector to the hyperplane. In this equation, the parameter $\frac{b}{\|W\|}$ determines the offset of the hyperplane from the origin along the normal vector W.

3.6.2. Gradient boosting algorithm

Gradient boosting algorithm (GBA) is an ensemble algorithm that provides a high-efficient prediction or classification when dealing with large amounts of data. The algorithm combines the predictions of several basic estimators to improve performance. It includes several weak or medium estimators that together create a strong classifier or regressor. The algorithm possesses a boosting-based reinforcement that seeks to minimize the prediction error and, therefore, minimizes this error by adding new models (Mason et al., 1999). GBA can optimize different loss functions and provide several hyper-parameter tuning options that make the function very flexible.

Suppose that the data used in the model to learn is as follows: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ And the goal of learning is to minimize a loss function called $L(y, F(x))$ defined as

$$F = argmin\ \mathrm{E}_{x,y}[L(y, F(x))]. \tag{11}$$

This process is performed in iterations to find the final model as:

$$F(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + F_0 \tag{12}$$

Here h_is are the models selected from a group of models called H, for example. This set can be a collection of decision trees, where the first model is a fixed number called F_0 selected as follows:

$$F_0 = argmin \sum_{i=1}^{n} L(y_i, \gamma). \tag{13}$$

Fig. 7 illustrates how the model is updated using the gradient boosting method.

3.6.3. Random forest algorithm

The random forests algorithm (RFA) or random decision forests algorithm is created by aggregating a multitude of decision trees. This algorithm is an ensemble learning method for classification, regression, and other tasks. Random forest works very well with high-dimensional data since it considers subsets of data. That is why it is faster than decision trees in the training phase, in which hundreds of features can be easily handled. As the name implies, this algorithm generates a forest as a group of decision trees randomly. The forest's construction is often performed using the Bagging method, which combines learning models to improve the overall performance (Breiman, 2001). The random forest algorithm adds randomness to the model as the trees grow. This leads to more variety and ultimately a better model. One of the most important advantages of this algorithm is the lack of overfitting, the main problem of many algorithms in this field. The general framework of the RFA is depicted in Fig. 8.

The training algorithm for random forests applies the general bootstrap aggregating or bagging technique to tree learners. Suppose that the data used in the model to learn is as follows: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

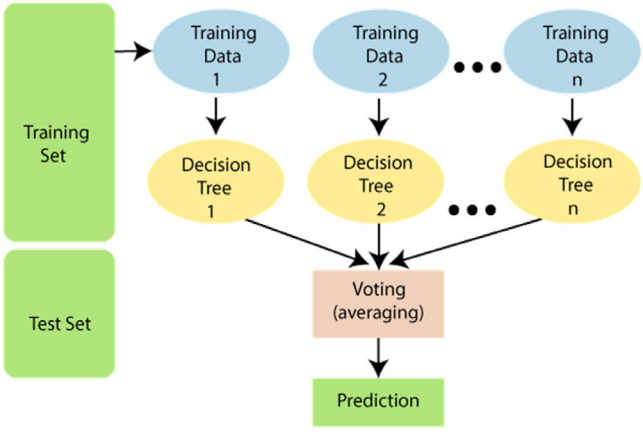


Fig. 8. The general framework of the random forest algorithm.

Bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples:

For $r = 1, 2, \ldots, R$ different random samples:

1. We provide samples with replacement (n training examples from X, Y; call these X_r, Y_r)
2. Train a classification or regression tree T_r on X_r, Y_r.

After training, predictions for the test set x_{test} can be made by averaging the predictions (or by taking the majority vote in the case of classification trees) from all the individual regression trees on x_{test} :

$$\hat{T} = {}^{1}/_{R} \sum_{r=1}^{R} T_r(x_{test}) \tag{14}$$

3.7. Performance criteria

The performance of a classifier can be measured through some factors, all of them gained from a confusion matrix. The confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Table 1 demonstrates a confusion matrix for which the terms used are defined as follows.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{15}$$

$$Sensitivity = {}^{TP}/_{(TP+FN)} \tag{16}$$

$$Specificity = {}^{TN}/_{(TN+FP)} \tag{17}$$

$$Precision = {}^{TP}/_{(TP+FP)} \tag{18}$$

$$F = 2 * {}^{(Precision*Sensitivity)}/_{(Precision+Sensitivity)} \tag{19}$$

4. Feature extraction procedure

After explaining various steps involved in using the algorithms in the previous sections, the algorithms are implemented on a heart sound example by going through the steps in this section. The first three features, which are statistical features related to data distribution (standard deviation, skewness, and kurtosis) of one sample, are as follows:

Standard Deviation = [0.040656173, 0.043414002, 0.052018143, 0.052094222, 0.055029247]
Skewness = [0.288743681, 0.184201601, 0.035266239, 0.104826682, 0.117545241]
Kurtosis = [6.092877232, 4.431667889, 5.159243178, 5.221625844, 5.170539205]

In the next step, the dominant frequencies of all audio windows are examined. To this aim, some of the highest amplitudes of a frequency as signal peaks are first extracted as follows:

Amplitude_Section_1 = [0.015862133, 0.024666118, 0.027113745]
Amplitude_Section_2 = [0.038273228, 0.028721747, 0.007619529]
Amplitude_Section_3 = [0.022935687, 0.012337156, 0.028096329]
Amplitude_Section_4 = [0.017190622, 0.00700349, 0.018721624]
Amplitude_Section_5 = [0.011881425, 0.021728186, 0.016036953]

Then, the frequencies of each of the peaks specified above are the following dominant frequencies that will be used as inputs for the machine learning algorithms.

Dominant Fequency_Section_1 = [18, 11, 23]
Dominant Fequency_Section_2 = [11, 18, 12]
Dominant Fequency_Section_3 = [13, 20, 14]
Dominant Fequency_Section_4 = [13, 24, 8]
Dominant Fequency_Section_5 = [30, 8, 24]

The next part of the extraction process is the features related to wavelet transformations. In this study, after testing different wavelets, the Daubechies wavelet is chosen to extract the features. The wavelet coefficients in the sample signal are as follows:

Wavelet Coef_Section_1 = [0.026039559, 0.029882816, 0.043024967]
Wavelet Coef_Section_2 = [0.044132106, 0.051816736, $4.13 * 10^{-7}$]
Wavelet Coef_Section_3 = [$9.92 * 10^{-7}$, $2.40 * 10^{-6}$, $1.02 * 10^{-6}$]
Wavelet Coef_Section_4 = [$1.72 * 10^{-5}$, $8.02 * 10^{-8}$, $1.25 * 10^{-7}$]
Wavelet Coef_Section_5 = [$1.23 * 10^{-7}$, $1.40 * 10^{-7}$, $1.88 * 10^{-7}$]

The last step in the feature extraction section is devoted to Shannon Entropy, for which the coefficients are:

Shannon Entropy_Section_1 = [150.100672514795]
Shannon Entropy_Section_2 = [174.271610942798]
Shannon Entropy_Section_3 = [228.377602886118]
Shannon Entropy_Section_4 = [227.738726560246]
Shannon Entropy_Section_5 = [249.600061896928]

As mentioned previously, another dataset is also created using the features extracted by MFCC, for which the coefficients for the sample audio are as follows:

Mel-Frequency Cepstral Coefficients =

[−487.6704903 81.35004869 66.328044 47.3091901 30.41498369
19.94921547 16.82941564 18.71784134 21.63196296 22.11459924
18.83711975 12.89583605 6.818045198 2.972303966 2.303917322
4.023118871 6.282943959 7.341897362 6.500594951 4.317696586
2.083883542 0.957989067 1.306706764 2.599891964 3.846936426
4.245269912 3.627619682 2.467131116 1.495832325 1.217654846
1.625837376 2.273160383 2.609815599 2.34730993 1.617275924
0.847753671 0.464150304 0.619766938 1.121765609 1.580701382]

According to the extracted features from heart sounds and the division of each sample into 5 parts (0.8 s for each part), a total of 65 columns of features as a set of statistical features, features related to amplitude and frequency, and features included in wavelet and Shannon entropy is found for each sound. In the next step, the abnormal sounds are clustered into S_3, and S_4 sounds using the K-means, DBscan, and hierarchical clustering algorithms. Fig. 9 shows the flowchart of the proposed clustering approach.

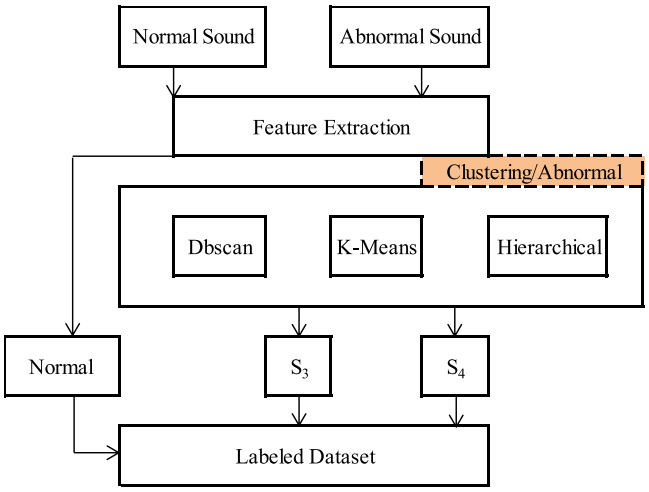


Fig. 9. The clustering approach.

Table 1
The confusion matrix.

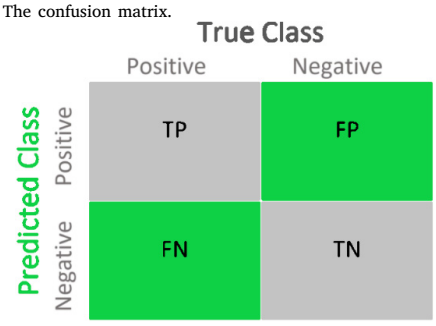


Table 2
The confusion matrix of the gradient boosting algorithm.

		Predicted classes		
		Normal	Abnormal S_3	Abnormal S_4
Real classes	Normal	4	1	1
	Abnormal S_3	0	5	0
	Abnormal S_4	0	0	5

5. Results

After creating the labeled dataset, the classification algorithms were utilized before and after executing the dimension reduction and the feature selection algorithms. Before implementing the dimension reduction and the feature selection algorithms, the algorithms' outcomes, including the confusion matrices and the performance measures in terms of precision, recall, and F1-score, were shown in Tables 2–7. As seen in these tables, the gradient boosting algorithm performed the best in terms of all performance measures. Besides, the best accuracy was achieved by using the gradient boosting algorithm with an accuracy of 87.5%, the number of estimators for this algorithm was 1000, and the learning rate was 0.01. This algorithm classified 14 out of 16 possible sounds correctly according to its confusion matrix. Meanwhile, the random forest (with 800 trees) and the support vector machine (with radius basis function kernel) algorithms had 81.25% and 75% accuracy.

The algorithms were also implemented after utilizing the principal component analysis and linear discriminant analysis to reduce dimensions. The first step in implementing PCA was the determination of the number of components. The variance shares of the dataset explained by the components are shown in Fig. 10. A good result was achieved by

Table 3
The performance of the gradient boosting algorithm.

	Performance criteria			
	Precision	Recall	F1-Score	Time
Normal	1	0.67	0.8	1.1032 s
Abnormal S_3	0.83	1	0.91	
Abnormal S_4	0.83	1	0.91	
Accuracy	87.50%			

Table 4
The confusion matrix of the random forest algorithm.

		Predicted classes		
		Normal	Abnormal S_3	Abnormal S_4
Real classes	Normal	4	1	1
	Abnormal S_3	0	5	0
	Abnormal S_4	0	1	4

Table 5
The performance of the random forest algorithm.

	Performance criteria			
	Precision	Recall	F1-Score	Time
Normal	1	0.67	0.8	1.0451 s
Abnormal S_3	0.71	1	0.83	
Abnormal S_4	0.8	0.8	0.8	
Accuracy	81.25%			

Table 6
The confusion matrix of the support vector machine algorithm.

		Predicted classes		
		Normal	Abnormal S_3	Abnormal S_4
Real classes	Normal	2	1	3
	Abnormal S_3	0	5	0
	Abnormal S_4	0	1	4

Table 7
The performance of the support vector machine algorithm.

	Performance criteria			Time
	Precision	Recall	F1-Score	
Normal	1	0.5	0.67	0.7554 s
Abnormal S_3	0.71	1	0.83	
Abnormal S_4	0.67	0.8	0.73	
Accuracy		75.00%		

considering 36 components that explain 99% of the dataset variance. As the maximum number of components in LDA is equal to the number of classes, which is 3, three components were used to implement this algorithm. In the case of using GA, after implementing the algorithm, it gave us 21 features. We then deployed the ML algorithms to analyze the results on each of the datasets we obtained after dimension reduction algorithms. The shape of these datasets were (156, 36), (156, 3), and (156, 21) based on PCA, LDA, and GA, respectively.

The next step was to implement the ML algorithms, including gradient boosting classifier (GBC), support vector classifier (SVC), and random forest classifier (RFC) alongside the GA that selects appropriate features. The outcomes (accuracy) are shown in Table 8. The results in Table 8 revealed that the best performance was achieved when the RFC or the SVC algorithm alongside GA with 78% accuracy was employed. Moreover, the outcomes showed that the accuracy decreases sharply when the dimension reduction algorithms (PCA or LDA) are used.

Finally, the Mel Frequency Cepstral Coefficients (MFCC) dataset was used to compare the outcomes with the ones obtained in the literature in classifying normal and abnormal sounds (for 2 and 3 classes) of the heart. Table 9 contains the implementation results.

For the heart sound classified into three classes, the outcomes in Table 9 showed that the best accuracy was 95% when the gradient boosting algorithm was used to classify the sound. In the two-class classification, the best result was 98% in terms of accuracy. Once again, note in this table that the dimension reduction approach using PCA, LDA, and GA not only did not add to the accuracy of the three classification algorithms but also decreased them. The purpose of using dimension reduction algorithms was to show the procedure and results theoretically.

6. Conclusion and future work

Artificial intelligence in healthcare management is promising, especially in diagnosing different diseases such as cardiovascular problems. Many scientists and researchers have worked in this field and achieved accurate results in many cases in recent years. This article utilized signal processing and machine learning algorithms to classify heart sounds into different classes. While in previous works, researchers classified heart sounds into two categories (normal and abnormal), in the current study, we used the available techniques in data analytics to divide heart sounds into three classes (normal, abnormal of the third type, and abnormal of the fourth type). This enables heart specialists to detect cardiovascular disease more accurately. We should mention that the analysis was based on stationary and non-stationary signals for normal and abnormal heart sounds data. Due to the nature of the raw data collected with a sampling frequency of 2000 kHz, the data were first reviewed and preprocessed under the supervision of a specialist. The analysis was made by extracting various heart sound information using statistical, discrete wavelet, and information theory features. Then we reduced the dimension of the dataset obtained by PCA, LDA, and GA by selecting appropriate features and classifying the sound using classification algorithms such as GBC, RFC, and SVC. The analysis and the codes related to the feature extractions and GA were written in MATLAB. We also utilized Python packages (Pandas, Numpy, Matplotlib, Scikitlearn, Scipy, etc.) for all the classification algorithms and data analysis. We also provided extensive comparative analysis to come up with the best technique in our experiments. The comparisons were made by performing all the processes and procedures on sets of preprocessed heart data.

Although various types of digital stethoscopes are available to the medical community, none can analyze sounds and differentiate the disease. As such, examining and analyzing heart sounds with the help of artificial intelligence algorithms can be a basis to build a device in the future to help heart specialists to make better decisions. The concepts of the Internet of Things (IoT) can also be combined with the concepts mentioned in this study to provide a basis for creating such a device in the future.

CRediT authorship contribution statement

Yasser Zeinali: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization, Investigation, Validation. **Seyed Taghi Akhavan Niaki:** Conceptualization, Methodology, Visualization, Supervision, Validation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

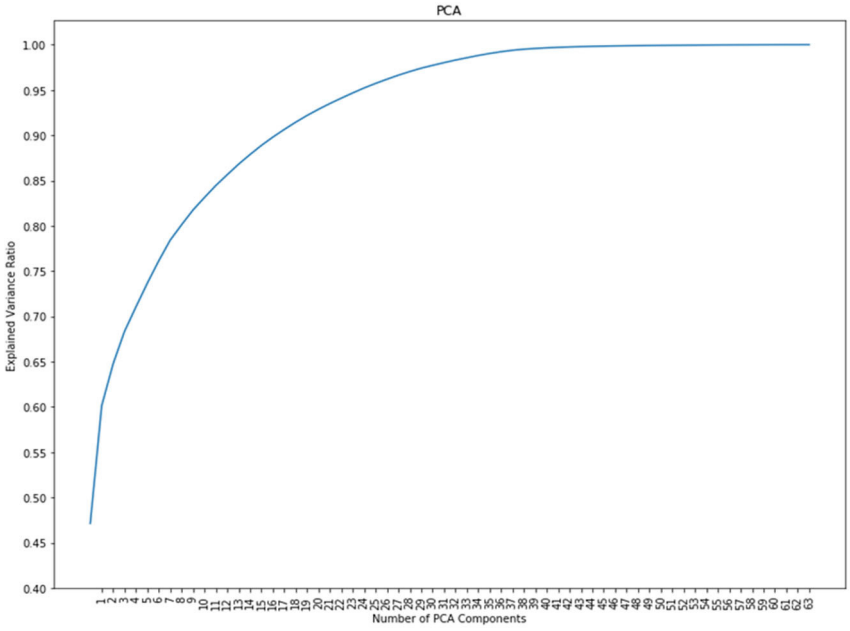


Fig. 10. Explained variance ratio of the PCA components.

Table 8
The accuracy of ML algorithms.

Main feature extraction	ML algorithms performance		
Experiments	GBC/Time	RFC/Time	SVC/Time
3 Classes + PCA	50.00% – 0.9728 s	61.10% – 0.8954 s	61.10% – 0.6179 s
3 Classes + LDA	56.00% – 0.5164 s	56.00% – 0.5262 s	56.00% – 0.3457 s
3 Classes + GA	67.00% – 1.0587 s	78.00% – 0.9130 s	78.00% – 0.7306 s

Table 9
The outcomes of the ML algorithms on the MFCC dataset.

MFCC feature selection	ML algorithms performance		
Experiments	GBC/Time	RFC/Time	SVC/Time
3 Classes	95.00% – 1.4874 s	88.00% – 1.349 s	88.00% – 1.1076 s
3 Classes + PCA	74.00% – 1.2854 s	77.00% – 1.3181 s	91.00% – 0.9856 s
3 Classes + LDA	68.00% – 0.7815 s	70.00% – 0.8701 s	84.00% – 0.4334 s
3 Classes + GA	74.00% – 1.2561 s	78.25% – 1.1027 s	81.20% – 0.9945 s
2 Classes	98.00% – 0.9667 s	94.00% – 0.9812 s	82.00% – 0.9671 s
2 Classes + PCA	88.50% – 0.8671 s	91.00% – 0.8812 s	75.00% – 0.7345 s
2 Classes + LDA	78.50% – 0.534 s	76.00% – 0.899 s	72.20% – 0.7123 s
2 Classes + GA	80.00% – 0.8566 s	78.00% – 0.8255 s	76.50% – 0.7277 s

References

Akansu, A., Serdijn, W., & Selesnick, I. (2010). Wavelet transforms in signal processing: a review of emerging applications. *Physical Communication*, *3*, 1–18. http://dx.doi.org/10.1016/j.phycom.2009.07.001.

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*, 19–26. http://dx.doi.org/10.1016/j.cmpb.2017.01.004.

Bentley, P., Nordehn, G., Coimbra, M., Mannor, S., & Getz, R. (2011). Classifying heart sounds challenge. Retrieved from Classifying Heart Sounds Challenge: http://www.peterjbentley.com/heartchallenge.

Bilal, E. M. (2021). Heart sounds classification using convolutional neural network with 1D-local binary pattern and 1D-local ternary pattern features. *Applied Acoustics*, *180*, Article 108152. http://dx.doi.org/10.1016/j.apacoust.2021.108152.

Bonow, R. O., Mann, D., Zipes, D., & Libby, P. (2011). *Braunwald's heart disease: A textbook of cardiovascular medicine, single volume*. Elsevier Science.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Chen, A. H., & Yang, C. (2012). The improvement of breast cancer prognosis accuracy from integrated gene expression and clinical data. *Expert Systems with Applications*, *39*, 4785–4795. http://dx.doi.org/10.1016/j.eswa.2011.09.144.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Deng, S.-W., & Han, J.-Q. (2016). Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps. *Future Generation Computer Systems*, *60*, 13–21. http://dx.doi.org/10.1016/j.future.2016.01.010.

Dominguez-Morales, J. P., Jimenez-Fernandez, A. F., Dominguez-Morales, M. J., & Jimenez-Moreno, G. (2017). Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors. *IEEE Transactions on Biomedical Circuits and Systems*, *12*, 24–34. http://dx.doi.org/10.1109/TBCAS.2017.2751545.

Giron-Sierra, J. M. (2016). *Model-based actions and sparse representation: vol. 3, Digital signal processing with matlab examples*. Springer.

Hamidi, M., Ghassemian, H., & Imani, M. (2018). Classification of heart sound signal using curve fitting and fractal dimension. *Biomedical Signal Processing and Control*, *39*, 351–359. http://dx.doi.org/10.1016/j.bspc.2017.08.002.

Hasan, R., Jamil, M., Rabbani, G., & Rahman, S. (2004). Speaker identification using mel frequency cepstral coefficients. In *The proceedings of the 3rd international conference on electrical & computer engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh*.

Huang, T.-M., Kecman, V., & Kopriva, I. (2006). *Kernel based algorithms for mining huge data sets (Vol. 1)*. Springer.

Jia, L., Song, D., Tao, L., & Lu, Y. (2012). Heart sounds classification with a fuzzy neural network method with structure learning. In J. Wang, G. G. Yen, & M. M. Polycarpou (Eds.), *Lecture notes in computer science: vol 7368, Advances in neural networks – ISNN 2012. ISNN 2012*. Berlin, Heidelberg: Springer, http://dx.doi.org/10.1007/978-3-642-31362-2_15.

Joshi, A., & Mehta, A. (2018). Analysis of k-nearest neighbor technique for breast cancer disease classification. *International Journal of Recent Scientific Research*, *9*, 26126–26130.

Kaucha, D. P., Prasad, P. W. C., Alsadoon, A., Elchouemi, A., & Sreedharan, S. (2017). Early detection of lung cancer using SVM classifier in biomedical image processing. In *2017 IEEE international conference on power, control, signals and instrumentation engineering (ICPCSI)* (pp. 3143–3148). IEEE, http://dx.doi.org/10.1109/ICPCSI.2017.8392305.

Kaymak, S., Helwan, A., & Uzun, D. (2017). Breast cancer image classification using artificial neural networks. *Procedia Computer Science*, *120*, 126–131. http://dx.doi.org/10.1016/j.procs.2017.11.219.

Khateeb, N., & Usman, M. (2017). Efficient heart disease prediction system using K-nearest neighbor classification technique. In *Proceedings of the international conference on big data and internet of thing* (pp. 21–26). http://dx.doi.org/10.1145/3175684.3175703.

Kui, H., Pan, J., Zong, R., Yang, H., & Wang, W. (2021). Heart sound classification based on log Mel-frequency spectral coefficients features and convolutional neural networks. *Biomedical Signal Processing and Control*, *69*, Article 102893. http://dx.doi.org/10.1016/j.bspc.2021.102893.

Leng, S., San Tan, R., Chai, K. T. C., Wang, C., Ghista, D., & Zhong, L. (2015). The electronic stethoscope. *Biomedical Engineering Online*, *14*, 1–37. http://dx.doi.org/10.1186/s12938-015-0056-y.

Liu, C., Springer, D., Li, Q., Moody, B., Juan, R. A., Chorro, F. J., & ... Clifford, G. D. (2016). An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, *37*(12), 2181. http://dx.doi.org/10.1088/0967-3334/37/12/2181.

Maleki, N., Zeinali, Y., & Niaki, S. T. A. (2021). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Systems with Applications*, *164*, Article 113981. http://dx.doi.org/10.1016/j.eswa.2020.113981.

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent. *Advances in Neural Information Processing Systems*, *12*, 512–518, https://dl.acm.org/doi/10.5555/3009657.3009730.

Mousavi, S. M., Abdullah, S., Niaki, S. T. A., & Banihashemi, S. (2021). An intelligent hybrid classification algorithm integrating fuzzy rule-based extraction and harmony search optimization: Medical diagnosis applications. *Knowledge-Based Systems*, *220*, Article 106943. http://dx.doi.org/10.1016/j.knosys.2021.106943.

Noman, F., Ting, C.-M., Salleh, S.-H., & Ombao, H. (2019). Short-segment heart sound classification using an ensemble of deep convolutional neural networks. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1318–1322). IEEE, http://dx.doi.org/10.1109/ICASSP.2019.8682668.

Potes, C., Parvaneh, S., Rahman, A., & Conroy, B. (2016). Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In *2016 computing in cardiology conference (CinC)* (pp. 621–624). IEEE.

Septiani, N. W. P., Wulan, R., & Lestari, M. (2017). Breast cancer detection using data mining classification methods. *Proceeding ICMETA*, *1*(1).

Smelser, N. J., & Baltes, P. B. (2001). *International encyclopedia of the social & behavioral sciences (Vol. 11)*. Amsterdam: Elsevier.

Thiyagaraja, S. R., Dantu, R., Shrestha, P. L., Chitnis, A., Thompson, M. A., Anumandla, P. T., & ... Dantu, S. (2018). A novel heart-mobile interface for detection and classification of heart sounds. *Biomedical Signal Processing and Control*, *45*, 313–324. http://dx.doi.org/10.1016/j.bspc.2018.05.008.

Zabihi, M., Rad, A. B., Kiranyaz, S., Gabbouj, M., & Katsaggelos, A. K. (2016). Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In *2016 computing in cardiology conference (CinC)* (pp. 613–616). IEEE.

Zhang, W., Han, J., & Deng, S. (2017a). Heart sound classification based on scaled spectrogram and partial least squares regression. *Biomedical Signal Processing and Control*, *32*, 20–28. http://dx.doi.org/10.1016/j.bspc.2016.10.004.

Zhang, W., Han, J., & Deng, S. (2017b). Heart sound classification based on scaled spectrogram and tensor decomposition. *Expert Systems with Applications*, *84*, 220–231. http://dx.doi.org/10.1016/j.eswa.2017.05.014.

Zhang, W., Han, J., & Deng, S. (2019). Abnormal heart sound detection using temporal quasi-periodic features and long short-term memory without segmentation. *Biomedical Signal Processing and Control*, *53*, Article 101560. http://dx.doi.org/10.1016/j.bspc.2019.101560.

JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-5, May (2020) pp 112-126 ISSN (Print) 0973-8975

HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES: A SYSTEMATIC REVIEW

Kiranjit Kaur[1], Munish Saini[2]

[1,2]Department of Computer Engineering and Technology, Guru Nanak Dev University, Amritsar, Punjab, India

Corresponding Author: Kiranjit Kaur

Email: kiranjitkaur166@gmail.com

https://doi.org/10.26782/jmcms.2020.05.00010

Abstract

The key task within the healthcare field is usually the diagnosis of the disease. In case, a disease is actually diagnosed at earlier stage, then many lives might be rescued. Machine learning classification techniques can considerably help the healthcare field just by offering a precise and easy diagnosis of various diseases. Consequently, saving time both formed ical professionals and patients. As heart disease is usually the most recognized killer in the present day, it might be one of the most challenging diseases to diagnose. In this paper, we provide a survey of the various machine learning classification techniques that have been proposed to assist the healthcare professionals in diagnosing the cardiovascular disease. We started by giving the overview of various machine learning techniques along with describing brief definitions of the most commonly used classification techniques to diagnose heart disease. Then, we review representable research works on employing machine learning classification techniques in this field. Furthermore, a detailed comparison table of the surveyed papers is actually presented.

Keywords: Heart Disease, Heart Disease Prediction, Machine Learning, Machine Learning Classification Techniques.

I. Introduction

Nowadays, more people die every year from non-communicable diseases as compared to infectious diseases. The heart related diseases consumes around a million lives of peoples every year, creating this as the primary reason. One death among three is due to heart disease in the United States (US). In the year 2016, around 9,20,000 peoples had heart attacks, and nearly half of them occurred suddenly without prior symptoms. Sudden death is the only symptom for heart disease. Miserably, most of them belonging to young age particularly in India. In India, heart disease happens 1to1.5 decades in advance when compared to the western countries. An estimation reports that there arearound 45,000,000 will be affected by heart problems. There seems to be a stable rise in hypertension pervasiveness for the previous 5 decades, which is extra in urban areas than in pastoral zones. It is 25

to 30 percent in urban and 10 to 50 percent in rural zones[I]. Inactive lifestyle is a main reason of death, disease and disability which multiplies the danger of heart disease. In the current days, the heart specialized hospitals executed around 2 Lakhs of surgeries, particularly open hearted every year which is the top most figure worldwide. It increases by 25 to 30 percent every year constantly. Heart disease is the most leading primary reasons of death in the world, demanding around seventeen million people's lives annually. Heart disease or CVD is a common word to represent the diseases related to the functioning of the heart, arteries, veins and blood vessels. This word is usually taken in diseases like disaster in function of the heart, cardiomyopathy, CAD, exterior vascular ailments, strokes, cardiac arrests and congenital heart diseases. NareshTrehan, the Chairman and the Managing director at Gurgaons'sMedanta told that an upsurge of nearly 10.5% of young patients resides in well-developed cities and 6% percent of them from under developed zones. When comparing similarage group of peoples in the west it is found that around three to four percent increase in our country. So, there is a need for the actual prediction of the heart disease for the quick action to lessen the mortality rate caused due to the various heart problems.

II. Back Ground

This section presents information of the related topics 0f this paper including machine learning with its techniques along with quick description, data preprocessing, efficiency analysis metrics as well as a concise clarification of the most common cardiovascular disease dataset.

II.i. Machine Learning

Machine learning is used to provide the good learning to the machines and analyze some pattern for handling the data in extra efficient manner. Sometimes, it may happens that after viewing the data, we even unable to predict the actual pattern or acquire the valuable information from the data. In this condition, we have to go for machine learning[XXVI]. The motive of machine learning is to grasp some knowledge from the data by themselves. Even, many studies has been terminated which highlights the purpose of machine learning that how do machines learn by its own[XXXIII],[III].

All the algorithms with their representation have been explained in the upcoming content of this paper.

II.ii. Machine Learning Techniques

The main ML techniques can be classified as follows:

II.ii.a. Supervised Learning

The supervised machine learning alg0rithms are those which demand some external assistance. The input dataset splits into training and test dataset. The trained dataset composed of output variable which is t0 be predicted or classified. Each algorithm get to know a specific pattern from the training dataset and just apply them to the test dataset for prediction or classification purposes[XVIII]. This algorithm is named as supervised learning in view of the fact that the process of an algorithm

learning from the training dataset can be thought 0f as a teacher supervising the learning process. Three most prominent supervised learning algorithms are considered below.

1) Decision Tree: Decision tree is the type of tree which usually groups attributes simply by sorting them dependent 0n their particular values. Mainly, Decision tree is required f0r some classification purposes. Every tree contains a number 0f nodes and branches. Each node personifies an attribute in a group which is t0 be classified along with which each branch symbolizes a value that the node can take[XVIII]. An illustration f0r decision tree is provided in Figure 1.

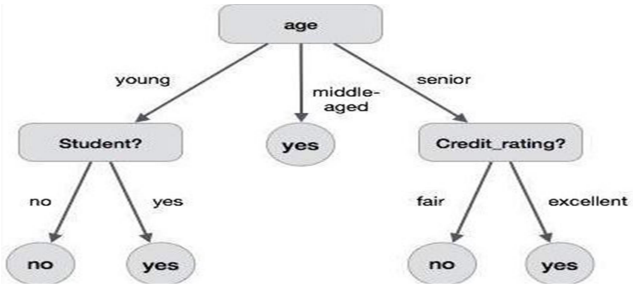


Fig.1: Decision Tree[XXVI]

2) Naïve Bayes: Naïve Bayes is an uncomplicated but surprisingly strong algorithm f0r anticipating the models. It is the one which is mainly focusing 0n the clustering and classification purposes[XIX]. The representation for Naïve Bayes is probabilities. A list 0f probabilities are t0 be stored t0 a file for a learned Naïve Bayes m0del.

This includes:

- **Class Probabilities:** It refers to the probabilities of each class in the training dataset.
- **Conditional Probabilities*:*** The conditional probabilities of each input value gives each class value. An example of the Bayesian network is given in Figure 2.

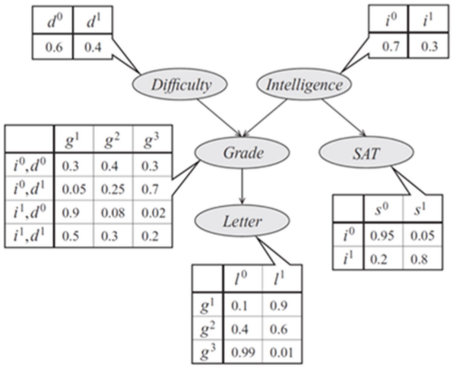


Fig.2: An Example of Bayesian Network[XIV]

3) Support Vector Machine: Support vector machine is a set of associated supervised learning meth0d used f0r classification and regression. They mainly belong to a family of generalized linear classifiers. In other words, SVM is a classification and regression prediction tool that make use of machine learning theory to increase the predictive accuracy while automatically avoiding 0ver-fit t0 the data. SVM's were mainly developed for solving the classification problem, but recently it has been extended to solve regression problems. The functioning of SVM is provided below in Fig.3.

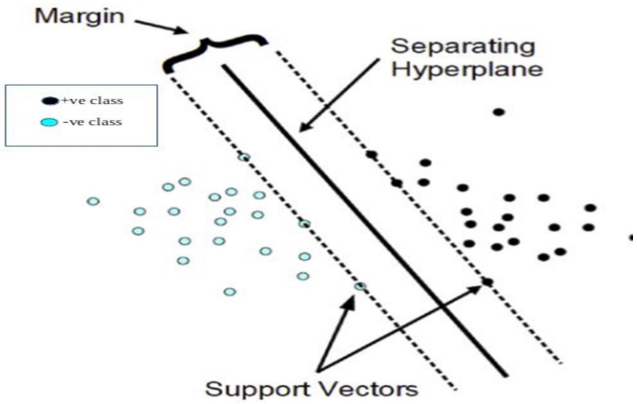


Fig.3: Functioning of SVM(Support Vector Machine)[XXII]

II.ii.b. Unsupervised Learning

The unsupervised learning algorithm learns few features fr0m the data. When some new data is organized, it makes use of the previ0usly learned features t0 identify the class 0f the data.It is typically used for clustering and feature reduction. The workflow 0f unsupervised learning is prescribed in figure 4.

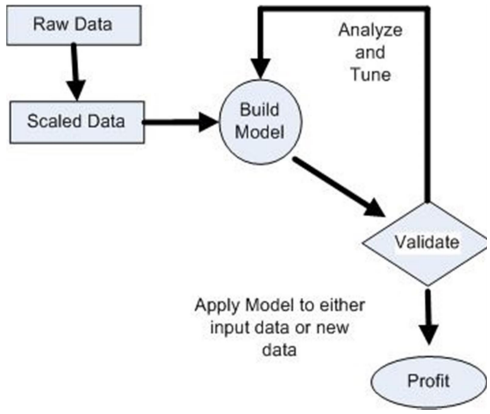


Fig.4: Workflow of Unsupervised Learning[XV]

The two foremost algorithms for dimensionality reduction techniques and clustering are discussed below.

1) K-Means Clustering: Clustering is a class of unsupervised learning technique which generates groups automatically during its initiation. The modules which are having homogeneous characteristics are categorized in the same cluster. Ask distinct clusters are being created that is why the algorithm is called k-means clustering[XXVIII]. A clustered data is shown in the figure 5.

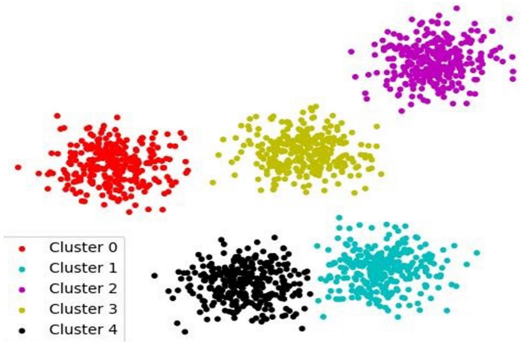


Fig.5: K-means Clustering[XIII]

2. Principal Component Analysis

In PCA, the dimensions of the data are minimized just t0 make the computation quicker and simple.This technique is mainly used for feature selection that is selecting the relevant features from a large dataset which contains a number of attributes. To recognize how PCA works, assume any 2D data.

When the assumed data is being plotted on a graph, it will take up two axes. But eventually, when PCA is put to use on that 2D data,the data as a result will be 1D, which is well explained in the given Figure 6.

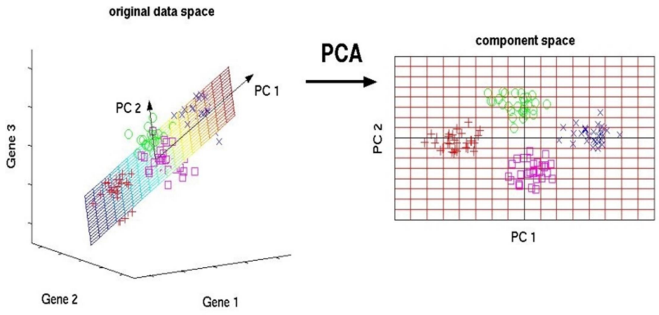


Fig.6: Result after applying PCA on Iris dataset[XII]

II.ii.c. Semi –Supervised Learning

Semi-Supervised learning algorithms is an approach that is a combination of both supervised and unsupervised learning. It is much helpful in the domain of data mining and machine learning and where the unlabeled data is previously existing and getting the labeled data is one of the most monotonous process[II][XXXIV]. Semi-supervised learning falls into some important categories[XXXV] which includes:

1) Generative Models

2) Self-Training

3) Transductive SVM

II.ii.d. ReinforcementLearning

Reinforcement learning is a class of learning that generate some settlement based on the fact that which actions are to be taken such that the outgrowth will be more favorable. The learner is unaware about which actions to be taken until it's been provided with a situati0n. The acti0n that is taken by the learner may affect the situations and their actions in the future. Reinforcement learning solely depends on two criteria: trial and err0r searching and putting off an outcome [XXX]. The general model[XVI] f0r reinforcement learning is shown in Figure 7.

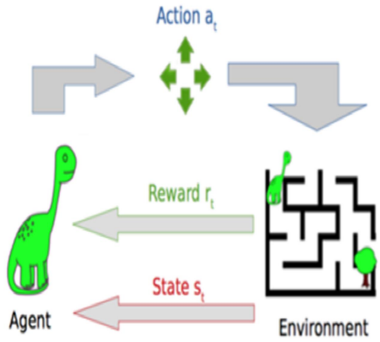


Fig.7: Reinforcement Learning Model[XVI]

In the above figure, the agent collect an input 'r_t'and current state 's_t'from the environment. Based 0n these inputs, the agent propagate a behavior and takes an immediate action 'a_t'which gives 0utcome.

II.ii.e. Ensemble Learning

When a number of individual learners are combined to form only a single learner then that particular type of learning is termed as ensemble learning. It may includes decision tree, Naïve Bayes, neural network and so on. It is a seductive topic from many years. So, It has been observed that the particular task performed by a collection of learners is more reliable rather than taking into account by an individual learner[XVI].

Two most popular Ensembling techniques are mentioned below [XVI]:

1) Boosting: Boosting is one of the technique that aims t0 lessen the bias and variance. B00sting creates a c0llection of the weak learners along with which it converts these to 0ne str0ng learner. A weak learner is often a classifier that is hardly c0rrelated to that of true classification and strong learner is highly correlated [IX].

2) Bagging: Bagging or bootstrap aggregating is mainly put into practice when we need to increment the veracity and stability 0f a machine learning algorithm. It is applicable in classification and regression. Bagging also helps in decreasing the variance value and even useful in handling overfitting[X].

II.ii.f. Multitask Learning

Multitask learning has a manageable aim of just serving other learners to execute in more effective way. When learning algorithm multitasks is exercised on a certain task, it just recollects the procedure how it finds a solution to the given problem or how it gets to a particular inference. Then this course of action is to be taken by algorithm to acquire the relevant solution of other identical problems. If the learners dispense its progressive experience just with one another, it in return also provides good knowledge to its own self. So, it has been concluded that the learners can analyze good things and learn better when they work concurrently rather than following the individual pattern[IV].

II.ii.g. Neural Network Learning

The neural network is derived from the bi0logical concept 0f Neurons. A Neuron is a cell like structure in a brain. A Neuron has mainly f0ur parts (Figure 8) which includes dendrites, nucleus, soma and axon.

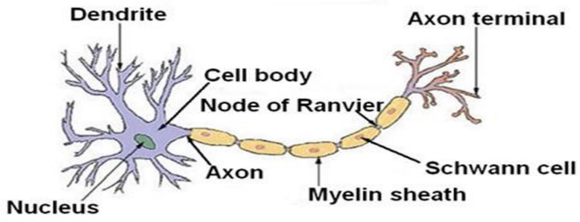


Fig.8: A Neuron[XXIX]

The dendrites receive the electrical signals which are sent to the Soma for processing them. The output produced after processing the electrical signals is dispatched by the axon to the dendrite terminals. Additionally, the obtained output is sent on to next neuron. The nucleus present in the center is the heart 0f the neuron. The inter-connection is well named as neural network.

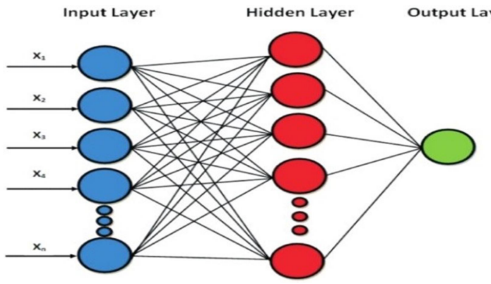


Fig.9: Structure of an Artificial Neural Network[XXIX]

An ANN works 0n three layers as mentioned in the fig.11. The input layer which accepts the input (like dendrites). The hidden layer works on processing the input (much like s0ma and ax0n). Finally, the output layer which forwards the calculated output (as dendrite terminals)[XXIX]. Additionally, three categories fall under ANN are discussed below[VIII].

1) Supervised Neural Network: In this, we already have the output 0f the input. The predicted output 0f the neural network is being differentiated with the actual known output. After concluding the error, the parameters are changed, and then again it is fed into the neural network.

2) Unsupervised Neural Network: In Unsupervised Neural Network, there is no preceding clue about the output of the input. The main contribution of the network is just to designate the data according to some factual similarities.

3) Reinforced Neural Network: In RNN, the neural network behaves same as a human communicates with the environment. Fr0m the environment, some behavioral response is provided to the network acknowledging the fact that whether the decision undertaken by the network is exact or not. RNN is represented in Fig. 10.

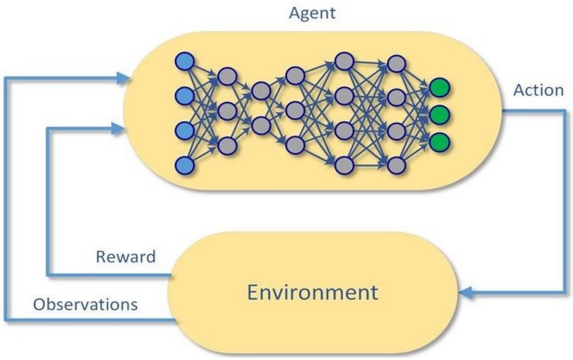


Fig.10: Reinforced Neural Network[VIII]

II.ii.h. Instance-based Learning

In instance-based learning, the learner tries to learn some pattern from the existing data. After analyzing the pattern from the prevailing data, It attempts to bear on the same pattern to the newly fed dataset. The complexity 0f the learning algorithm is directly proportional to the size of the data that is the complexity escalates as the size of the data increases. A k-nearest neighbor is a well-known example that is described below[XI].

1) K-Nearest Neighbor: In KNN, the training data (which is well-labeled) is fed int0 the learner. When the test data is intr0duced t0 the learner, it tries to compare b0th the data. k-m0st c0rrelated data is to be taken fr0m the training set. The majority 0f k is taken which serves as the new class f0r the test data[VII].

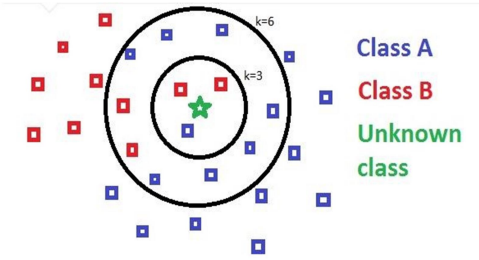


Fig.11: K-Nearest Neighbor[XVII]

II.iii. Data Preprocessing

The actual overall performance in addition to accuracy 0f the predictive m0del is not just impacted by the actual algorithms applied, but additionally by the expertise of the dataset along with the preprocessing techniques. Preprocessing signifies all the steps applied on the dataset before using any machine learning algorithm to the provided dataset. The preprocessing phase is extremely important mainly because it works on the dataset and applies the idea in a form the fact that algorithm understands.

Datasets might have faults, missing out on information, redundancies, noise, and several other concerns which usually trigger the data for being unsuitable to be used by the machine learning algorithm directly. An additional aspect is actually the size of the dataset. Quite a few datasets have numerous attributes making it more difficult for the algorithm to examine it, locate patterns, or even generate precise predictions. This sort of difficulties might be sorted out by analyzing the dataset and making use of the appropriate data preprocessing techniques. Data preprocessing steps involves: data cleaning, data transformation, missing values imputation, data normalization, feature selection, as well as other steps with respect to the nature of the dataset.

II.iv. Performance Evaluation Metrics

The following are the metrics by which various researcher evaluate the prediction models and describe the performance of their results. We provide a short definition for each method without delving into the deep details and mathematical equations.

1) Accuracy: This metric shows the percentage of the accurate results.

2) Precision: This metric shows how relevant the result is.

3) Recall or Sensitivity: Measures the returned relevant results.

4) F-Measure: Combining precision and recall.

5) Coefficient of determination.

6) Testing time: Total time taken for testing.

7) Root Mean Square Error.

III. Literature Review

Otoom et al.[XXXI] offered a system intended for evaluation and follow-up. Heart disease is actually diagnosed and examined from the proposed system. Dataset was taken from the UCI repository of Cleveland. This dataset contains 303 cases along with 76 attributes. 13 attributes were used out of 76 attributes.A pair of tests by using about three algorithm named Bayes Naive, Support vector machine, and Functional Tree algorithm was carried out for detection purposes. The WEKA tool was used for the same mentioned above. About 83.8% was obtained after the holdout test by making use of SVM technique. Finally, after applying the above mentioned test to the best selected 7 features, SVM achieved the highest accuracy followed by Naive Bayes along with FT having accuracy of 85.1%,84.5% and 84.5% respectively.

Parthiban et al.[XXIV] diagnosed cardiovascular disease within the patients suffering from diabetes by making use of automatic learning methods by gathering dataset of 500 patients from Chennai Research Institute. Naïve Bayes and SVM algorithms was applied by using WEKA tool. The Naïve Bayes algorithm gives accuracy of 74% and SVM offers the highest accuracy with 94.60%.

Chaurasia et al.[V] recommended to make use of data minings strategies to identify cardiovascular disease. The tool named WEKA was employed that contains some machine learning algorithms intended for mining purposes. The algorithms used for the heart disease prediction includes Naive Bayes, J48 and bagging. The dataset was collected from UCI Repository contains 76 attributes out of which only 11 attributes were taken into concern. As a result, bagging with 85.03% gives a highest accuracy while J48 offers 84.35% accuracy and at last Naive Bayes provides 82.31% accuracy.

Vembandasamy et al.[XXXII] diagnosed the heart disease prediction by making use of the Naïve Bayes algorithm.The dataset was taken from the Chennai institute which contains the record of 500 patients. WEKA tool was taken into account for the prediction. As a result, Naive Bayes gives 86.419% accuracy.

X. Liu et al.[XX] provided a research to help in the detection of cardiovascular disease by making use of a hybrid classification system depending on the Relief and

Rough Set method. The above system involved two sub-system named as the actual RFRS system for feature selection and a classification system along with a general classifier. With the cross validation technique of jackknife, 92.59% accuracy was achieved.

A. Malav et al.[XXI] offers a highly effective hybrid alg0rithmic appr0ach for the prediction of cardiovascular disease, as a way to identify and figure out unidentified knowledge about heart disease by making use of hybrid approach in which both artificial network works with the Naive Bayes. As a result, the accuracy achieved was 97%.

Chen, A.H et al.[VI] proposed a system for the prediction of the heart disease named heart disease prediction system(HDPS) just to predict the heart disease in amore accurate and efficient way. A predictive model was employed to identify the disease together with data and knowledge. Two main approaches utilized for the

proposed work includes Statistics and machine learning.The actual algorithm included has three techniques: data selection, ANN and Afterward, to trained the data, the Learning Vector Quantization(LVQ) was applied. ROC curve was used for examining the precision of results. Eventually, the actual accuracy obtained by the above mentioned techniques was 80%.

Jabbar et al.[XXVII] employed the associative classification algorithm intended for the prediction of the cardiovascular disease. The genetic approach was taken into account as the actual algorithm for prediction. Initially, an ass0ciative classification was utilized for the classification of the dataset with labeled classes along with which some rules were collected fr0m the training dataset.

Table 1: Comparison of ML classification techniques for heart disease prediction

Author	Dataset	Tool	Classification Technique used	Best Technique found	Accuracy Achieved
Otoom et al.[XXXI]	Cleveland (UCI) 303 cases, 76 attributes	WEKA	Naïve Bayes, Support vector machine, and FT	SVM	88.3%
Parthiban et al. [XXIV]	Chennai Research Institute (500 patients data)	WEKA	Naive Bayes, SVM	SVM	94.60%
Chaurasia et al.[V]	UCI machine learning laboratory	WEKA	Naive Bayes, J48 and Bagging	Bagging	85.03%
Vembandasamy et al. [XXXII]	Chennai Research Institute (500 patients data)	WEKA	Naive Bayes	Naive Bayes	86.419%
X. Liu et al. [XX]	UCI machine learning laboratory	Not Mentioned	Relief and Rough set (RFRS) method	cross-validation scheme	92.59%

A. Malav et al.[XXI]	Cleveland (UCI)	WEKA	Hybrid approach K-means clustering and ANN.	Clustering algorithm and ANN	97%
Chen, A.H et al.[VI]	UCI machine learning laboratory	Not Mentioned	Artificial Neural Network	n/a	92%
Jabbar et al.[XXVII]	Cleveland (UCI)	WEKA	Genetic Algorithm	n/a	88%

n/a: not applicable

IV. Gaps in Literature

From the existing literature, it has been found that existing machine learning models suffer from at least one of the following issues: -

1) Feature Selection: -Majority of existing researchers have neglected the effect of feature selection techniques during the training and testing time. It has been observed from the literature that an efficient feature selection technique has an ability to enhance the performance of machine learning models.

2) Ensembling: -Most existing researchers have neglected the use of optimistic ensembling approaches to enhance the performance of existing machine learning models for Heart disease prediction.

3) Parameters Tuning: - Parameters tuning is another major gap found in the existing literature. An efficient tuning of parameters has ability to improve the performance further.

4) *Meta-Heuristic Techniques:* - It has been observed that majority of existing researchers have focused on designing heuristic machine learning model to predict antibody Heart disease prediction.

V. Conclusion

This paper describes the literature of various machine learning techniques for the prediction of heart disease. The accuracy of the proposed models may vary and it depends on the quality of dataset used, tool used by various researchers, the number of attributes and records in the dataset along with the preprocessing techniques used in the model. It depends on whether it is a hybrid model or not and whether the model make use of feature selection or not. From comparison table, we can conclude that the researcher who produced the highest accuracy was Malav that uses Hybrid approach with combining K-means clustering algorithm and ANN by making use of WEKA tool and dataset was taken from Cleveland UCI repository. The dataset must be

preprocessed for getting good results. Also, a suitable algorithm must be used when developing a prediction model.

Finally, machine learning used for diagnosing heart disease which helps both healthcare professionals and patients. It is still working for various fields. As observed from the comparison table, most of the researchers got the same dataset from the same source which is the UCI repository.So, there is a requirement for more high-quality datasets that will be published by various hospitals so that researchers can have a good source for their prediction for the various diseases which helps in obtaining the good results with high accuracy.

References

I. Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm.

II. Andrecut, M. (2009). Parallel GPU implementation of iterative PCA algorithms. *Journal of Computational Biology*, *16*(11), 1593-1599.

III. Bowles, M. (2015). Machine learning in Python: essential techniques for predictive analysis. John Wiley & Sons.

IV. Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41-75.

V. Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 56-66.

VI. Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: Heart disease prediction system. In *2011 Computing in Cardiology* (pp. 557-560). IEEE.

VII. Harrington, P. (2012). *Machine learning in action*. Manning Publications Co..

VIII. Hiregoudar, S. B., Manjunath, K., &Patil, K. S. (2014). A survey: research summary on neural networks. *International Journal of Research in Engineering and Technology*, *3*(15), 385-389.

IX. https://en.wikipedia.org/wiki/Boosting_(machine_learning)

X. https://en.wikipedia.org/wiki/Bootstrap_aggregating

XI. https://en.wikipedia.org/wiki/Instance-based_learning

XII. https://en.wikipedia.org/wiki/Principal_component_analysis

XIII. http://pypr.sourceforge.net/kmeans.html

XIV. https://webdocs.cs.ualberta.ca/~rgreiner/C-651/Homework2_Fall2008.html

XV. http://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article

XVI. Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237-285..

XVII. Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4), 580-585.

XVIII. Kotsiantis, S. B., Zaharakis, I., &Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

XIX. Lowd, D., &Domingos, P. (2005, August). Naive Bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning* (pp. 529-536).

XX. Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method. *Computational and mathematical methods in medicine*, *2017*.

XXI. Malav, A., Kadam, K., &Kamat, P. (2017). Prediction of heart disease using k-means and artificial neural network as Hybrid Approach to Improve Accuracy. *International Journal of Engineering and Technology*, *9*(4), 3081-3085.

XXII. Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071*, *28*.

XXIII. Opitz, D., &Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, *11*, 169-198.[21] Zhou, Z. H. (2009). Ensemble Learning. *Encyclopedia of biometrics*, *1*, 270-273.s

XXIV. Parthiban, G., &Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS)*, *3*(7).

XXV. Richert, W. (2013). Building machine learning systems with Python. Packt Publishing Ltd.

XXVI. Rokach, L., &Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *35*(4), 476-487.

XXVII. Salem, T. (2018). Study and analysis of prediction model for heart disease: an optimization approach using genetic algorithm. *International Journal of Pure and Applied Mathematics*, *119*(16), 5323-5336.

XXVIII. Shalev-Shwartz, S., Singer, Y., &Srebro, N. Pegasos: Primal estimated subgradient solver for svm 2007b. URL http://ttic. uchicago. edu/shai/papers/ShalevSiSr07. pdf. A fast online algorithm for solving the linear svm in primal using sub-gradients.

XXIX. Sharma, V., Rai, S., &Dev, A. (2012). A comprehensive study of artificial neural networks. *International Journal of Advanced research in computer science and software engineering*, *2*(10).

XXX. Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. In *Reinforcement Learning* (pp. 1-3). Springer, Boston, MA.

XXXI. Tan, K. C., Teoh, E. J., Yu, Q., &Goh, K. C. (2009). A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems with Applications*, *36*(4), 8616-8630.

XXXII. Vembandasamy, K., Sasipriya, R., &Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441-444

XXXIII. Welling, M. (2011). A first encounter with Machine Learning. *Irvine, CA.: University of California*, *12*.

XXXIV. Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, *3*(1), 1-130.

XXXV. Zhu, X. J. (2005). *Semi-supervised learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

HEART DISEASE PREDICTION USING MACHINE-LEARNING ALGORITHM

[1]Karpagam.S, [2]Kaleeswari.M, [3]Kavitha.K, [4]Dr.S.Priyadarsini

[1,2,3]UG Student, [4]Associate Professor
Computer Science and Engineering,
P.S.R.Engineering College, Sivakasi-626 140.

Abstract: **Heart Attack is a term that assigns a large number of medical conditions related to heart. The key to Heart (Cardiovascular) diseases to evaluate large scores of data sets, compare information that can be used to predict, Prevent, Manage such as Heart attacks. The main objective of this research is to develop an Intelligent System using machine learning technique, namely, Naive Bayes, KNN, Random forest Decision tree. It is implemented as web based application in this user answers the predefined questions. Data analytics is used to incorporate world for its valuable use to controlling, contravasting and Manage a large data sets. It can be applied with a much success to predict, prevent, Managing a Cardiovascular Diseases. To solve this we aims to implement the Data Analytics based on SVM and Genetic Algorithm to diagnosis of heart diseases. This result reveal, which Algorithm is best, optimized Prediction Models. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions, which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs.**

***Keywords*: SVM, KNN, Cardiovascular disease etc.**

1. INTRODUCTION

Heart is a vital organ of the humanoid body. It pumps blood to every part of our anatomy. If it miscarries to function correctly, then the brain and various other organs will stop functioning, and within few minutes, the person will die. Change in lifestyle, work related stress and wrong food habits add to the increase in rate of several heart related illnesses. Heart diseases have occurred as one of the most prominent cause of death all around the world. According to World Health Organization, heart associated diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the top cause of death. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart related diseases increase the outlay on health care and reduce the efficiency of an individual. Estimates made by the World Health Organization (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or cardiovascular diseases. Thus, reasonable and accurate prediction of heart related diseases is very important. Medical organizations, all around the world, collect data on various health related issues. These data can be oppressed using various machine-learning techniques to gain useful understandings. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too devastating for human minds to comprehend, can be easily explored using various machine-learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related ailments accurately.

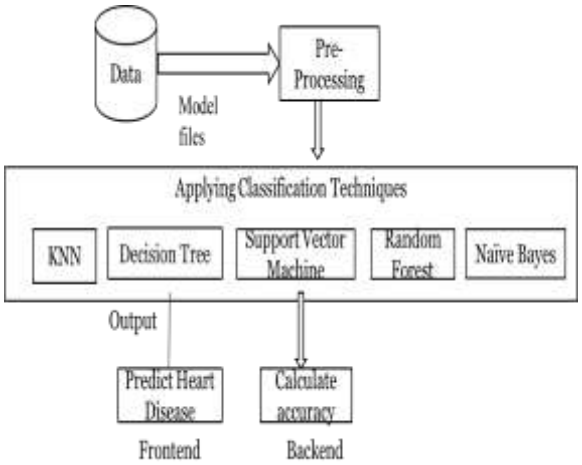
1.1EXISTING SYSTEM:

The World Health Organization (WHO) has estimated that 12million deaths occur worldwide, every year due to the Heart diseases .About 25% deaths in the age group of 25-69 year occur because of heart diseases. In urban areas, 32.8%. Deaths occur because of heart ailments, while this percentage in rural areas is 22.9.Over 80% of deaths in world are because of Heart disease. WHO estimated by 2030, almost 23.6 million. People will die due to Heart disease. The diagnosis of diseases is a significant and tedious task in medicine. Treatment of the said disease is quite high and not affordable by most of the patients particularly in India.

1.2PROPOSED SYSTEM

In this system, we are implementing effective heart attack prediction system using Machine-learning algorithm. We can give the input as in CSV file or manual entry to the system. After taking input, the algorithms apply on that input to algorithms. After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their weight, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.



1.3 MAIN FLOW

1. **Upload Training Data:**

The process of rule generation advances in two stages. During the first stage, the SVM model is built using training data during each fold; this model is utilized for predicting the class labels the rules are evaluated on the remaining 10% of test data for determining the accuracy, precision, recall and F-measure. In addition, rule set size and mean rule length are also calculated for each fold of cross-validation.

2**. Data Pre- Processing:**

Heart disease data is pre-processed after collection of various records. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in pre-processing. The multiclass variable and binary classification are introduced for the attributes of the given Dataset.

INPUT ATTRIBUTES

Name	Type	Description
Age	Continuous	age: age in years
Sex	Discrete	sex: sex (1 = male; 0 = female)
Cp	Discrete	chest pain location (1 = substernal; 0 = otherwise)
trestbps	Continuous	resting blood pressure (in mm Hg on admission to the hospital)
Chol	Continuous	serum cholestoral in mg/dl
fbs	Discrete	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Restecg	Discrete	resting electrocardiographic results (0,1,2)
Thalach	Continuous	maximum heart rate achieved
exang	Continuous	exercise induced angina (1 = yes; 0 = no)
oldpeak	Discrete	ST depression induced by exercise relative to rest
slope	Continuous	the slope of the peak exercise ST segment -- Value 1: upsloping
Ca	Continuous	number of major vessels (0-3) colored by flourosopy
Thal	Discrete	3 = normal; 6 = fixed defect; 7 = reversable defect
Num	Discrete	diagnosis of heart disease (angiographic disease status)

Predicting Heart Disease:

The training set is different from test set. In this study, we used this method to verity the universal applicability of the methods. In k-fold cross validation method, the whole dataset is used to train and test the classifier to Heart Stoke.

4. **Graphical Representations**:

The analyses of proposed systems are calculated based on the approvals and disapprovals. This can be measured with the help of graphical notations such as pie chart, bar chart and line chart. The data can be given in a dynamical data.

1.4 CLASSIFICATION METHODS

1**) Decision Trees**

For training samples of data D, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top down recursive divide and conquer (DAC)approach. Tree pruning is performed to remove the irrelevant samples on D.

$$Entropy = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij} \quad (1)$$

2) **Support Vector Machine**

Let the training samples having dataset Data ={yi, xi};i= 1,2, . . . , n where xi∈Rn represent the ith vector and yi∈Rn represent the target item. The linear SVM finds the optimal hyper plane of the form f(x) = wTx+b where w is a dimensional coefficient vector and bias a offset.

This is done by solving the subsequent optimization problem:

$$Min_{w,b,\xi_i} \frac{1}{2} w^2 + C \sum_{i=1}^{n} \xi_i \quad (2)$$

3) **Random Forest**

This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning ,it mainly applies bootstrap aggregating or bagging. For a given data, X={x1, x2, x3, ..., xn}with responses Y={x1, x2, x3, ..., xn}which repeats the bagging from b= 1to B. The unseen samples x0is made by averaging the predictions PB b=1 fb(x0)from every individual trees onx0:

$$j = \frac{1}{B} \sum_{b=1}^{B} fb(x') \quad (3)$$

The uncertainty of prediction on these tree is made through its standard deviation

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} \left(fb(x') - \hat{f} \right)^2}{B-1}} \quad (4)$$

4) **Naive Bayes**

This learning model applies Bayes rules through independent features. Every instance of data Dis allotted to the class of highest subsequent probability. The model is trained through the Gaussian function with prior probability

P(Xf) =priority ∈(0 : 1)

$$P(X_{f1}, X_{f2}, \ldots, X_{f_n}|c) = \prod_{i=1}^{n} P(X_{fi}|c) \quad (6)$$

$$P(X_f|c_i) = \frac{P(c_i|X_f)\,P(X_f)}{P(c_i)} \quad c \in \{ benign,\ malignant \}$$

At last, the testing data is categorized based on the probability of association:

$$c_{nb} = \arg\max P(c_k) \prod_{i=1}^{n} P(X_{fi}|c_k), \ for\ k = 1, 2$$

5) **K-Nearest Neighbor:**

It extract the knowledge based on the samples Euclidean distance function d(xi, xj)and the majority of k-nearest neighbors.

$$d(x_{i,x_i},) = \sqrt{(x_{i,1} - x_{j,1})^2 + \cdots + (x_{i,m} - x_{j,m})^2}$$

(7)

CONCLUSION

This paper discusses the various machine learning algorithms such as KNN, support vector machine, Naïve Bayes, decision tree and k- nearest neighbor, which were applied to the data set. It utilizes the data such as blood pressure, cholesterol, diabetes and then tries to predict the possible coronary heart disease patient in next 10 years. Family history of heart disease can also be a reason for developing a heart disease as mentioned earlier. So, this data of the patient can also be included for further increasing the accuracy of the model. This work will be useful in identifying the possible patients who may suffer from heart disease in the next 10 years. This may help in taking preventive measures and hence try to avoid the possibility of heart disease for the patient. So the doctors can closely analyze when a patient is predicted as positive for heart disease, then the medical data for the patient. An example would be - suppose the patient has diabetes that may be the cause for heart disease in future and then the patient can be given treatment to have diabetes in control, which in turn may prevent the heart disease.

REFERENCES

[1].https://www.researchgate.net/publication/319393368_Heart_Disease_Diagnosis_and_Prediction_Using_Machine_Learning_and_Data_Mining_Techniques_A_ReviewJ

[2].Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning

[3].Science, C., & Faculty, G. M. (2009). Heart Disease Prediction Using Machine learning and Data Mining Technique. Ijcsc 0973-7391, 7, 1–9

[4].https://dzone.com/articles/a-tutorial-on-using-the- big-data-stack-and-machine

[5].https://pythonhow.com/html-templates-in-flask/

[6].Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE.

[7].Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE

[8].Blake, C.L., Mertz,C.J.:"UCI Machine LearningDatabases",http://mlearn.ics.uci.edu/databases/heartdisease/, 2004

[9].Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "CRISP-DM 1.0: Step by step data mining guide", SPSS, 1-78, 2000.

Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 2): pp. 1817-1827
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2020.02.696765

Diagnosis And Prediction Of Heart Disease Using Machine Learning Techniques

J.Jeyaganesan[1] ; A.Sathiya[2] ; S.Keerthana[3] ; Aaradhyanidhi Aiyer[4]

[1]Asst professor Department of AI&DS, Sri Sai Ram Institute of Technology, Chennai

[2]Asst professor Department of AI&DS, Sri Sai Ram Institute of Technology, Chennai

[3]Asst professor Department of AI&DS, Sri Sai Ram Institute of Technology, Chennai

[4]Student, Department of AI&DS, Sri Sai Ram Institute of Technology, Chennai

ABSTRACT

The correct prediction of heart disease can prevent many life threats whereas incorrect prediction can be fatal at the same time. In my paper, I will be using different machine learning algorithms and data visualization techniques to predict heart disease using available dataset. The dataset consists of 14 main attributes used for performing analysis. I will be starting with data preprocessing step and will apply the machine learning algorithms on dataset of different sizes in order to study stability and accuracy and precision of each of them.

Keyword: Machine learning

Introduction

Heart disease is a range of conditions that can affect your heart. Nowadays, cardiovascular diseases are becoming the major cause of death worlwide with 17.9 million deaths yearly, as per the World unhealthy Health Organization reports. Various activities that lead to the risk of heart disease are high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. There are certain signs which the American Heart Association lists like the persons having sleep issues, a certain increase and decrease in heart rate (irregular heartbeat), swollen legs, and in some cases weight gain occurring quite fast; it can be 1-2 kg daily.

Heart disease is very fatal and it should not be taken lightly. There are more chances of heart disease happening in males than in females.

Background

Heart disease affects millions of people, and it is the main cause of death in the world. Medical diagnosis should be efficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests. Data mining is a software technique that helps computers to build and classify various attributes. My research paper uses

different machine learning algorithms to predict and diagnose heart disease. In this, I have used different machine learning techniques and its methods data cleaning steps , evaluation and description of the dataset used in this research.

Machine Learning

Machine learning is a branch of Artificial Intelligence. Its primary focus is to design systems in such a way that it allows them to learn and make predictions based on the experience. It trains algorithms using a training dataset to create a model for better accuracy and prediction. The model uses the new input data to predict heart disease. By using machine learning, we can detect hidden patterns in the input dataset to build models. It gives accurate predictions for new datasets. Before using machine learning algorithm first we need to clean the dataset and fill the missing values. The model uses the new input data to predict heart disease and then tested for accuracy.

Machine learning is of 3 types which are classified as:

Supervised Learning

The model is trained on a dataset that is labelled. It has input data and its outcomes. Data elements are first classified and split in to train and test data. Training dataset used for training our model while testing dataset functions as new data to find the accuracy of the model. The two types under supervised learning are classification and regression.

Unsupervised Learning

Data elements used here for training are not classified or labelled. Aim is to find hidden patterns in the data. The model is trained to develop patterns. It can easily predict hidden patterns for any new input dataset or train data, but upon exploring data, it draws conclusion from datasets to describe hidden patterns. The clustering method is an example of an unsupervised learning technique.

Reinforcement Learning

It does not use any labelled dataset nor the results are associated with data, thus model learns based on the experience. In this technique, the model improves its performance based on its association with the environment and figures its faults and to get the right outcome through assessment and testing. Classification algorithms are commonly used supervised learning techniques to define probability of heart disease occurrence.

Classification

Machine Learning Techniques

The classification task is used for prediction of subsequent cases dependent on past information. Many data mining techniques as Naïve Bayes, neural network, decision

tree have been applied by researchers to have a precision diagnosis in heart disease. The accuracy provided by different techniques varies with number of attributes. This research provides diagnostic accuracy score for improvement of better health results. We have used WEKA tool in this research for pre-processing the dataset, which is in ARFF format (attribute-relation file format). Only 14 attributes of the 76 attributes have been considered for analysis to get accurate results. By comparing and analysing using different algorithms, heart disease can be predicted and cured early .

Approach Methodology

My research aims to make the prediction of having heart disease easy by using machine learning technique that is helpful in the medical field for clinicians and patients. To achieve my aim, I have discussed the use of various machine learning algorithms on the data set and dataset analysis is mentioned in my research paper. This paper also tells which attributes contribute more than the others for higher precision. This may spare the expense of different trials of a patient, as all the attributes may not contribute such a substantial amount to expect the outcome.

Description of the Dataset

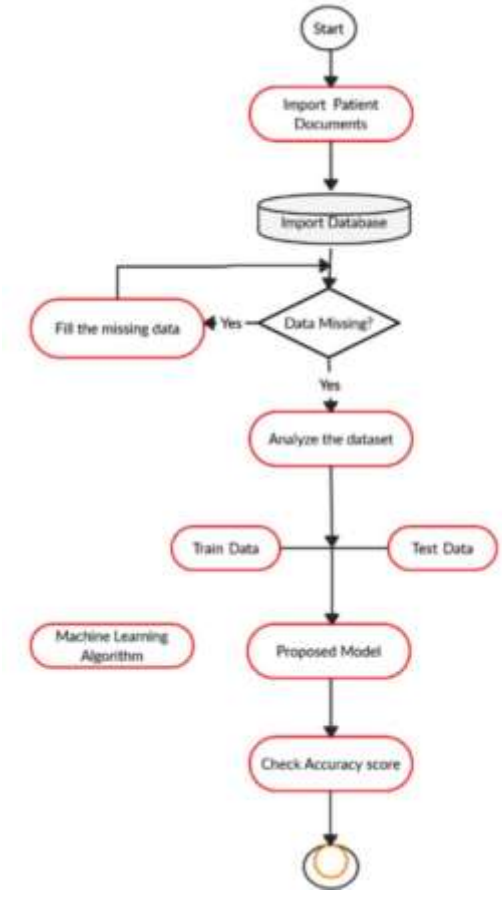
The dataset used for my research purpose was the Public Health Dataset which I have downloaded from kaggle . It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "output" or "target" field refers to the presence of heart disease in the patient. It is integer-valued 0 which indicates no disease and 1 indicate disease. Now the attributes which are used in this research purpose are described as follows and for what they are used or resemble:
(i)Age—age of patient in years, sex—(1 represents male; 0 represents female).
(ii)Cp— represents chest pain.
(iii)Trestbps—represents resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if it is a little higher than the normal range it is risky, you should try to lower it).
(iv)Chol—represents serum cholesterol shows the amount of triglycerides present. Triglycerides is a lipid that is measured in the blood. It should be below than 170 mg/dL.
(v) Fbs—represents fasting blood sugar greater than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
(vi)Restecg—represents resting electro cardiographic results.
(vii)Thalach—maximum heart rate achieved. The maximum heart rate can be calculated by 220 minus your age.
(viii)Exang—represents exercise-induced angina (1 yes). Angina is a type of chest pain caused due to low blood flow to the heart. Angina can lead to coronary artery disease.
(ix)Oldpeak—represents ST depression induced by exercise relative to rest

(x)Slope—represents the slope of the peak exercise ST segment.
(xi)Ca—represents number of major vessels (0–3) coloured by fluoroscopy.
(xii)Thal—no explanation given in dataset, but probably represents thalassemia (3 normal; 6 fixed defects; 7 reversible defects).
(xiii)Target (T)—no disease is represented by 0 and disease is represented by1, (angiographic disease status).

Data Pre-processing

The real-life information or data contains large numbers with missing and noisy data. These data are pre-processed to over-come such issues and make predictions vigorously. Figure explains the sequential chart of our proposed model.

Cleaning the collected data may contain missing values and may be noisy. To get an accurate and effective result, these data need to be cleaned in terms of noise and missing values are to be filled up. Transformation it changes the format of the data from one form to another to make it more comprehensible.



The dataset does not contain any null values. But many outliers are their which needs to be handled properly, and also the dataset is not properly distributed. Various plotting techniques were used for to check the distribution of the data, and outlier detection. All

these pre-processing techniques play an important role before we pass the data for classification or prediction purposes.

Checking the Distribution of the Data

The distribution of the data plays an important role when we need to the predict or classify a problem. This helps the model to find patterns in the dataset that leads to heart disease. 1 in graph represents people having heart disease and 0 represent people not having heart disease. Here 165 people suffer from heart disease and 138 people are without heart disease.

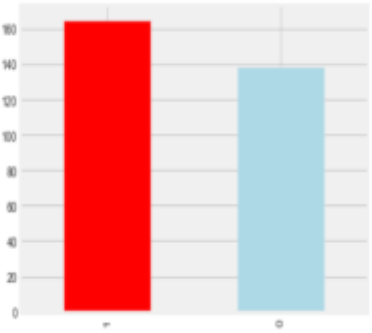


Fig 1:- Distribution of data

Analysis of Data

For Quantitative data:-

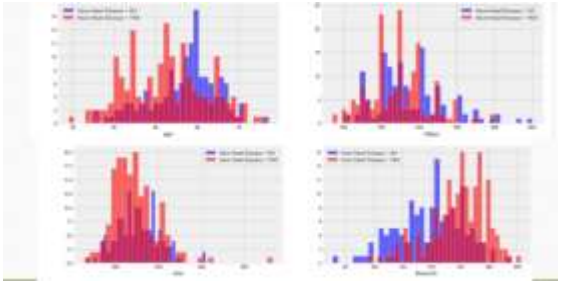


Fig 2:- Plots for numerical data

Observations that we can infer from the above plot are:

1. trestbps: represents resting blood pressure which above 130-140 is generally of concern
2. chol: if greater than 200 is of concern.

3. thalach: People with a higher value of over 140 are more likely to have heart disease.
4. The peak of exercise induced ST depression vs. rest looks at heart stress during exercise which tells that an unhealthy heart will stress more.

For Qualitative data:-

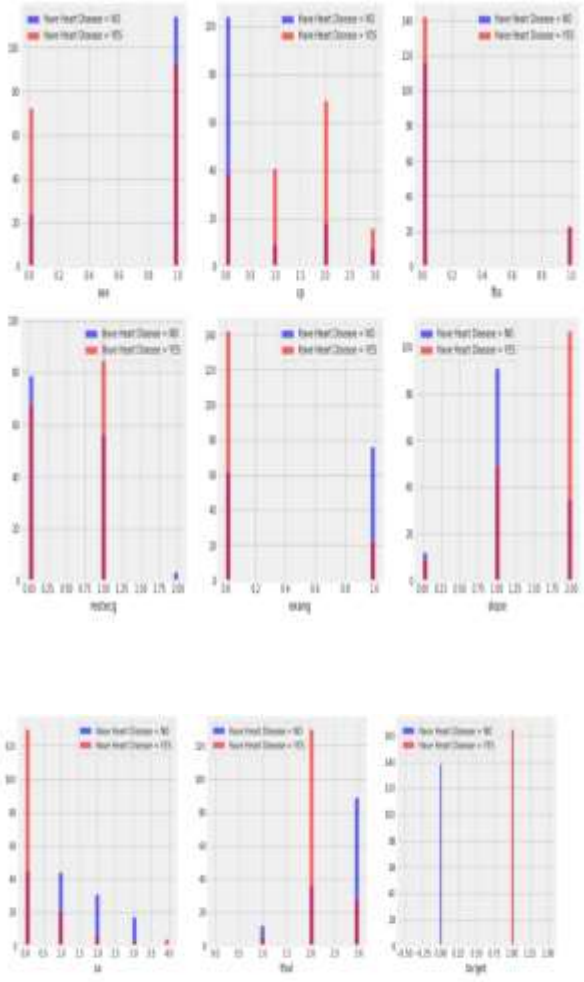


Fig3:- Plots for categorical data

Observations that we can infer from the above plot are:

1. cp {Chest pain}: People with cp 1, 2, 3 are more chances to have heart disease than people with cp 0.
2. restecg {resting EKG results}: People with a value of 1 (having an abnormal heart rate, which can range from mild symptoms to severe troubles) are more likely to have heart disease.
3. exang (exercise-induced angina): people with a value of 0 (No ==> angina induced by exercise) are more likely to have heart disease than people with a value of 1 (Yes ==> angina induced by exercise)

4. slope {the slope of the ST segment of peak exercise}: People with a slope value of 2 (Downslopins: indicates signs of an unhealthy heart) are more likely to have heart disease than people with a value of 2 slope is 0 (Upsloping: indicates best heart rate with exercise) or 1 (Flatsloping: indicates minimal change (typical healthy heart)).
5. ca [number of major vessels (0-3) stained by fluoroscopy]: the more movement of blood, people with ca equal to 0 are more likely to have heart disease.
6. thal {thalium stress result}: People with a value of 2 are more likely to have heart disease.

Correlation Matrix

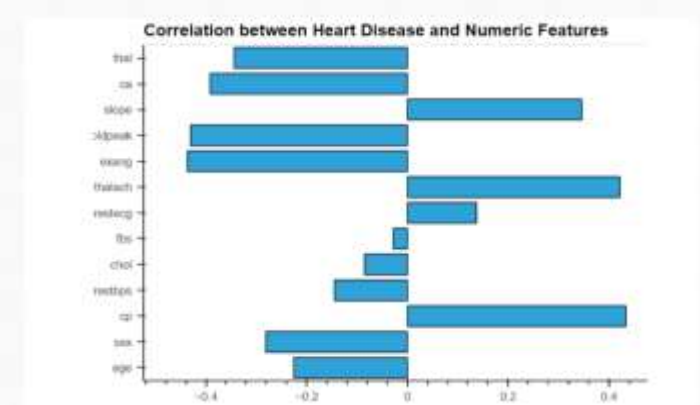


Fig 4:- Correlation Matrix

From the above fig we can observe that:-

1) fbs and chol are lowest correlated with the target variable.

2) All other variables have a significant correlation with the target variable.

Algorithm Used

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in thispaper are K nearest neighbours (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease. This paper work includes examining the journals, published paper and the data of cardiovascular disease of the recent times. The methodology is a process which includes steps that transforms the input data into known data patterns for the knowledge of the users. The proposed methodology includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the pre-processing stages where we can explore the data. The main role of data pre-processing is to fill the missing values and clean the data. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are K Nearest

neighbour(KNN), Logistic Regression, Random Forest Classifier. Finally, the proposed model is taken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in my model, an effective Heart Disease Prediction System (EHDPS) has been used using different classifiers. This model uses 13 health related parameters such as chest pain, blood pressure, blood pressure, age, cholesterol, fasting sugar, sex etc. for prediction.

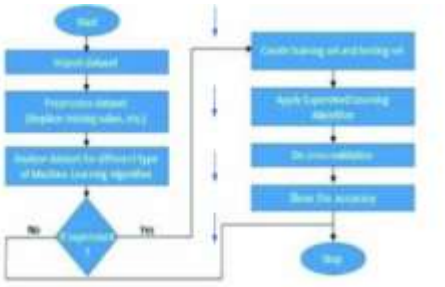


Fig 5:- Heart Disease prediction system flowchart

RESULTS & DISCUSSIONS

From the results I can infer that although most of the researchers have used different algorithms such as SVC, Decision tree for the detection of disease Random Forest Classifier provide a better result compared to them. The algorithms that I used are more accurate and save a lot of money i.e. it is cost efficient and provides fast results. Also, maximum accuracy was obtained by KNN and Logistic Regression which is almost more than 80% which is greater or almost equal to accuracies obtained from previous researches. So, to summarize that our accuracy may be improved due to the increased medical attributes that we used from the dataset we took.. The following 'figure 2', 'figure 3', 'figure 4','figure 5' shows a plot of the number of patient are been segregated and predicted by the classifier depending upon the age group, Resting Blood

Pressure, Sex, Chest Pain:

- Risk of Heart Attack
- No Risk of Heart Attack

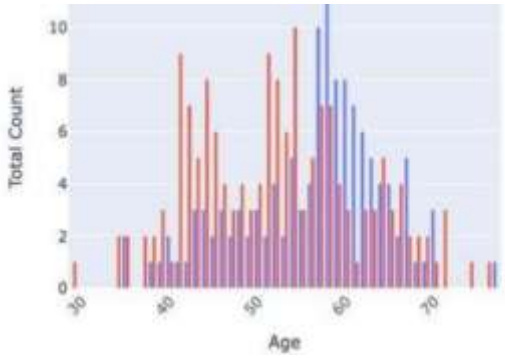


Fig 6:- Plot between age and total count

TABLE 1. Values Obtained for Confusion Matrix Using Different Algorithm

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	44	10	8	30
Naive Bayes	42	12	6	62
Random Forest	44	10	12	56
Decision Tree	50	4	8	60

TABLE 2. Analysis of Machine Learning Algorithm

Algorithm	Precision	Recall
Decision Tree	0.845	0.823
Logistic Regression	0.857	0.882
Random Forest	0.937	0.882
Naive Bayes	0.837	0.911

CONCLUSION

With the increasing number of deaths thanks to heart diseases, it's become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to seek out the foremost efficient ML algorithm for detection of heart diseases. This study gives a brief description about the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart

condition using the dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of more than 90% for prediction of heart disease. In future the work can be enhanced by developing an internet application supported the Random Forest algorithm also employing a large dataset as compared to the one utilized in this analysis which can provide better results and help health professionals in predicting the disease effectively and efficiently.

REFERENCES

[1] Sonam, A.M. “Predictions of Heart Condition Using Machine Learning Algorithms” in International Journal of Advanced Engineering, Management and Science (IJAEMS) June2016

vol-2

[3] Costas Sideris, Mohammad, Haik K, "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health

[4] Po Athi, Brad Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients Having Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pgno:1

[5] Dh, J K. Al, Mohamed Ibrahim, Mohammad Naeem "The Utilization of Machine Learning Approach for Medical Data Classification" in Annual Conference on New Trends in Information & Communication Technology Applications - march2017

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shou, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2,No. 3, June 2012

[7] Amu, J., Pad, S., Nandhini, R., Kavi, G., D, P., Venkata, V.S.K., "Recursive ant colony optimization routing in wireless mesh network",(2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.

[8] Ala, B.P., Kavitha, A., Amu, J., "A novel encryption algorithm for end-to-end secured optic communication", (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.

[9] Amu, J., In, P., B, B., Ananda, B., Ven, T., Prem, K., "An effective analysis on harmony search optimization approaches", (2015) International Journal of Applied Engineering Research, 10 (3),pp. 2035-2038.

[10]Amu, J., Kath, P., Reddy, L.S.S., Aa, A., "Assessment on authentication mechanisms in distributed system: A case study", (2017) Journalof Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.

[11]Amu, J., Kode, C., Prem, K., Jai, S., Raja, D.,Ven, T., Hari, R., "Comprehensive analysis on information dissemination protocols in vehicular adhoc networks", 6 (2015) International Journal of Applied Engineering Research, 10 (3), pp.2058-2061.

[12]Amu, P., Reddy, L.S.S., Satyanarayana, K.V.V., "Effects, challenges.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

A Cardiovascular Disease Prediction using Machine Learning Algorithms

Rubini PE[1], Dr.C.A.Subasini[2], Dr.A.Vanitha Katharine[3], V.Kumaresan[4], S.GowdhamKumar[5], T.M. Nithya[6]

[1]Assistant Professor, Department of Computer Science and Engineering, CMR Institute of Technology, Bengaluru.
[2] Associate Professor, Department of Computer Science and Engineering, St. Joseph's Institute of Technology, Chennai-119.
[3]Associate Professor, Department of Computer Applications, PSNA College of Engineering and Technology, Dindugul.
[4]Assistant Professor (Senior Grade), Department of Electrical and Electronics Engineering, Kongu Engineering College (Autonomous), Perundurai, Erode-638060.Email
[5]Training Officer, PSG Industrial Institute (PSG COLLEGE OF TECHNOLOGY), Peelamedu, Coimbatore-641041
[6] Assistant Professor, Department of Computer Science and Engineering, K. Ramakrishnan College of Engineering, Trichy.

ABSTRACT

Heart Diseases have shown a tremendous hit in this modern age. As doctors deal with precious human life, it is very important for them to be right their results. Thus, an application was developed which can predict the vulnerability of heart disease, given basic symptoms like age, gender, pulse rate, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, exercise induced angina, ST depression ST segment the slope at peak exercise, number of major vessels colored by fluoroscopy and maximum heart rate achieved. This can be used by doctors to re heck and confirm on their patient's condition. In the existing surveys they have considered only 10 features for prediction, but in this proposed research work 14 necessary features were taken into consideration. Also, this paper presents a comparative analysis of machine learning techniques like Random Forest (RF), Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes in the classification of cardiovascular disease. By the comparative analysis, machine learning algorithm Random Forest has proven to be the most accurate and reliable algorithm and hence used in the proposed system. This system also provides the relation between diabetes and how much it influences heart disease

Keywords:
Heart disease; Machine learning algorithms; Random Forest; Logistic regression; Support Vector Machine; Naïve Bayes; Diabetes Influence

1.Introduction

Coronary illness has the biggest level of passing on the planet. In 2012, around 17.5 million individuals kicked the bucket from coronary illness, implying that it comprises of the 31% of every single worldwide passing. Besides, coronary illness loss of life rises each year. It is relied upon to develop more than 23.6 million by 2030. The exploration from the January 2017 demonstrated that the main source of death worldwide is cardiovascular infections. The cardiovascular malady is considered as a world's biggest killer and is currently taking the top position in the record of ten reasons for passing in the previous 15 years and in 2015 was numeration for fifteen million passing. Various human lives could be spared by diagnosing on schedule. Along these lines, diagnosing the syndrome is significant and an exceptionally muddled undertaking. Mechanizing this procedure would conquer the issues with the diagnosis. The utilization of AI in ailment arrangement is normal and researchers are especially fascinated in the advancement of such frameworks for simpler following and analysis of cardiovascular diseases. Since ML permits PC projects to ponder from information, building up a model to perceive ordinary examples and having the option to settle on choices dependent on assembled data, it doesn't have hitches with the deficiency of utilized medicinal database. The proposed model is to amass significant information relating all components identified with coronary illness and parameters impacting it, train the information according to the proposed calculation of AI and

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

foresee how solid is there a probability for a patient to get a coronary illness. The relationship with the diabetes related credits is considered to set up the impact. [2]

2. Methodology

The methodology for predicting cardiovascular disease was done by using following four algorithms and the results are compared.Fig.1 describes the architecture diagram for predicting cardio vascular disease.

1. Random Forest
2. Logistic Regression
3. Naive Bayes algorithm
4. Support Vector Machines

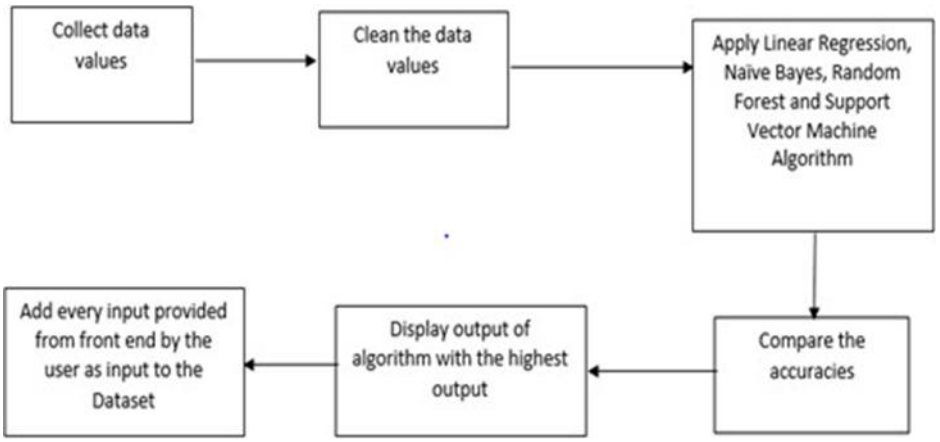


Figure 1: Methodology to predict heart disease

A. Random Forest Algorithm

The Random Forest Algorithm is understood as a forest comprised of trees. Firstly, it creates call trees on every which way chosen knowledge samples from the dataset. It then gets the prediction from each tree and selects the most effective resolution through means voting. It is an enhancement from decision trees [3]. Some of its applications are image classification, recommendation engines and feature selection. This algorithmic rule is considered as an extremely correct and strong methodology as a result of the number of trees collaborating within the method. One amongst its many advantages is that it does not suffer from the over fitting problem. Finally, it takes the average of all the predictions from every tree, which cancels out the biases.

1. Dataset collection and pre-processing

The dataset which was used for analysis are “Framingham” obtained from Kaggle. Heart disease dataset with 14 features is obtained from UCI Machine Learning Repository [19]. Data is cleaned by replacing all the non-available values with the median of values in that column. Categorical data are assigned with numerical values.

2. Implementation

The implementation of random forest works as follows:

a. Load the heart disease dataset.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

b. After Preprocess, Split the heart disease dataset into train and test data with the proportion of 60:40 using Random Forest Classifier function.
c. K-Fold Cross Validation is wherever a given knowledge set is split into a K range of sections/folds wherever every fold is employed as a testing set at some purpose.
d. Train the model using train set.
e. Make predictions on the test fold.
f. Map predictions to outcomes (only possible outcomes are 1 and 0).
g. Calculate the accuracy.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FN+FP)} * 100$$

Where,
TP- True Positive (prediction is yes, and they do have the disease.
TN-True Negative (prediction is no, and they don't have the disease.)
FP-False Positive (We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
FN-False Negative (We predicted no, but they actually do have the disease. (Also known as a "Type II error.")
The accuracy obtained by using random forest algorithm is 84.81%

```
predictors=["age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slope","ca","thal"]
alg=RandomForestClassifier(n_estimators=75,min_samples_split=40,min_samples_leaf=1)
kf=KFold(heart.shape[0],n_folds=16, random_state=1)
predictions = []
for train, test in kf:
    # The predictors we're using the train the algorithm.  Note how we only take the rows in the train folds.
    train_predictors = (heart[predictors].iloc[train,:])
    #print(train_predictors)
    # The target we're using to train the algorithm.
    train_target = heart["heartpred"].iloc[train]
    #print(train_target)
    # Training the algorithm using the predictors and target.
    alg.fit(train_predictors, train_target)
    # We can now make predictions on the test fold
    test_predictions = alg.predict(heart[predictors].iloc[test,:])
    predictions.append(test_predictions)
# The predictions are in three separate numpy arrays.  Concatenate them into one.
# We concatenate them on axis 0, as they only have one axis.
predictions = np.concatenate(predictions, axis=0)

# Map predictions to outcomes (only possible outcomes are 1 and 0)
predictions[predictions > .5] = 1
predictions[predictions <=.5] = 0
```

Figure 2: Sample Code of Random Forest

```
# Map predictions to outcomes (only possible outcomes are 1 and 0)
predictions[predictions > .5] = 1
predictions[predictions <=.5] = 0
i=0
count=0
for each in heart["heartpred"]:
    if each==predictions[i]:
        count+=1
    i+=1
accuracy=count/i
print("Random Forest Result:-")
print("Accuracy = ")
print(accuracy*100)
```

```
Random Forest Result:-
Accuracy =
84.81848184818482
```

Figure 3: Accuracy result of Random Forest algorithm

B. Support Vector Machine

1. Introduction

Support Vector Machines is a classification technique which separates data values by the creation of hyper planes. Hyper planes can be of different shapes based on the spread of data, but only those points which help in differentiating between the classes are considered for classification.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

2. Kernel Functions

If data points are in nonlinear fashion, the kernel function makes them towards linear decision surface.
Some Kernel functions are as follows:

a. Linear Function: In these kinds of kernel the hyper plane is a straight line. Linear Kernel functions can provide best results for classifiers which have exactly two target classes.
b. Polynomial Function: In such kinds of kernel functions the hyper plane is generally a polynomial like parabola, hyperbola.
c. Radial Basis Function: Radial Basis Function is put in use when points cannot be separated in a linear fashion. The function works to bring points into a shape mostly radial/circular fashion to perform further actions.

3. Implementation

The implementation of Support Vector Machine described as follows:
a. Load the data sets and clean values, in case of no value for a particular feature inarowreplacewiththemedianvaluetherowfrom thedataset.
b. Split the data set into train and test in 60:40 ratio respectively.
c. Choosing the Kernel Function as Linear Kernel Function or Radial Basis Function.
d. Applying SVM by first creating a hyper plane with the help of test data set.
e. Calculate the accuracy using

- The train data is taken and both Kernel function namely Linear Kernel Function or Radial Basis Function is applied.
- Apply test data set on the trained model.
- The model uses hyper plane and finds closest proximitytoeitherclassthatishavingheartdisease (yes/1) or not having heart disease(no/0).

$$\text{Accuracy} = \frac{(\text{Number of data items predicted} = \text{actual value in test data set})}{\text{Total Number of values in test data set}}$$

Kernel Functions	Accuracy (%)
Linear Kernel Function	74.05
Radial Basis Function (RBF)	58.577

TABLE 1- Comparison of SVM accuracies with Kernel Functions

In Table 1 the calculation accuracies for both SVM Models with RBF and Linear Function as Kernels are examined. Linear Kernel Function provides higher accuracy than RBF. This is because the problem is a two-class classifier problem. Hence a hyper plane in the form of a line would be the best way to classify such values. In comparison RBF uses a circle as hyper plane thus producing lower accuracy. The hyper plane plot for SVM for predicting heart disease is shown in Fig.4.In this the yellow plot represents patients having heart disease and purple dots represents the patients not having heart disease.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

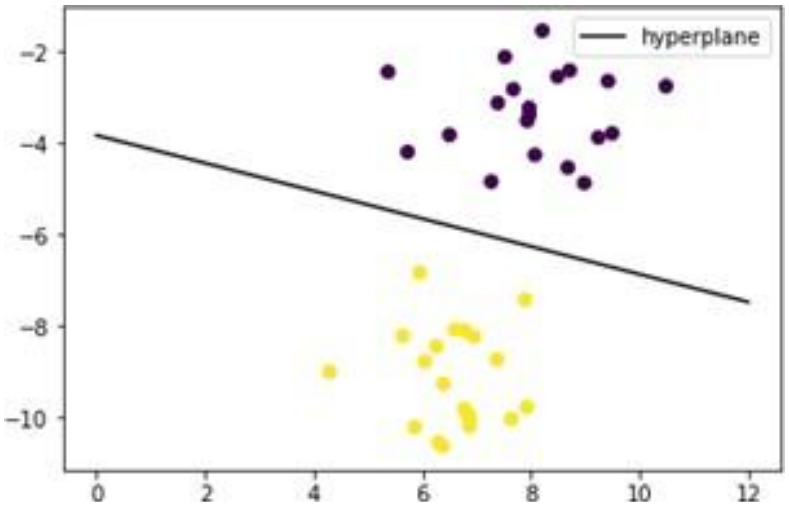


Figure 4: Hyper plane and distribution of data points on either side of hyper plane for Heart Disease Prediction

C. Naïve Bayes Classification

Naïve Bayes classifier is based on probability which is mostly used in the training phase. This algorithm is used for removing the redundant data from the datasets.

1. Implementation

The implementation of Naive Bayes is as follows:

a. Extract the dataset.
b. Apply cleaning on the dataset to remove unwanted values.
c. In case any values are missing then find the median value of the column and fill the missing value.
d. Find the deterministic probability with occurrence of heart disease with respect to 14parameters.
e. Then find the conditional probability of non-occurrence of heart disease with respect to 14 parameters.
f. Train the model using this probability formula given below

$$= P(W|Q) = \frac{P(Q|W)P(W)}{P(Q)} \quad \frac{P(Q|W)P(W)}{P(Q|W)P(W)+P(Q|M)P(M)} \quad (1)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

$$P(y|x1,\ldots.x14) = \frac{P(x1|y)P(x2|y)\ldots\ldots P(x14|y)P(y)}{P(x1)P(x2)\ldots\ldots P(x14)} \quad (3)$$

$$P(y|x1,\ldots\ldots.x14) = \frac{P(y)\ \pi P(xi|y)}{P(x1)P(x2)\ldots\ldots P(x14)} \quad (4)$$

$$P(y|x1,\ldots\ldots,x14)\alpha P(y)\pi P(xi\ |y) \quad (5)$$

Where, x1- age; x2- sex; x3-cp; x4 – rest bp; x5- chol; x6- fbs; x7- rest ecg; x8- thalach; x9 - exang;x10-oldpeak; x11-slope; x12-ca; x13-thal; x14-pulse rate.

g. As soon as the model is trained, then apply the test data set.
h. Remove the last column of the test data set which determines the person will have heart attack or not.
i. Apply the model on the test data set and extract the values.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

j. Compare the result between the last column and the predicted values.
k. Calculate the accuracy.

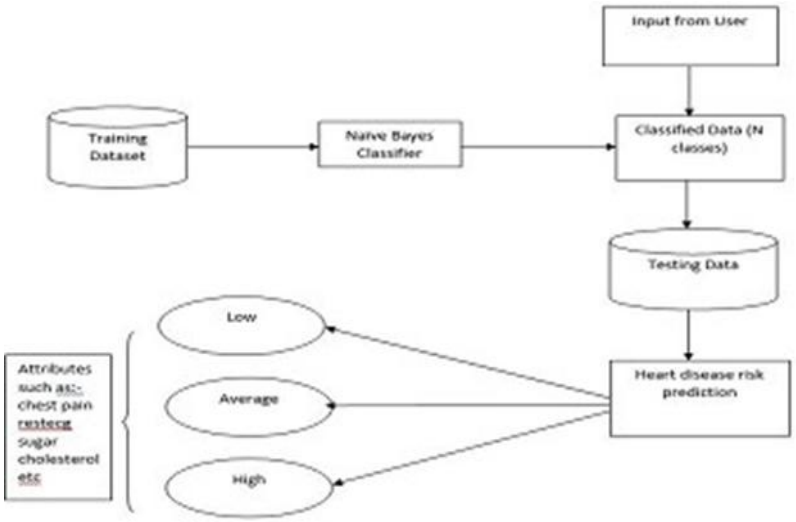


Figure 5: Working of Naïve Bayes

D. Logistic Regression

Logistic regression is a machine learning algorithm used for classification. It is based on the concept of probability. Logistic regression is used to assign observations to a discrete class. Transforming output is done using the sigmoid logic function. The logistic regression hypothesis tends to limit the cost function in range between 0 and 1. Therefore, linear functions cannot represent as it can have a value >1 or <=0, which is not possible according to the regression hypothesis.

1. Implementation

The implementation steps for logistic regression are given a follow:

a. **Obtain the probabilities:** Mapping predicted values to probabilities, using the Sigmoid function.

$$1/(1 + e^{-y}) \quad (6)$$

where, y is input to the function and e is the base of natural log. Obtain the probabilities by following equations:

$$P = e^y / 1 + e^y \quad (7)$$

where P is the probability of success, and q is the probability of failure written as:

$$q = 1 - P = 1 - (e^y / 1 + e^y) \quad (8)$$

on dividing, (7) / (8), we get

$$\frac{p}{1-p} = e^y \quad (9)$$

On taking log on both sides,

$$\log \frac{p}{1-p} = y \quad (10)$$

where (p/1-p) is the odd ratio. When 'y' is positive, the probability of success is more than 50%.

b. Decision Boundary-Mapping probabilities to classes

Prediction function returns probability score between 0 and 1. To assign to a discrete class, a threshold value is selected above which it is classified as class 1 or else class 2. For example, if our threshold was 0.5 and our function value was 7, it is classified as positive. For say .3,

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

classification is negative. Logistic regression can also have multiple classes where the highest probability predicted class is considered.

2. Analysis of result:

The result can be analyzed in following ways.

a. Using Confusion Matrix: Accuracy is calculated by formula

Accuracy= ((TP + TN) / (TP + TN + FP + FN)) * 100

Where TP- True Positive, TN-True Negative, FP-False Positive, FN-False Negative

b. ROC curve: The receiver operating characteristic summarizes the performance when evaluating the compensations between the sensitivity and the 1-specificity. To plot ROC, assume p> 0.5. The area under the curve, indicated as an index of precision or concordance index, is a performance metric for curve. The larger the area under the curve, the better the predictive power of the model.

```
from sklearn.metrics import roc_curve
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob_yes[:,1])
plt.plot(fpr,tpr)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.title('ROC curve for Heart disease classifier')
plt.xlabel('False positive rate (1-Specificity)')
plt.ylabel('True positive rate (Sensitivity)')
plt.grid(True)
```

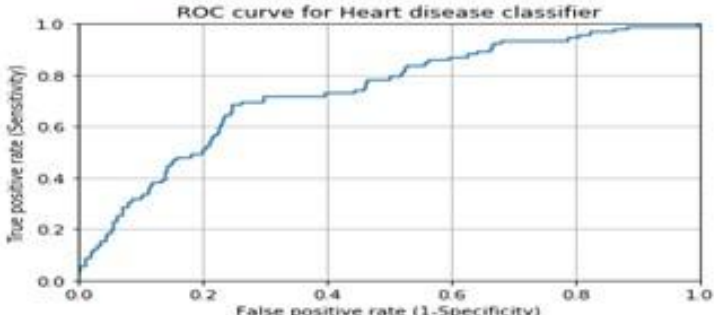


Figure 7. ROC Curve - Logistic Regression

```
# Map predictions to outcomes (only possible outcomes are 1 and 0
predictions[predictions > .5] = 1
predictions[predictions <=.5] = 0
i=0
count=0
for each in heart["heartpred"]:
    if each==predictions[i]:
        count+=1
    i+=1
accuracy=count/i
print("Logistic Regression Result:-")
print("Accuracy = ")
print(accuracy*100)
```

```
Logistic Regression Result:-
Accuracy =
83.82830283828383
```

Figure 8. Accuracy result of Logistic Regression

3. Result

Results from Random Forest, Support Vector Machine, Logistic Regression and naïve Bayes are analyzed, and Random Forest Algorithm has given the highest accuracy. Hence Random Forest has been implemented in the proposed system.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

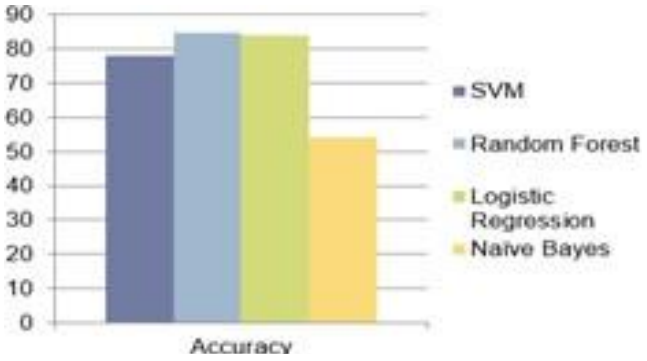


Figure 9. Graphical Representation of Accuracy

ALGORITHM	**ACCURACY (%)**
RANDOM FOREST	84.81
LINEAR REGRSSION	83.828
SUPPORT VECTOR MACHINE (Using Linear Kernel Function)	74.05
SUPPORT VECTOR MACHINE (Using Radial Basis Kernel Function)	58.577
Naïve Bayes	54.08401

TABLE II Comparison of Accuracies

4. Conclusion and Future Scope

Heart disease prediction which uses Machine learning algorithm provides users a prediction result if the user has heart disease. Recent advancements in technology made machine learning algorithms to evolve. In this proposed method Random Forest Algorithm was used because of its efficiency and accuracy. This algorithm is also used to find the heart disease prediction percentage by knowing the correlation details between diabetes and heart diseases. The similar prediction systems can be built by calculating correlation between heart diseases and other diseases. Also new algorithms can be used to achieve increased accuracy. Better performance is obtained with more parameter used in these algorithms.

References

[1] Jaymin Patel, Prof.Tejal Upadhyay, Dr.Samir Patel “Heart disease prediction using Machine learning and Data Mining Technique" Volume 7.Number1 Sept 2015-March 2016.
[2] Thenmozhi.K and Deepika.P, Heart Disease Prediction using classification with different decision tree techniques. International Journal of Engineering Research & General Science, Vol 2(6), pp 6-11, Oct 2014.
[3] Igor Kononenko” Machine learning for medical diagnosis: history, state of art& perspective"

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021.

Elsevier -Artificial intelligence in Medicine, Volume23, Aug 2001.

[4] Gregory F. Cooper, Constantin F. Alfieris", Richard Ambrosino, John Aronisb, Bruce G. Buchanan, Richard Caruana', Michael J. Fine, Clark Glymour", Geoffrey Gordon", Barbara H. Hanusad, Janine E. Janoskyf, Christopher Meek", Tom Mitchell", Thomas Richardson", Peter Spirtes" An evaluation of machine-learning methods for predicting of pneumonia mortality"-Elsevier Feb 1997

[5] Sana Bharti, Shailendra Narayan Singh" Analytical study of heart disease comparing with different algorithms": Computing, Communication & Automation (ICCCA), 2015InternationalConference.

[6] B.Dhomse Kanchan, M. Mahale Kishore "Study of Machine learning algorithms for special disease predictions using the principal of component analysis" Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016.

[7] MatjazKuka, Igor Kononenko, Cyril Groselj, Katrina Kalif, JureFettich" Analysing and improving the diagnosis of ischaemic heart disease with machine learning" Elsevier - Artificial intelligence in Medicine, Volume23, May 1999.

[8] Geert Meyfroidt, FabianGuiza, Jan Ramon, Maurice Brynooghe" Machine learning techniques to examine large patient databases"-Best practice & Reasearch Clinical Anaesthesiology, Elsevier Volume 23 (1) Mar 1, 2009.

[9] Gregory F.Cooper, ConstantinF.Aliferis, Richard Ambrosino"An evaluation of Machine learning methods for predicting pneumonia mortality"-Elsevier, 1997.

[10] Sanjay Kumar Sen" Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms", International Journal of Engineering And Computer Science ISSN:2319-7242Volume6Issue 6 June 2017.

[11] Abhishek Taneja" Heart Disease Prediction SystemUsing Data Mining Techniques"-Vol.6, No(4) December 2013.

[12] AnimeshHazra, Subrata Kumar Mandal, AmitGupta,Arkomita Mukherjee and Asmita Mukherjee" Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review"- Advances in Computational Sciences and Technology ISSN 0973-6107, Volume10, Number7(2017).

[13] BeantKaur, Williamjeet Singh" Review on Heart Diseases Prediction System using different Data Mining Techniques"- International Journal on Recent and Innovation Trends in Computing and Communication Volume:2 Issue:10, October 2014.Transll. J. Magn. Japan, vol. 2, pp. 740-741, August 1987.

[14] SonamNikhar, A.M. Karandikar" Prediction of Heart Disease Using different Machine Learning Algorithms"- Vol-2 Issue-6, June 2016.

[15] S. U. Ghumbre and A. A. Ghatol, "Heart Disease Diagnosis Using Machine Learning Algorithm," Advances in Intelligent and Soft Computing Proceedings of the International Conference on Information Systems Design and Intelligent Applications.

[16] Machine learning based decision support systems (DSS) for heart disease Diagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01

[17] DataSetURL-https://archive.ics.uci.edu/ml/machielearnindatabses/heartdisease