Predicting Cardiovascular Disease with Machine Learning

Mr.C.Rajeshkumar[1]. Dr.K.Ruba Soundar[2], Dr.M.Vargheese[3], Dr.G.Nallasivan[4], A.Selvakumar[5]

1Assistant Professor/CSE,PSN College of Engineering and Technology,Tirunelveli,India,
[2]Associate Professor/CSE,Mepco Schlenk Engineering College,Sivakasi,India,
[3]Professor/CSE, PSN College of Engineering and Technology,Tirunelveli,India,email:
[4]Professor/CSE, PSN College of Engineering and Technology,Tirunelveli,India,email:
[5]CSE Student, PSN College of Engineering and Technology, Tirunelveli,India,

Abstract:Machinelearninghasmany applicationsin today'smodern environment.In themedical field, there are no exceptions. Predicting the presence of problems in locomotion,cardiac function, and other areas may be greatly aided by machine learning. Such information, ifpredicted in advance, may provide vital intuitions to physicians, allowing them to tailor theirdiagnosis and treatment plan to each individual patient. Here, patient records are read from acomma-separated values (CSV) file. After obtaining the data, the procedure is executed, and avaluable heart attack level is produced. We're training machine learning algorithms to identifythose at risk for getting heart disease. Classifiers such as decision trees, logistic regression,supportvectormachines,and randomforestsarecompared and contrasted inthisresearch.

Keywords:Machinelearning,HeartDiseasePrediction,Decisiontreealgorithm,RandomForest,SupportVectorMachine,LogisticRegression.

Published in IJIRMPS (E-ISSN: 2349-7300), Volume 11, Issue 2, March-April 2023

INTRODUCTION

According to the World Health Organization, around 12 million people worldwide losetheir lives each year due to cardiovascular disease. Heart disease is a prominent cause of deathand disability across the world. Predicted cardiovascular disease is one of the most importantissues in the field of data analysis. Recent years have seen a dramatic increase in the globalincidence of cardiovascular disease. Since death from heart disease often occurs suddenly andwithout warning,it has earned the nickname "silent killer." An early diagnosis of heart diseaseis critical for helping high-risk patients make decisions about whether or not to make lifestylechangesthatreducetheseverityofthe condition.

In this case, the use of machine learning techniques may be quite useful. Despite the factthat heart disease may appear in a variety of ways, there is a common set of fundamental riskfactors that determine whether or not someone will eventually be at risk for heart disease. Wemay say that this method is well suited to accomplishing heart disease prediction by collectingdata from various sources, classifying them under suitable categories, and then analyzing toacquire therequireddata.

Motivation

The major purpose of this research was to suggest a model for predicting the onset ofcardiovascular disease. In addition, this study's objective is to zero in on the best approach toheart disease categorization. Supporting this work is a comparative study and evaluation of fourclassificationalgorithms:SVM,DecisionTree,LogisticRegression,andRandomForest.Despite the widespread use of machine learning techniques, accurately forecasting heart illnessremains a mission critical challenge. As a result, the three algorithms are evaluated using a widerange of criteria and testing procedures. Because of this, researchers and doctors will be able todevelopamoreeffective.

LITERATURESURVEY

Recent years have seen a flurry of noteworthy articles detailing the results of trials and researchintotheintersectionofmedicalscience andmachinelearning.

Using hill climbing and decision tree algorithms, Purushottam et al. suggested a "Efficient HeartDisease Prediction System" in their study [1]. They took the Cleveland dataset, and before usingclassification methods, they preprocessed the data. To complete the Knowledge Extraction, anopen-source data mining method called Evolutionary Learning (KEEL) is used to infer themissing variables. A decision tree is built from the top down. One node is chosen at each stageof the hill climbing algorithm based on the results of a test. The used parameters and values maybe trusted. It has a confidence interval of at least 0.25. The method has an approximate 86.7%rateofaccuracy.

Prediction of cardiac disease using machine learning algorithms was suggested by SanthanaKrishnan et al. in [2], namely the use of a decision tree and the Naive Bayes method. In adecision tree algorithm, branches are constructed according to criteria that determine whether anode is True or False. Algorithms like support vector machine and k-nearest neighbor usedependentfactorstodeterminewhethertodividethedataverticallyorhorizontally.Incontrast,a decision tree is a tree-like structure in which each node, or branch, is based on a previouschoice. Using a decision tree may also provide light on which data points are most crucial.Cleveland has also been utilized in their analysis. Some approaches divide the dataset such that70% is used for training and 30% is used for testing. The accuracy of this method is 91%. NaiveBayes, another classification method, comes in at number two. Because it is capable of dealingwith complex, nonlinear, dependant data, it is deemed appropriate for the heart disease dataset.Thealgorithmachieves anaccuracyof87%.

A neural network approach called multilayer perceptron is used for both training and evaluatingthe dataset in the study "Prediction of Heart Disease Using Machine Learning" presented byAditi Gavhane et al. There will be one input layer and one output layer in this algorithm, withone or more hidden levels in between. Each input node is linked to the output layer throughhidden layers. Some arbitrary weights are put on this link. The connection between the nodesmay be feedforward or feedback, and the other input is termed bias, which is assigned withweightdependingonnecessity.

AvinashGolande et al. published "Heart Disease Prediction Using Effective Machine LearningTechniques" in [4], which used a limited set of data mining methods to aid physicians in makingdistinctions between various forms of heartdisease. K-nearestneighbor, decision tree, andNaive Bayes are the most often used methods. Packing calculation, Part thickness, consecutivenegligible streamlining and neural systems, straight Kernel self-arranging guidance, and SVM(Boltzmann Machine) are some more examples of characterization-based techniques that areused.

More risk factors for cardiovascular disease are taken into account in the "Machine LearningTechniques for Heart Disease Prediction" suggested by Lakshmana Rao et al. in [5]. Thus,distinguishing cardiac illness is challenging. Various neural networks and data mining methodsare utilizedtoassessthe severityofheartdiseaseinindividuals.

In [6], Abhay Kishore et al. introduced a heart attack prediction system that employs Deeplearning methods and makes use of a Recurrent Neural System in order to forecast the likelyelements of the patient's heart-relatedillnesses. In order to provide the mostaccurate modelwith the fewest possible errors, this model employs deep learning and data mining. This workservesasagoldstandard againstwhichotherheartattackprediction modelsmaybejudged.

The primary goal of the "Effective Heart Disease Prediction Using Hybrid Machine LearningTechniques" proposal by Senthil Kumar Mohan et al. in [7] is to increase precision in predictingcardiovascular diseases. KNN, LR, SVM, and NN are the algorithms used in the heart diseaseprediction model with hybrid random forest with linear model (HRFLM), which results in anenhancedexhibitionlevelwithaprecisionlevelof88.7percent.

Performance of prediction for two categorization models is studied and compared to prior work,as proposed by Anjan N. Repaka et al. in [8]. The experimental findings demonstrate that oursuggestedstrategyoutperformscompetingmodelsintermsofaccuratelypredictingtheproportionatrisk.

4. METHODOLOGY

To get a feel for how far ML has come in its application to heart disease prediction, we combedthrough the existing literature. Finding research gaps and providing direction for further studiesin a topic are two key benefits of doing a comprehensive literature review. All relevant papersare retrieved from online resources, synthesized, and presented utilizing a method in SLR toanswer the research challenges described in the study. Newcomers to the field may get insightinto the state of the art thanks to the newperspectives and knowledge provided by an SLRstudy.

DatasetCollection:

Webegin by amassing a dataset to serve as the backbone of ourheartdisease predictionalgorithm. After collecting data, we split it into a training and a testing set. The training datasetis used to teach the prediction model, while the testing dataset is used to assess its performance.In this project, we only put 70% of the data through its training paces and 30% through itstesting paces.The project'sdatacamefromthe UCIHeartDisease Database.

Selectionofattributes:

Attribute or feature selection encompasses the process of deciding which characteristics will beused in the prediction system. The purpose of this is to improve the efficiency of the system.Multiple patient factors, such as gender, the kind of chest pain, fasting blood pressure, serumcholesterol, etc., are selected for the prediction. In this approach, a correlation matrix is used todecide whichattributes toinclude.

Pre-processingofData:

Oneof themostimportantstepsin creatingamachinelearningmodelisthepre-processing of data. Inaccurate findings might be the consequence of poorly cleaned or formatteddata being fed into the model. In order to get the information we need, we must first "pre-process" it. It is used to deal with the ambiguity, redundancy, and missing values included in thedataset. Data pre-processing includes tasks like importing datasets, splitting datasets, scalingattributes,etc.Improvingthemodel'sprecisionrequirespreprocessingthe data.

BalancingofData:

Itispossibletobalanceanunbalanceddatasetintwoways.Bothproblems—underandoversampling—are present.

UnderSampling

DatasetbalanceinUnderSamplingisachievedbyreducingthesizeofthelargeclass.
Whenthereisenoughdata,thisprocedureistaken intoaccount.

OverSampling

When oversampling, the tiny samples are enlarged to restore data balance. This method is consideredwhen there is insufficientdata. whenwe are itching to dive headfirstinto studying amassive datasetand building an ML model. We keep getting a "out of memory" issue whenever we try to load thedatasetintoourcomputer.When dealingwithmassive datasets,thishappensall the time.When itcomes to data science, one of the biggest obstaclesis handling "big datasets" on computationallylimited devices (a problem that may be solved, of course, with more resources). The term "sampling"referstoa statisticaltechniqueusedtomanageabiggerdataset.

PredictionofDisease

For classification, a variety of machine learning algorithms are employed, includingSVM, Decision Tree, Random Tree, Logistic Regression, Ada-boost, and XG-boost. Algorithmsarecompared,andtheonethatpredictsheartdiseasewiththebestdegreeofaccuracyischosen.

diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
0	195	106	70	26.97	80	77
0	250	121	81	28.73	95	76
0	245	127.5	80	25.34	75	70
0	225	150	95	28.58	65	103

Table:1Heartdiseasedataset

Step 1:LoadingthecsvDataset

Step2:PlottingHistogramandcreatingaheatmap

Step3: PerformingSmotingprocesstoclassifythe PositiveandNegativeResults.

Significance: Thesmotingresultisconcernedwiththefouralgorithms.

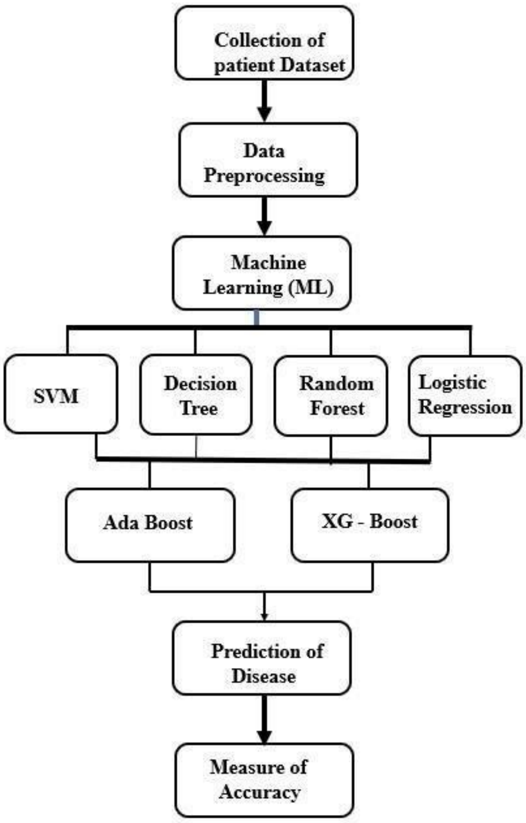
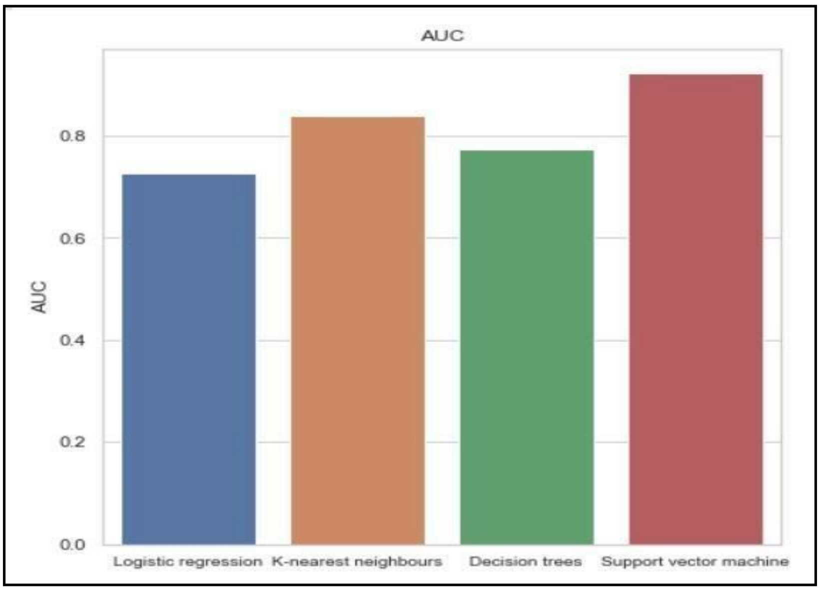


Figure:ArchitectureDiagram

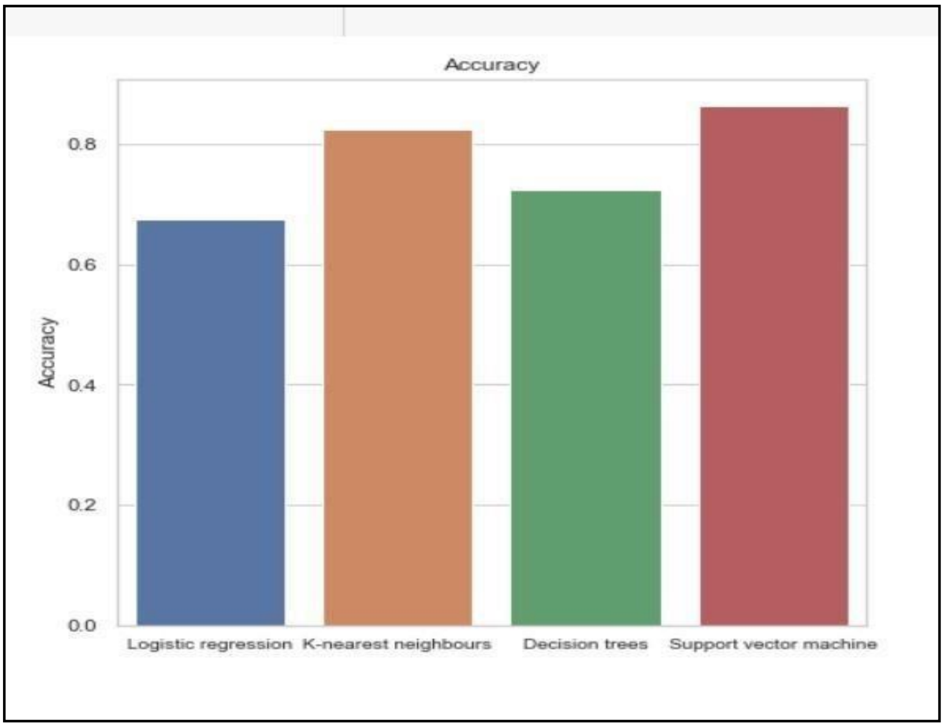
EXPERIMENTANDANALYSIS

AccuracycalculationsandprojectexecutionforHeartDiseasearecomplete.Fouralgorithms are used to specify the user'sheartrate,blood pressure,glucoselevel,smokingstatus,andbodymassindex,withthebestresultexpectedtobeHighestAccuracy.Th ealgorithmis carriedoutinthefollowingstages:

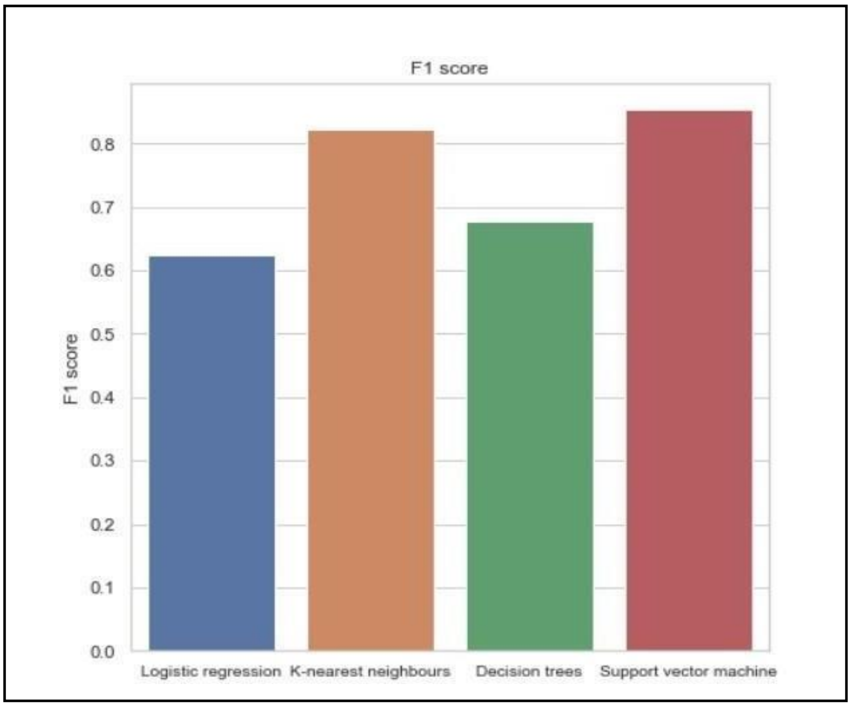
The results are accurate, as you can see here. Data from the Heart Disease PredictionModelarereadintotheinputdataset.



TheabovefigureshowstheAreaUnderROCCurvefortheLogisticRegression,K–NearestNeighbour,DecisionTrees,SupportVectorMachine.



The abovefigureillustratesthe AccuracyofDataset.



TheabovefigureillustratestheF1scoremeanoftheDataset.

CONCLUSION

As one of the main causes of mortality in India and throughout the globe, heart disorderswould have a profound societal effectif promising technologies like machine learning wereused to the early prediction of heart issues. Significant medical progress may be made with theearly diagnosis of heart illness, which may aid high-risk individuals in deciding on preventativelifestyle changes. Heart disease is becoming more prevalent among the general population.Because of this,promptidentification and treatmentare essential.Using the rightkind oftechnological aid in this area might be very beneficial for both doctors and patients. SVM,Decision Tree, Random Forest, and Logistic Regression are just some of the seven machinelearning methods put to the test here. The dataset was subjected to both Adaptive Boosting andExtreme GradientBoosting.

The76-featuredatasetrepresentstheexpected factors thatleadtoheartdiseaseinpeople,and14relevantfactorsareselectedfromitforevaluationpurposes.Takingintoconsideration all of the system's characteristics results in less efficiency for the author. The goalof attribute selection is to boost productivity. Here, n features are selected for use in evaluatingthecorrectnessofdifferentmodels.Becauseoftheirhighassociationswithothervariables, somepropertiesofthedatasetarediscounted.Ifeveryattributeinthedatasetisused,performance suffers dramatically.

REFERENCE

1. SoniJ,AnsariU,SharmaD&SoniS(2011).Predictivedataminingformedicaldiagnosis:anoverviewofheart diseaseprediction.InternationalJournalofComputerApplications,17(8),43-8
2. Dangare C S &Apte S S (2012). Improved study of heart disease prediction system usingdata mining classification techniques. International Journal of Computer Applications, 47(10),44-8.
3. Ordonez C (2006). Association rule discovery with the train and test approach for heartdisease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.
4. Shinde R, Arjun S, Patil P &Waghmare J (2015). An intelligent heart disease predictionsystem using k-means clustering and Naïve Bayes algorithm. International Journal of ComputerScience andInformationTechnologies,6(1),637-9.
5. BashirS,QamarU&JavedMY(2014,November).Anensemble-baseddecisionsupportframeworkforintelligentheartdiseasediagnosis.InInternationalConferenceonInformation Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series:Materials ScienceandEngineering1022(2021) 012072IOP Publishingdoi:10.1088/1757-899X/1022/1/0120729
6. Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronaryheartdiseaseprediction model:TheKoreanHeart Study.BMJopen,4(5),e005025.
7. Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J &Ingelsson E(2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis,thrombosis,andvascularbiology,33(9),2267-72.
8. Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease predictionusing lazy associative classification. In 2013 International Mutli- Conference on Automation,Computing,Communication,ControlandCompressed Sensing (iMac4s)(pp.40-6).IEEE.
9. Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths fromacute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register.BMJ,315(7101),159-64.

10. FolsomAR,PrineasRJ,KayeSA&SolerJT(1989).Bodyfatdistributionandselfreported prevalence of hypertension, heart attack, and other heart disease in older women.Internationaljournalofepidemiologyy,18(2)
11. Ruba Soundar K Rajathi L V, "An advancement in energy efficient clustering algorithm using cluster coordinator-based CH election mechanism (CCCH)", Measurement: Sensors, Vol.25, pp.1-6, 2023
12. Ruba Soundar K Kavitha N, "Automatic Identification of Cyber Predators Using Text Analytics and Machine Learning", River Publishers, pp.41-54, 2023
13. K.Muthamil Sudar, K.Ruba Soundar, P.Vinoth, P. Nagaraj, V. Muneeswaran, "Detection of DDoS Attack Using Machine Learning Techniques in Software Defined Networking", IGI Global Publishers, 2023