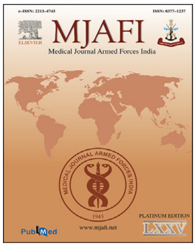
Available online at www.sciencedirect.com

ScienceDirect

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



Original Article

Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India

Ekta Maini [a,*], *Bondu Venkateswarlu* [b], *Baljeet Maini* [c], *Dheeraj Marwaha* [d]

[a] *Research Scholar (Computer Science & Engineering), Dayananda Sagar University, Bengaluru, India*
[b] *Associate Professor (Computer Science & Engineering), Dayananda Sagar University, Bengaluru, India*
[c] *Professor Pediatrics, Teerthanker Mahaveer Medical College & Research Centre, Moradabad, India*
[d] *Senior Software Engineer, Microsoft India, Hyderabad, India*

ARTICLE INFO

Article history:
Received 29 January 2020
Accepted 14 October 2020
Available online xxx

Keywords:
Affordable healthcare
Cardiovascular diseases
Early diagnosis
Machine learning

ABSTRACT

Background: In India, huge mortality occurs due to cardiovascular diseases (CVDs) as these diseases are not diagnosed in early stages. Machine learning (ML) algorithms can be used to build efficient and economical prediction system for early diagnosis of CVDs in India.

Methods: A total of 1670 anonymized medical records were collected from a tertiary hospital in South India. Seventy percent of the collected data were used to train the prediction system. Five state-of-the-art ML algorithms (k-Nearest Neighbours, Naïve Bayes, Logistic Regression, AdaBoost and Random Forest [RF]) were applied using Python programming language to develop the prediction system. The performance was evaluated over remaining 30% of data. The prediction system was later deployed in the cloud for easy accessibility via Internet.

Results: ML effectively predicted the risk of heart disease. The best performing (RF) prediction system correctly classified 470 out of 501 medical records thus attaining a diagnostic accuracy of 93.8%. Sensitivity and specificity were observed to be 92.8% and 94.6%, respectively. The prediction system attained positive predictive value of 94% and negative predictive value of 93.6%. The prediction model developed in this study can be accessed at http://das.southeastasia.cloudapp.azure.com/predict/

Conclusions: ML-based prediction system developed in this study performs well in early diagnosis of CVDs and can be accessed via Internet. This study offers promising results suggesting potential use of ML-based heart disease prediction system as a screening tool to diagnose heart diseases in primary healthcare centres in India, which would otherwise get undetected.

* *Corresponding author.*
E-mail address: ekta.marwaha@gmail.com (E. Maini).

https://doi.org/10.1016/j.mjafi.2020.10.013
0377-1237/© 2020 Director General, Armed Forces Medical Services. Published by Elsevier, a division of RELX India Pvt. Ltd. All rights reserved.

Please cite this article as: Maini E et al., Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India, Medical Journal Armed Forces India, https://doi.org/10.1016/j.mjafi.2020.10.013

Introduction

Cardiovascular diseases (CVDs) are the foremost reason of disease burden and mortality all over the world. Approximately 30% of total deaths (17.9 million) occurred due to CVDs globally in 2016.[1] The situation is critically serious in low- and middle-income countries like India. During the past three decades, the number of deaths due to CVDs has increased significantly from 15.2% to 28.1% in India.[2] Prevalence of CVDs was observed to be as high as 54.5 million cases in 2016.[3] CVDs are often detected in advanced stages amongst the underprivileged patients. Due to various reasons, Indian public healthcare system is still not capable in effectively preventing non-communicable diseases like CVDs. Efficient healthcare in terms of affordability, accessibility and quality is still far from being within reach of many.[4] Shortage of facilities in rural areas hampers medical diagnostic and therapeutic help in the initial stage of disease. Despite the government initiatives of health insurance (which are mainly for therapeutic care only) for poor people, the major section of Indian population does not have preventive health check-up benefits.[5] All these reasons lead to delayed treatment and increase in morbidity and mortality.[6]

Amalgamation of ML-based prediction system in primary healthcare centres can potentially aid hugely in the prevention of CVDs in India. Recent advancements in the field of computer science have proved that machine learning (ML) algorithms can generate huge meaningful information from the immense data generated by the healthcare sector.[7] This information can be used for the diagnosis of diseases at initial stage, which can thus aid in the prevention of diseases. ML-based tools for efficient healthcare are fetching a huge attention globally. As an example, ML-based tools have recently been successfully implemented in the fields of ophthalmology and oncology in the United States.[8,9] Available literature reveals the development of highly accurate prediction systems for CVDs.[10–13] All the research studies done for early detection of CVDs as reported in the literature so far are based on the freely available online data set provided by ML repository of University of California, Irvine.[14] This data set provides information about 76 medical attributes of 303 medical records gathered from hospitals of Western countries. Information obtained from diagnostic tests like electrocardiogram, treadmill test, fluoroscopy test etc. is available in the above-mentioned data set. However, these medical tests are neither accessible nor affordable to a major section of Indian population.[15] Thus, the prediction systems developed so far are not suitable for Indian population. Moreover, some of the primary risk factors responsible for heart diseases in India, like, obesity, lack of physical activity, physiological stress, smoking and alcohol consumption etc. have not been considered in the ML-based studies so far.

As the importance of early detection of CVD is increasingly being realized, there is a definite need of developing ML-based prediction system for CVDs specifically suitable for Indian scenario.

This study was carried out with the following objectives: a) Development of a high-performance and cost-effective ML-based heart disease prediction system using routine clinical data specifically suited for Indian population and b) Deployment of the prediction system in public cloud to ensure easy accessibility via Internet particularly beneficial for rural areas in India.

Table 1 – Description of attributes used in the study.

Attributes	Description	Categorical/ numeric
Age	Years	Numeric
Weight	Kilograms	Numeric
Height	Centimetres	Numeric
Total cholesterol levels	mg/dL	Numeric
Gender	Male/female	Categorical
Hypertension	Yes/no	Categorical
Diabetes	Yes/no	Categorical
Alcohol	Yes/no	Categorical
Smoking	Yes/no	Categorical
Exercise	Yes/no	Categorical
Stress	Yes/no	Categorical
Family history of cardiovascular disease (CVD)	Yes/no	Categorical
Healthy diet	Yes/no	Categorical
Risk of CVD	High/low	Categorical

Materials and methods

Study setting

This study is an interdisciplinary research work carried out by collaboration of data scientists and specialist doctors. The study was approved by institutional ethical committee. Members of ethical committee deemed that data privacy is ensured by using anonymized medical records of existing/retrospective cohort.

Data collection

By a random selection after applying the exclusion criteria, anonymized medical records of heart patients as well as of healthy persons were collected from a tertiary hospital in South India. Anonymization ensured data privacy as the personal details of the patients were not collected for the study.

Exclusion criteria: 1) Medical data sets corresponding to pregnant females, 2) patients reporting chronic kidney disease, severe mental illness, atrial fibrillation, 3) patients who reported the prolonged use of anti-depressants, antibiotics and medicines for asthma, tuberculosis and cancer, 4) patients who are prescribed oral corticosteroids, antipsychotic drugs and immunosuppressants and 5) patients younger than 20 years or older than 100 years.

After applying these exclusion criteria, the final data set comprised of 1670 medical records belonging to people between the age 30–79 years. Study population included 881 males and 789 females. Ethnicity of all records in this study was observed to be Asian. Eight-hundred and seventy-four records did not have

Table 2 – Study Population Descriptive Characteristics.

Risk factor attribute	Unit	Total records (n = 1670)		P-value
		Cardiovascular disease (CVD) (n = 893)	No CVD (n = 777)	
Age	Years (SD)	66.2 (11.2)	57.3 (12.4)	<0.001
Weight	Kilograms (SD)	85.4 (9.2)	69.4 (10.1)	<0.001
Height	Centimeters (SD)	165.7 (9.1)	162.3 (13.4)	0.23
Total cholesterol levels	mg/dL (SD)	267.7 (14.1)	218.4 (13.9)	<0.001
Gender (female)	N (%)	244 (27.3)	545 (70.1)	<0.001
Hypertension (yes)	N (%)	614 (68.7)	182 (23.4)	<0.001
Diabetes (yes)	N (%)	630 (70.5)	318 (40.9)	<0.001
Alcohol (yes)	N (%)	623 (69.7)	305 (39.2)	<0.001
Smoking (yes)	N (%)	570 (63.8)	258 (33.2)	<0.001
Exercise (yes)	N (%)	412 (46)	737 (94.8)	<0.001
Stress (yes)	N (%)	568 (63.6)	352 (45.3)	<0.001
Family history of CVD (yes)	N (%)	592 (66.2)	299 (38.4)	<0.001
Healthy diet (yes)	N (%)	496 (55.5)	398 (51.2)	0.077

*p-value < 0.05 is statistically significant.

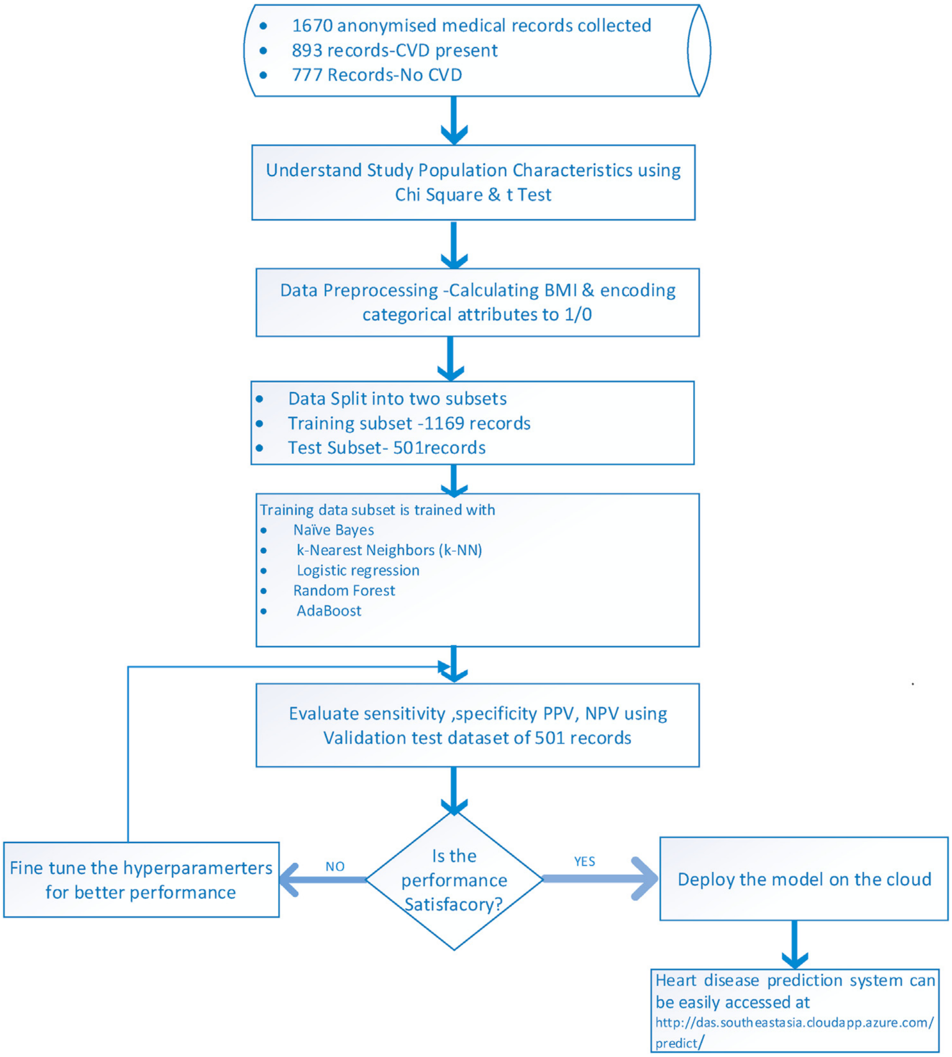

1670 anonymised medical records collected
893 records-CVD present
777 Records-No CVD
Understand Study Population Characteristics using Chi Square & t Test
Data Preprocessing -Calculating BMI & encoding categorical attributes to 1/0
Data Split into two subsets
Training subset -1169 records
Test Subset- 501records
Training data subset is trained with
Naïve Bayes
k-Nearest Neighbors (k-NN)
Logistic regression
Random Forest
AdaBoost
Evaluate sensitivity ,specificity PPV, NPV using Validation test dataset of 501 records
Is the performance Satisfacory?
NO
YES
Fine tune the hyperparamerters for better performance
Deploy the model on the cloud
Heart disease prediction system can be easily accessed at
http://das.southeastasia.cloudapp.azure.com/predict/

Fig. 1 – Workflow diagram of the study. This figure depicts the complete workflow of the study. The medical data set of 1670 records were gathered (in random fashion). Seventy percent data samples used to train the models. Test subset comprised the rest 30% of medical records. Five machine learning algorithms are applied to train the training subset. The prediction system was hosted on the public cloud for easy accessibility.

Please cite this article as: Maini E et al., Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India, Medical Journal Armed Forces India, https://doi.org/10.1016/j.mjafi.2020.10.013

Table 3 – Details of training and test subsets.

Class	Training subset (70%)	Test subset (30%)	Total records
High-risk cardiovascular disease (CVD)	656	237	893
Low-risk CVD	513	264	777
Total records	1169	501	1670

hypertension, while rest 796 reported hypertension. Of 1670 records, 928 reported to consume alcohol. Eight-hundred and twenty-eight records belonged to smokers. Nine-hundred and twenty records complained of stress and anxiety in life. Of 1670 records, 893 records (53.47%) were diagnosed with CVDs and remaining 777 records (46.53%) were of healthy persons with no CVDs. The persons who visited the hospital for routine check-ups and were not diagnosed with any heart disease are referred to as healthy persons (CVD risk: low) in this study.

Risk factor attributes

People living in rural parts of the country are usually unaware of the potential risk factors of heart diseases. They usually neglect the early signs of heart disease. Since the study has been carried out especially for rural areas, the clinical attributes already known to be the potential risk factors of CVDs along with lifestyle attributes associated with heart disease were chosen for this study. These attributes include age, gender, weight, height, total cholesterol levels, smoking habits, alcohol, diabetes, hypertension, family history of CVDs, intake of healthy diet, physical activity/exercise habits and stress/anxiety in life. Body mass index (BMI) was calculated internally by the software. Table 1 represents the details of risk factors considered in this study.

Diagnostic procedures like treadmill test and fluoroscopy (used extensively in similar studies done so far) were not considered relevant for this study to ensure cost-effectiveness. Tests for triglycerides, serum creatinine, C-reactive protein, serum fibrinogen, gamma glutamyl transferase, lipoprotein, apolipoprotein B, homocysteine, insulin test etc. although associated with the risk of heart diseases, were also not considered in this study as these medical tests are not feasible/affordable for rural population for which the research is aimed for.

Table 4 – Performance of machine learning algorithms on validation set of 501 records.

Algorithm	True positive	True negative	False negative	False positive	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy
k-Nearest Neighbours	211	230	26	34	89%	87.1%	86.1%	89.8%	88%
Naïve Bayes	210	232	27	32	88.6%	87.8%	86.7%	89.5%	88.2%
Logistic Regression	215	240	22	24	90.7%	90.9%	89.9%	91.6%	90.8%
AdaBoost	218	246	19	18	91.9%	93.1%	92.3%	92.8%	92.6%
Random Forest	220	250	17	14	92.8%	94.6%	94%	93.6%	93.8%

Study population characteristics

Out of 1670 records, 893 were positive cases of CVD while remaining 777 records were negative cases of CVD ensuring that the data set is balanced and is not skewed in favour of any class. In the data set, mean age of patients with heart disease is 66.2 years while mean age of healthy people was 57.3 years. Mean total cholesterol for healthy people was 188 mg/dL while the mean total cholesterol for heart patients was high at 237.7 mg/dL. Mean weight of heart patients was observed to be 85.4 kg while the mean weight of healthy people was 69.4 kg. It was observed that only 27.3% of heart patients were females. Nearly 95% of healthy people reported that they used to exercise regularly. Chi-square test of independence and t-test were carried out in the study subjects on 'prior basis' to determine the statistical significance of categorical and numeric input attributes, respectively, in determining heart disease.[16] These tests were used to ensure the validity of data of study variables, since the performance of AI algorithms is affected by the data of variables used to train the algorithms. Descriptive characteristics of these study population variables have been represented in Table 2.

Methodology

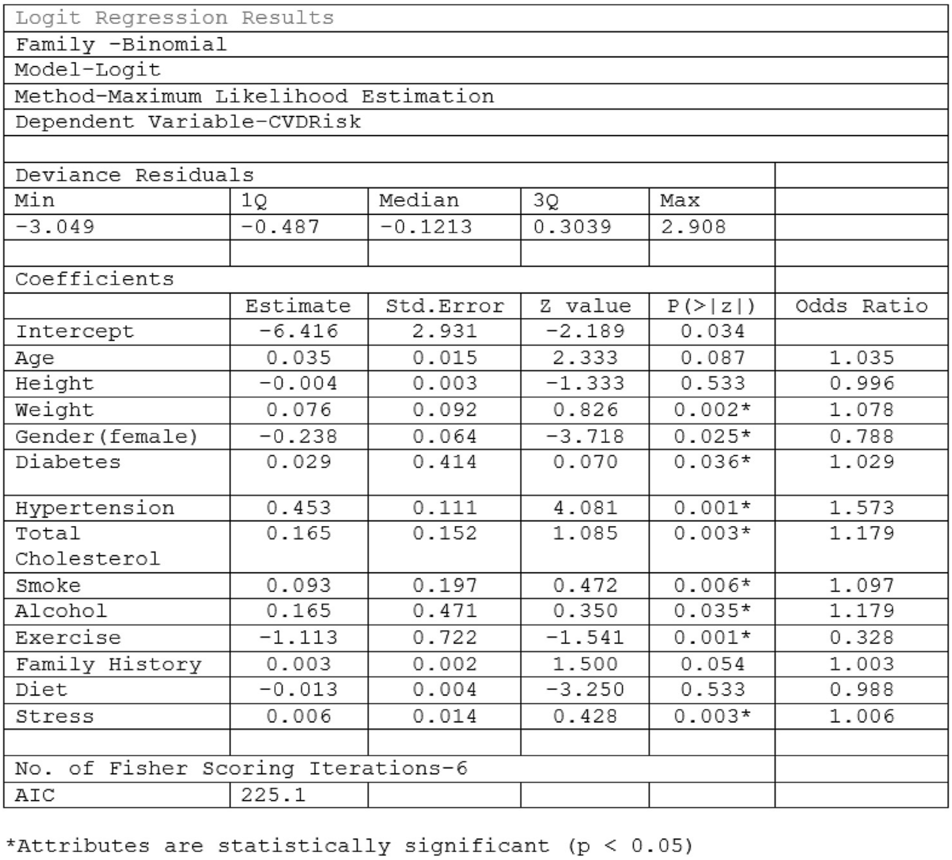
Python 3.7 programming language was used for building ML-based heart disease prediction system. Powerful software libraries supported by Python namely NumPy, Pandas, Seaborn, Statsmodels.api, SciPy and Sklearn etc. were used for exploratory analysis of data[17] and implementing five ML algorithms namely k-Nearest Neighbours (k-NN), Naïve Bayes (NB), Logistic Regression (LR), AdaBoost (AB) and Random Forest (RF). This study also has typical binary classification where 13 input attributes are observed to determine if there is a high risk of heart disease in a patient (risk of CVD = high) or not (risk of CVD = low). Fig. 1 shows the workflow diagram of complete project.

Data pre-processing

It was observed that there were no missing values or outliers in the data.

Since the ML algorithms can process only numerical data, the categorical attributes were label encoded. Gender female was encoded as 1 while male as 0. For all the other categorical variables like diabetes, stress, exercise etc., the presence (yes) was encoded as 1 while absence (no) was encoded as 0. High risk of CVD was encoded as 1 while low risk of CVD was encoded as 0.

Please cite this article as: Maini E et al., Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India, Medical Journal Armed Forces India, https://doi.org/10.1016/j.mjafi.2020.10.013



Logit Regression Results					
Family -Binomial					
Model-Logit					
Method-Maximum Likelihood Estimation					
Dependent Variable-CVDRisk					
Deviance Residuals					
Min	1Q	Median	3Q	Max	
-3.049	-0.487	-0.1213	0.3039	2.908	
Coefficients					
	Estimate	Std.Error	Z value	P(>|z|)	Odds Ratio
Intercept	-6.416	2.931	-2.189	0.034	
Age	0.035	0.015	2.333	0.087	1.035
Height	-0.004	0.003	-1.333	0.533	0.996
Weight	0.076	0.092	0.826	0.002*	1.078
Gender(female)	-0.238	0.064	-3.718	0.025*	0.788
Diabetes	0.029	0.414	0.070	0.036*	1.029
Hypertension	0.453	0.111	4.081	0.001*	1.573
Total Cholesterol	0.165	0.152	1.085	0.003*	1.179
Smoke	0.093	0.197	0.472	0.006*	1.097
Alcohol	0.165	0.471	0.350	0.035*	1.179
Exercise	-1.113	0.722	-1.541	0.001*	0.328
Family History	0.003	0.002	1.500	0.054	1.003
Diet	-0.013	0.004	-3.250	0.533	0.988
Stress	0.006	0.014	0.428	0.003*	1.006
No. of Fisher Scoring Iterations-6					
AIC	225.1				

*Attributes are statistically significant (p < 0.05)

Fig. 2 – Study population characteristics mean (standard deviation) of numerical attributes along with p-values of t-test to indicate the statistical significance for two groups: high risk/low risk of cardiovascular disease (CVDs). Count (%) of categorical attributes in two groups: high risk/low risk of CVDs.

Building the model

Using the train_test_split function supported by scikit learn library, the complete medical data set was randomly split into two portions in the ratio 70:30 referred as training and test/validation subset, respectively. Out of total 1670 records, training subset had 1169 records while test subset had 501 records. Detailed information about the training and test subsets is provided in Table 3. The total number of records in the training data set were 1169, of which 656 records correspond to CVDs while 513 records belonged to healthy people not diagnosed with CVDs.

ML algorithms with well demonstrated performance for classification namely NB, LR and k-NN were applied to build the prediction model.

Applying ensembling algorithms for better performance

Research has proved that the performance of a ML-based prediction system can be improvised using ensembling techniques.[18] Ensembling is a union of individual classifying algorithms. Bagging ensemble algorithms namely RF and boosting ensemble algorithms namely adaptive boosting AB were also implemented for enhanced performance.

Testing the performance of the model

The performance of prediction models developed using k-NN, NB and LR algorithms was analysed using the validation subset of 501 records as shown in Table 4. Of these records, 237 were confirmed cases of CVDs while remaining 264 records correspond to healthy people not diagnosed with CVDs. Prevalence of disease in validation subset was 237/501 = 47.3%

Analysis of confusion matrix is a standard way to check the performance of ML-based prediction system. Confusion matrix has four components namely true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs).

TPs: Heart patients who are predicted correctly to have heart diseases.

TNs: Healthy persons who are predicted correctly to be healthy.

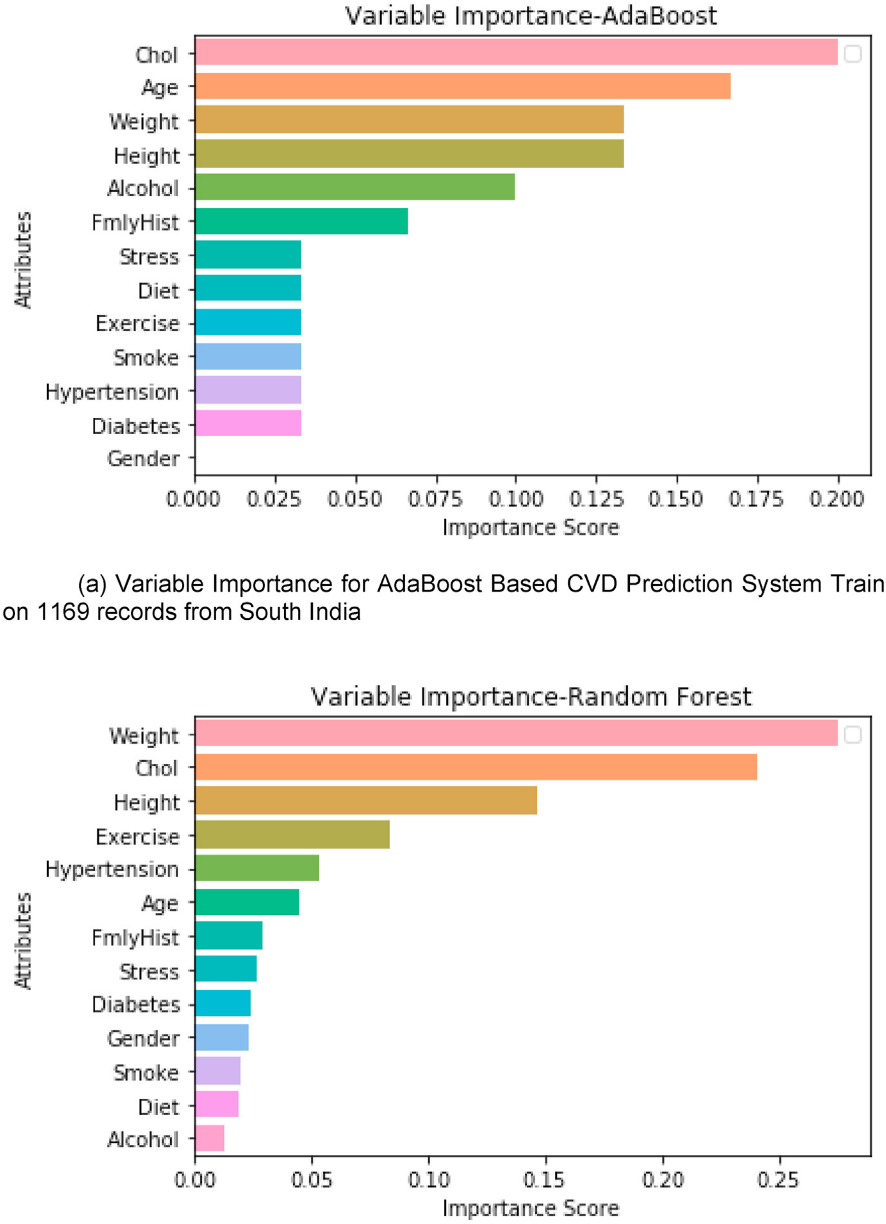
FPs: Healthy persons predicted incorrectly to have heart diseases (Type 1 error).

FNs: Heart patient predicted incorrectly to be healthy (Type 2 error).

These values are used to calculate accuracy, specificity, sensitivity, positive predictive value (PPV) and negative predictive value (NPV). PPV and NPV depend on the prevalence of disease.

A brief description of these parameters is given below.

i. Classification accuracy: This parameter represents that part of total predictions that were correct. Accuracy = (TN + TP)/(TN + FN + FP + TP)



(a) Variable Importance for AdaBoost Based CVD Prediction System Trained on 1169 records from South India

(b) Variable Importance for Random Forest Based CVD Prediction System Trained on 1169 records

Fig. 3 – Variable importance. (a) Variable importance for AdaBoost-based prediction model. (b) Variable importance for Random Forest–based prediction model.

ii. Sensitivity: This parameter reflects the ratio of cases that were accurately predicted with heart disease to the total number of actual cases of heart disease. Mathematically, sensitivity = TP/TP + FN

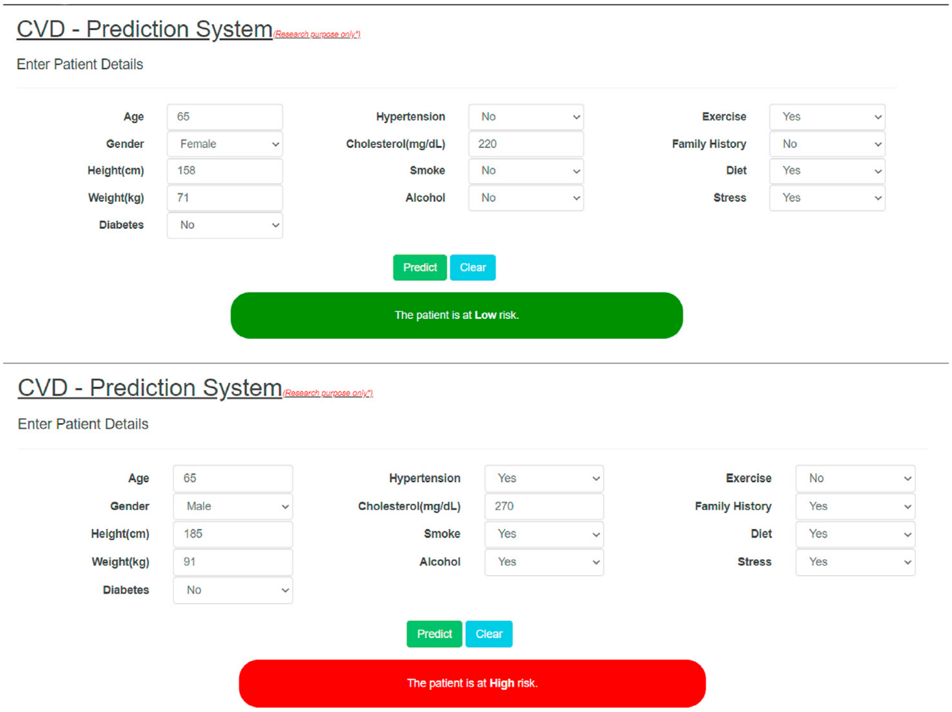
iii. Specificity: This parameter calculates the ratio of cases that are correctly predicted with no heart disease to the entire count of actual cases with no heart disease. Mathematically, Specificity = TN/FP + TN

iv. PPV: This parameter reflects the ratio of cases that are correctly predicted with heart diseases to the total count of cases predicted to have heart disease. Mathematically, PPV = TP/TP + FP

v. NPV: This parameter reflects the ratio of cases correctly predicted to be healthy to the total count of cases predicted to be healthy. Mathematically, NPV = TN/TN + FN

Fine tuning of hyperparameters

Grid Search for cross-validation was used to identify the best hyperparameters for the learning algorithms. Grid Search CV class from sklearn library was used for this purpose.



CVD - Prediction System (Research purpose only*)

Enter Patient Details

Age 65
Gender Female
Height(cm) 158
Weight(kg) 71
Diabetes No
Hypertension No
Cholesterol(mg/dL) 220
Smoke No
Alcohol No
Exercise Yes
Family History No
Diet Yes
Stress Yes

The patient is at **Low** risk.

CVD - Prediction System (Research purpose only*)

Enter Patient Details

Age 65
Gender Male
Height(cm) 185
Weight(kg) 91
Diabetes No
Hypertension Yes
Cholesterol(mg/dL) 270
Smoke Yes
Alcohol Yes
Exercise No
Family History Yes
Diet Yes
Stress Yes

The patient is at **High** risk.

Fig. 4 – Using cardiovascular disease (CVD) prediction model to test the risk of CVDs. The medical practitioner enters the patient's clinical parameters as well as attributes related to his lifestyle to predict the risk of CVD.

Deployment on the public cloud

The best performance prediction system built using RF model was deployed in Microsoft Azure cloud for better accessibility.[19] '*Pickle*' and '*Flask*' software libraries of Python programming language were used for this purpose.[20] Hosting the prediction system on cloud enables it to be easily accessed from anywhere in the world via Internet. This is highly useful feature for healthcare sector of India, which faces the major issue of shortage of medical facilities especially in rural areas. Accessing this prediction system is as easy as accessing an e-mail via Internet.

Results

CVD prediction system was developed by applying five well-established ML algorithms on the training data set. The performance was tested on the validation test set of 501 records. Prevalence of disease in validation subset was 237/501 = 47.3% Performance metrics namely accuracy, sensitivity, specificity, PPV and NPV were calculated for each algorithm. The performance results of all classifiers are given in Table 4.

The best hyperparameters for k-NN (n_neighbors = 12) resulted in a performance of sensitivity 89%, specificity 87.1%, PPV 86.1%, NPV 89.8%. The performance of NB was found to better than k-NN. Sensitivity 88.6%, specificity 87.8%, PPV 86.7%, NPV 89.5% were achieved by NB.

LR with hyperparameters (C = 1, penalty = l2) performed well in classifying people with low risk or high risk of CVDs. LR correctly classified 455 out of 501 records, thus attaining a classification accuracy of 90.8%. Sensitivity 90.7% and specificity were 90.7% and 90.9%, respectively. PPV was observed to be 89.9% while NPV was 91.6%.

Models built using ensemble techniques (RF and AB) performed better than LR. AB model was trained with Stage-wise Adaptive Modelling using a Multi-class Exponential loss function (n_estimators = 30) while RF based on 'gini index' with n_estimators = 150 resulted in the best performance. Sensitivity and specificity of AB model was 91.9% and 93.1%, respectively, while RF reported 92.8% sensitivity and 94.6% specificity. PPV 94% and NPV 93.6% were achieved by RF–based prediction model.

Interpretation of ML-based models is not easy, and these are usually considered as 'black boxes. However, logistic regression–based models are quite interpretable. Logistic regression was implemented using the Logit function (Binomial family) based on maximum likelihood estimation method to predict CVD risk using statsmodels.api library of Python. Fig. 2 shows the summary of results obtained.

Male gender, diabetes, hypertension, high cholesterol level, smoking and alcohol were significantly associated with CVD. Lack of exercise and stress were observed to be more prevalent in CVD group (p value < 0.05).

Estimate column in the summary reflects the natural logarithm of odds ratio of getting diagnosed with high risk of heart disease keeping all other features constant. Due to negative values of log (odds ratio) it is inferred that females had a low risk of CVDs compared with males. Regular exercise and intake of healthy diet were observed to be associated with low risk of CVDs; on the other hand, diabetes, hypertension, stress, smoking and family history tend to result in high risk of CVDs.

Please cite this article as: Maini E et al., Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India, Medical Journal Armed Forces India, https://doi.org/10.1016/j.mjafi.2020.10.013

The odds ratio column in the summary suggests how the odds ratio of being detected with high risk of CVD change if all other attributes are kept constant. Hypertension tends to increase the odds ratio of high risk of CVDs by 1.573 while the odds ratio drops significantly to 0.328 with regular physical exercise. Odds ratio of high risk of CVD for females is 0.788 compared with males.

Ensemble algorithms (RF and AB) are based on decision trees and attribute importance is graded according to selection occurrence frequency of an attribute as a decision node decided based on information gain and entropy. Variable importance for boosting algorithm was decided based on the impurity-based scores using feature_importances_ from sklearn library of Python. Attributes exercise, weight, total cholesterol, hypertension and age were the top five important attributes for AB algorithm. In case of RF prediction system, variable importance scores for attributes weight, exercise, total cholesterol, hypertension, and gender were found to be maximum for predicting CVDs. Variable importance for AB algorithms and RF is represented graphically in Fig. 3(a) and (b), respectively.

RF-based CVD prediction model (trained on 1169 records and tested on 501 records) is hosted on cloud and can be easily accessed at das.southeastasia.cloudapp.azure.com/predict/

The input attributes of the patient are entered into the system. The system predicts if the patient has low risk of CVDs or high risk. Sample screenshots of the result obtained using the prediction system are shown in Fig. 4.

Discussion

In the recent years, substantial research studies have been carried out to build methods for diagnosing heart diseases in early stages. Various feature selection techniques were applied in the research carried out by Takci.[21] (2018), and the resulting prediction system attained an accuracy of 84.81%. Similar study was carried out by Kausar et al. and an accuracy of 88.41% was obtained.[22] Prediction system developed by Khalid Raza using ensembling technique (2019) attained an accuracy of 88.88%.[23] A similar accuracy level of 89% was achieved by the prediction system developed by Haq et al. in 2019.[24] Using artificial neural network to design a prediction system Alic et al. achieved an accuracy of 91% in their research study.[25] But importantly, the prediction system developed in all of these studies do not work effectively well for Indian population as these models are based on data collected from Western countries and do not take into consideration lifestyle-related risk factors responsible for CVDs (lack of physical activity, family history, alcohol etc.). Moreover, these systems rely on the results of medical tests like ECG, treadmill test, fluoroscopy tests etc., which are not feasible in Indian primary health centres in the existing scenario.

The accuracy attained in the present study is 93.8%. The prediction system developed in this research uses 13 clinical parameters and identifies the risk of a person to have heart disease. Compared with the studies done so far, this study has been carried out on Indian population, and the potential risk factors like high body weight, lack of exercise, psychological stress, family history, smoking and alcohol consumption habits have been considered in this study (unlike the studies quoted previously). It is worth noting that the system developed in this study is highly cost-effective compared with earlier studies as expensive tests like fluoroscopy and treadmill tests have not been taken into consideration. Easy accessibility of the prediction system via Internet is also an added remarkable feature of this study, which was not reported by earlier studies. It is worth mentioning here that prediction model developed in this pilot study predicts output depending on the study population attribute trends it was trained on. Once the ML models are trained and tested on voluminous data sets, it can be used as a screening tool in rural India and can help in the prevention of CVDs.

Cost-effectiveness, excellent performance and easy accessibility of the prediction system via Internet defend the use of ML-based prediction system as a screening tool for CVD detection in India.

To the best of our knowledge, this study was first of its kind in Indian context. Developed countries like the United Kingdom and the United States are investing their resources to carry to research for developing ML-based prediction models for diagnosing heart diseases in primary healthcare centres.[26,27] It is recommended that similar studies should be promoted in India. The current national health policy (2017) of our government, laying stress on preventive health will be more meaningful and fruitful if advancement in this field is made as early as possible.[28] We propose larger studies of multicentric nature for development of AI prediction systems for CVD screening in our country, which is facing ever increasing load of morbidity and mortality due to CVD being detected in late advanced stages. Premier institutes of medicine and technology can collaborate in this regard to diagnose other lifestyle diseases and non-communicable diseases like malignancies. Cardiological Society of India (CSI) can help in this regard. Other modern techniques like artificial neural networks can be applied to further improve the performance of the system.

Limitations of the study

This study used a data set of 1670 patients reporting to a tertiary care private setup in a south Indian metropolitan city where largely the higher income group seeks the medical care. This potentially may seem biased in reader's mind, but this study was aimed only to detect the robustness of a prediction model based on ML. The results obtained from the prediction system developed in this study are based on the attribute trends of the study population on which the model is trained on. In future the model needs to be trained on huge data sets collected from diverse regions before using it as a screening tool.

Conclusions

The study portrays the capability of ML algorithms to predict CVDs in Indian population. Issues of affordability and accessibility in healthcare sector of India can be addressed using

Please cite this article as: Maini E et al., Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India, Medical Journal Armed Forces India, https://doi.org/10.1016/j.mjafi.2020.10.013

ML-based models, which can be easily accessed via Internet even in the rural parts of the country. It is proposed to build and test the performance of similar systems using voluminous cardiac data sets belonging to all economic sections of the society collected from various regions of India. We recommend similar studies of multicentric nature across entire country. To achieve the sustainable development goals laid down by World Health Organization, it is high time, we as a country do take timely advantage of ML-based prediction systems in improving preventive care aspect of public healthcare system.[29]

What is already known?

ML-based tools have shown remarkable performance in diagnosing various serious diseases in initial stages in healthcare centres of developed countries.

What does this study add?

An indigenous high-performance ML-based CVD prediction system easily accessible via Internet is proposed for existing Indian healthcare system. Healthcare in India can be made more affordable and accessible using ML-based prediction systems.

Disclosure of competing interest

The authors have none to declare.

Acknowledgements

The authors express their heartfelt gratitude to Sagar Hospitals, Jayanagar, Bengaluru, for providing anonymized information of patients' health parameters for carrying out this study. No funding was received for this project.

REFERENCES

1. [Internet]. *Noncommunicable Diseases Country Profiles*. World Health Organization; 2018, 2019 [cited 17 December 2019]. Available from: https://www.who.int/nmh/publications/ncd-profiles-2018/en/.
2. [Internet]. Healthdata.org [cited 17 December 2019]. *Institute for Health Metrics and Evaluation (IHME). Findings from the Global Burden of Disease Study 2017*. Seattle, WA: IHME; 2018. Available from: http://www.healthdata.org/sites/default/files/files/policy_report/2019/GBD_2017_Booklet.pdf;2019.
3. Prabhakaran D, Jeemon P, Sharma M, et al. The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990–2016. *Lancet Glob Health*. 2018. https://doi.org/10.1016/s2214-109x(18)30407-8.
4. Kasthuri A. Challenges to healthcare in India - the five A's. *Indian J Community Med*. 2018;43(3):141–143. https://doi.org/10.4103/ijcm.IJCM_194_18.
5. George A, Badagabettu S, Berra K, George LS, Kamath V, Thimmappa L. Prevention of cardiovascular disease in India: barriers and opportunities for nursing. *J Clin Prev Cardiol*. 2018;7:72–77.
6. Sangar S, Dutt V, Thakur R. Why people avoid prescribed medical treatment in India? *Indian J Publ Health*. 2019;63:151–153.
7. Maini Ekta, Venkateswarlu Bondu. Artificial intelligence-futuristic pediatric healthcare. *Indian Pediatr*. 2019;56:796.
8. Van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2017;96(1):63–68. https://doi.org/10.1111/aos.13613.
9. [Internet]. ibm.com [cited 17 December 2019]. *IBM Watson Health in Oncology. Scientific Evidence 2019*; 2019. Available from: https://www.ibm.com/downloads/cas/0ZRYPWL9.
10. Alexander CA, Wang L. Big data analytics in heart attack prediction. *J Nurs Care*. 2017;6(2). https://doi.org/10.4172/2167-1168.1000393.
11. Maini E, Venkateswarlu B, Gupta A. Applying machine learning algorithms to develop a universal cardiovascular disease prediction system. In: *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. 2018:627–632.
12. Shafenoor Amin M, Kia Chiam Y, Dewi Varathan K. Identification of significant features and data mining techniques in predicting heart disease. *Telematics Inf*. 2018. https://doi.org/10.1016/j.tele.2018.11.007.
13. Maini E, Venkateswarlu B, Gupta A. Determination of significant features for building an efficient heart disease prediction system. *Int J Recent Technol*. 2019;8(2):4500–4506.
14. [Internet]. Archive.ics.uci.edu. *UCI Machine Learning Repository: Heart Disease Data Set*; 2019 [cited 17 December 2019]. Available from: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
15. [Internet]. The Times of India. *Seven Tests to Diagnose Heart Diseases- the Times of India*; 2020 [cited 4 January 2020]. Available from: https://timesofindia.indiatimes.com/7-tests-to-diagnose-heart-diseases/listshow/43215326.cms.
16. [Internet]. Techopedia.com. *Why is Python so popular in machine learning?*; 2020 [cited 23 January 2020]. Available from: https://www.techopedia.com/why-is-python-so-popular-in-machine-learning/7/32881.
17. [Internet]. Learntech.uwe.ac.uk. *Data Analysis - Pearson's Correlation Coefficient*; 2020 [cited 23 January 2020]. Available from: http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442.
18. Gupta N, Ahuja N, Malhotra S, Bala A, Kaur G. Intelligent heart disease prediction in cloud environment through ensembling. *Expet Syst*. 2017;34, e12207. https://doi.org/10.1111/exsy.12207.
19. [Internet]. Salesforce UK Blog. *Why Move to the Cloud? 10 Benefits of Cloud Computing*; 2020 [cited 23 January 2020]. Available from: https://www.salesforce.com/uk/blog/2015/11/why-move-to-the-cloud-10-benefits-of-cloud-computing.html.
20. [Internet]. Medium. *Flask and Heroku for Online Machine Learning Deployment*; 2020 [cited 23 January 2020]. Available from: https://towardsdatascience.com/flask-and-heroku-for-online-machine-learning-deployment-425beb54a274.
21. Takci H. Improvement of heart attack prediction by the feature selection methods. *Turk J Electr Eng Comput Sci*. 2018;26:1–10.

22. Kausar N, Abdullah A, Samir B, Palaniappan S, AlGhamdi B, Dey N. Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. *J Med Imag Health Inform*. 2016;6(1):78–87.
23. Raza K. *Improving the Prediction Accuracy of Heart Disease with Ensemble Learning and Majority Voting Rule. U-Healthcare Monitoring Systems*. 2019:179–196. https://doi.org/10.1016/b978-0-12-815370-3.00008-6.
24. Haq A, Li J, Memon M, Nazir S, Sun R. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Inf Syst*. 2018;2018:1–21.
25. Alic B, Gurbeta L, Badnjevic A. Machine learning techniques for classification of diabetes and cardiovascular diseases. In: *2017 6th Mediterranean Conference on Embedded Computing (MECO)*. 2017.
26. Weng S, Reps J, Kai J, Garibaldi J, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS One*. 2017;12(4), e0174944.
27. Ambale-Venkatesh B, Yang X, Wu C, et al. Cardiovascular event prediction by machine learning. *Circ Res*. 2017;121(9):1092–1101.
28. [Internet]. nhp.gov.in [cited 14 january 2020]. *National Health Policy 2017*; 2020. Available from: https:www.nhp.gov.in/nhpfiles/national_health_policy_2017.pdf.
29. [Internet]. Who.int.. *From MDGs to SDGs, WHO Launches New Report*; 2020 [cited 14 January 2020]. Available from: https://www.who.int/en/news-room/detail/08-12-2015-from-mdgs-to-sdgs-who-launches-new-report

Please cite this article as: Maini E et al., Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India, Medical Journal Armed Forces India, https://doi.org/10.1016/j.mjafi.2020.10.013