Identification of Cardiovascular diseases (CVDs) using machine learning and analysis on risk factors for CVD

Vikramjit S. Dhaliwal[1] and Arunjeet Kaur[2]
[1,2]Fortis Hospital Mohali, India
[1]Vsd12111973@gmail.com
[2]arunjeet.kaur@fortishealthcare.com

Abstract

Background: Cardiovascular diseases (CVDs) comprise a range of dis-eases impacting heart and blood vessels. The early detection of CVDs in high-risk patients can certainly help them to recover from fatal diseases and to resume their routine life. The healthcare industry has repository of medical data which is sufficient enough to make predictions of cardiovascular diseases using machine learning algorithms at early stages. In this research paper, the data has been collected from Fortis healthcare centres. The data pre-processing techniques have been used to remove noisy data, to handle missing values, fill-ing default values if applicable and then machine learning based algorithms have been applied to predict whether the patient is suffering from Cardiovas-cular diseases or not. We have also presented analysis on the risk attributes with respect to the sample population in this paper.

Results: The performance of the ML based algorithms have been evaluated using confusion matrix, ROC, F1 score, precision and recall scores. The potential attributes have been highlighted and discussed that contribute in the occurrence of CVDs amongst the patients. We have analyzed the risk-factors such as consumption of alcohol and tobacco, blood pressure levels, blood sugar levels, cholesterol measures, other medical history such as diabetes, family his-tory, age and weight of the diseased, and measure of per day physical activity for analyzing the impact of each risk factor on the CVD patients.

Conclusions: In this paper, we have collected data of 70,000 patients from North-India region for the research study. We have attempted to apply classification algorithms for the identification of the patients suffering from cardiovascular diseases. It is observed that automated machine learning techniques produce accurate results and can assist the doctors to diagnose the diseased patients in a fast manner. We have also analyzed the risk-factors

related to CVDs such as diabetes, consumption of alcohol, smoking, obesity, blood pressure, cholesterol levels, physical activity and family history with chronic CVDs. We have also observed the impact of each respective risk-attribute with respect to cardiovascular diseases. This study will assist the physicians for identifying cardiovascular diseases in patients at early stage using proposed automated system and in analyzing the risk-factors with respect to CVDs. This study also provides insights into gender-wise analysis of cardiovascular cases.

Keywords : Heart attack, Congenital heart disease, Cardiovascular diseases, Stroke, Machine learning, CVD.

1 Background

Cardiovascular diseases (CVDs)

Cardiovascular diseases are the conditions affecting the blood vessels or hearts [1], [2], [3]. CVDs are amongst the potential diseases that affect numerous people in India [4], [5]. CVDs are prevalent in people who have bad life styles such as smoking, drinking and bad dietary habits [6], [7]. The population of millions died of CVDs around the world every year. When fatty deposits are found inside the arteries, these deposits give rise to blood clots [8]. The CVDs can also damage the

arteries. Nowadays CVDs are among the diseases that causes deaths and disability in India. In this section, we are proving in-sights into the types of CVDs, causes of CVDs, machine learning and CVDs, existing work, highlights of our research contribution and brief description on the structure of the paper.

Types of CVD: There are mainly 4 types of CVDs as are described below.

1. *Coronary heart disease:* It occurs during the reduction or blockage of ow of blood rich in oxygen to the heart muscle which results into strain on the human heart, and can cause:

   (a) Angina Chest pain : it happens due to low flow of blood to the heart muscles.

   (b) Heart failure: The condition when heart is not able to supply enough blood to the body.

   (c) Heart attacks: The condition when ow of blood to the heart gets blocked suddenly.

2. *Strokes and transient ischaemic attacks:* A stroke is a condition when the supply of blood to the brain gets restricted that as a result the brain gets damaged and possibly a person can lose his life [9]. A transient ischaemic attack is also a similar condition, but the supply of blood to the brain is interrupted temporarily.

3. *Peripheral arterial disease(PAD):* PAD occurs the blockage is found in the arteries connected to the limbs and resultant into the following:

   (a) hair loss on the feet and the legs.

   (b) numbness in the legs.

   (c) pain in legs from mild to severe.

4. *Aortic disease:* These are a group of diseases affecting the aorta (the largest blood vessel supplies blood from heart to other body parts).

Causes of CVD: There exist many factors that can cause CVDs in human beings as outlined below.

- *Smoking:* The usage of tobacco is a potential risk factor for CVD. It can narrow and damage blood vessels.

- *High blood pressure (BP):* High BP is another significant risk factor; high BP can also damage blood vessels.

- *High cholesterol:* It is a fatty substance which can increase the risk of development of blood clot.

- *Diabetes:* It is an ailment that keeps blood sugar levels very high which can narrow down the blood vessels. People with diabetes type-2 are mostly obese and obesity is also one of the risk factor for CVD.

- *Inactivity:* Regular walking and exercise are key factors of good health. Inactive person can have high BP, high cholesterol and obesity.

- *Obesity:* It increases the risk of having diabetes, high BP, and ultimately CVDs.

- *Family history:* The family history with CVD, is also a risk factor for having CVD to next generation. The people with family history should take more preventive measures.

CVDs are increasing at a rapid rate among the population of developing countries due to bad lifestyle and bad dietary habits. The report generated by WHO states that CVDs are responsible for 34% mortalities in India during 2019. The Global Burden of cardiovascular diseases states that the mortality rate is 276 per 100,000 people at present in India whereas it is 235 per 100,000 people in rest of the countries. The prevalence of CVDs have increased by 51% till 2019 from 1990 in India. Hence, it is important to propose an automated system that can diagnose CVDs at early stages and can predict from the attributes of a person whether he/she can suffer from CVDs in future. It is also necessary to study the attributes responsible for cardiovascular diseases to create awareness among the patients by showing them the impact of attributes such as smoking and diabetes on CVDs.

To predict whether a person is Non-CVD or a CVD patient is not a chal-lenge for doctors but to make them aware how their life style can make change in their lives is a challenge of this era. In order to analyze the impact of risk factors of CVDs and to apply machine learning techniques to predict whether the patient is suffering from CVD, we have collected primary data of the patients visited at Fortis Healthcare from 2014 to 2019. We have attempted to make use of machine learning automated techniques to classify the cardiovascular diseases from the featureset of the patients and we have also analyzed the impact of each attribute considered for study on the patients with respect to cardiovascular diseases. The proposed mechanism also predicts whether the patients with particular attributes can have CVD in future. We believe that our system would help in classifying the patients with CVDs and would assist the patients to escape from fatal diseases by predicting whether the patient can develop CVD in future from the featureset of a patient.

This paper is organized into five sections. The first section of the paper provides background details of Cardiovascular diseases and also highlights the existing work in the respective eld. The second section elaborates the proposed methodology and research design. The third section discusses about results obtained and analysis of the attributes affecting CVDs. The fourth section of this paper concludes our work and also provides future directions in the area of our research work.

1.1 Related Work

The research studies with respect to cardiovascular diseases in India may have slight differences in the outcomes due to different sample sizes and different regions selected for the study. We are attempting to present the existing literature in this section where machine learning and computational methods have been used for the detection and diagnosis of cardiovascular diseases.
These are many approaches available in the literature where people from technical background have proposed automated solutions by using machine learning and soft computing but the analysis of attributes is missing which affect CVDs. The works presented in [10] and [11] are based on neural net-works where authors have proposed computerized automated systems for the prediction of CVDs. In [12], neural networking based techniques are analyzed for the prediction of CVDs. In [13], authors have presented a classification approach for ECG Arrhythmia by making use of RNN. In [14], authors have presented comparative study on computerized techniques for CVD diagnosis.

In [15], authors have used data mining methods for the diagnosis of CVDs. In [16], authors have analyzed various data mining techniques for the diagnoses of CVDs. Dinesh et al. [17] have presented ML based methods for CVD diagnosis. Chen et al. [18] have proposed a model to identify CVD risk factors. Latha and Jeeva [19] have used ensemble classifiers for the detection of CVDs. Jain and Singh [20] have presented a review on CVD detection techniques. In [21], authors used deep learning (DNN) based approach for CVD classification. Other authors have also proposed techniques for classification of heart diseases as cited in [22], [23], [24], [9].

Though there exist many approaches to classify the CVDs and predicting the heart problems but with the advent of machine learning automated techniques, the classification of diseases have become easier and accurate. There is still a need to explore newer algorithms which are capable

enough to segregate the Cardiovascular diseases to diagnose the disease at early stage and to predict the disease in an accurate way. Hence, we are proposing dynamic machine learning based approaches to classify the CVDs.

2 Methods

Aim of the research study

- The first objective is to study the risk-factors related to CVD and analyze the impact of each risk factor on cardiovascular diseases.

- The second objective is to make use of machine learning based approaches for predicting the presence of cardiovascular diseases in patients based on the training dataset considered for research study.

2.1 Classification of CVDs using ML algorithms

Machine Learning approaches are being used to solve problems of various do-mains in real world scenarios. There is no exception in case of healthcare industry where artificial intelligence and machine learning techniques are playing a major role in prediction the diseases, performing surgeries and recommending the medicines to the diseased patients. We are making use of machine learning based approaches in our case study to predict the presence of Cardiovascular diseases in patients. Most of the CVDs can be identified in advance by using ML techniques; the healthcare physicians can prescribe treatment on patient to patient basis in a fast manner and can save the lives of patients who are suffering from fatal CVDs. The data used in this clinical research is original data collected at various Fortis healthcare hospitals in Northern India. In the given methodology, first of all the collected data has been processed to make it usable from our machine learning based model. The data visualization graphics have been used in our research work to see whether the data is balanced or not.

The parameters for the clinical study are mentioned in Table 1.

S.No	Attribute	Description
1.	Age	Age of the patient
2.	Height	Height of the patient
3.	Weight	3 categories,0: normal, 1: more than normal, 3: obese
4.	Gender	Specifies categorical code - male or female
5.	ap hi	It specifies systolic blood pressure
6.	ap low	It tells diastolic blood pressure
7.	Glucose	3 categories,1: normal, 2: more than normal, 3: very high
8.	Smoking	Represents binary feature
9.	Alcohol intake	Specifies binary feature
10.	Physical activity	Depicts binary feature
11.	Family history of CVDs	represents binary feature
12.	Cholesterol	3 categories,1: normal, 2: more than normal, 3: very high
13.	Diabetes	3 categories,1: type-1, 2: type-2, 0: normal blood sugar

Table 1: Parameters considered for classification of Cardiovascular diseases

The dataset obtained from Fortis repository has been used for the experiments which comprises 70,000 records. We have used 65% data for training the algorithms and 35% data for testing the ML based system.

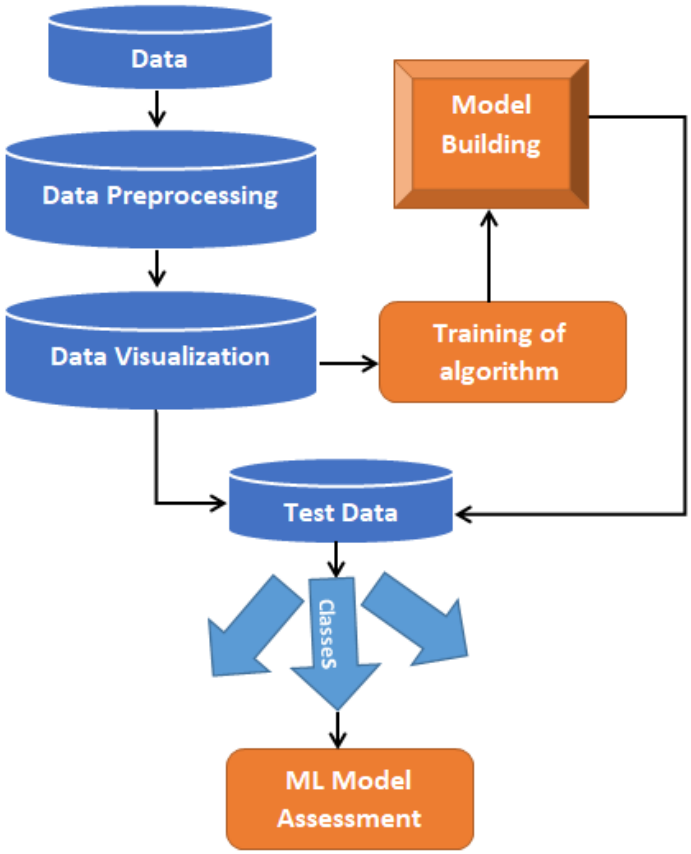

Fig. 1: Proposed ML model for classification of CVDs

Our ML based classification system and remained 35% for the testing of the proposed model. Patients considered under this case study are from 30 to 85 years of age. The male patients are nominated with gender value 1 and on the contrary, females by gender value 2. The next parameter is the reading of the blood pressure (BP), the systolic blood pressure range should be below 140, if its more than or equal to 140, the patient is suffering from high blood pressure. High BP is a condition in which the force of blood against the artery walls is quite high that it can cause Cardio-vascular diseases. The diastolic BP, is the pressure when the heart takes rest between the beats. The diastolic BP should be 80 and higher reading means a patient has high BP and can develop CVDs in future. The next parameter considered for research study is cholesterol. Total cholesterol should be lesser than 200 mg/dL and a measurement between 200 and 239 mg/dL can be seen as borderline and measurement over 240 mg/dL is considered high and can become a potential reason to get cardiovascular diseases.

Glucose levels are taken in 3 categories and assigned a value as 1 if the sugar is below 100 mg/dl, assigned value 2 if sugar level is below 140 and assigned value 3 (indicated very high) if the sugar level is above 140 mg/dl. Smoking parameter is based on binary values, if the value is 0 then the patient is not smoking and if the value 1, it indicates that the person is a smoker. Alcohol parameter is also based on binary values, value 0 indicates the person is non-alcoholic and value 1 indicates that the person is alcoholic. Next parameter is family history with binary values. If the patient's father or mother has a history of CVDs then there are likely chances that the patient may also suffer

from CVDs in future. Family history parameter is also based on binary values 0 for no history of CVD and 1 for presence of CVD in family history.

2.2 Classification using ensemble ML algorithms

We have used supervised learning methods for our problem statement. This paper proposes supervised ML procedure for the prognosis of cardiovascular diseases and an attempt is made to add some innovation in the existing algorithms for improving upon the classification accuracy. The supervised ML based procedures such as XGBoost and SVM with different kernals have been implemented with modifications for classifying the patients suffering from CVDs. The performance of the supervised ML techniques is evaluated by using confusion matrix, accuracy score, receiver operating characteristic (ROC) curve, sensitivity score, and F1 Score.

SVM Classifier: The first algorithm applied by us for classification of CVDs is binary SVM (Support Vector Machines). SVM is a supervised ML classifier where each data item is pointed as a point in n-dimensional space with the value of each feature. Then, the classification is performed by finding the hyper-plane that is able to differentiate the classes. Our objective is to test robustness of various kinds of kernels for binary SVM classifier and to find a plane that has the maximum margin from hyperplane. Support vectors impact the orientation as well as position of the hyperplane and basically they represent data points that are nearer to the hyperplane. In our case study, binary SVM is classifying the data into two classes and the classifier is using several kernels namely, linear, rbf, poly, and sigmoid; and the accuracy obtained by linear kernal is 81%, poly kernal is 71%, rbf kernal is 83% and sigmoid kernels is 79%. On the basis of kernal, the hyperplane is decided. The performance evaluation of rbf kernal based SVM is presented using confusion matrix as shown in Fig. 2, and the other parameters such as precision, recall and F1 score are shown in Table 2.

Random Forest (RF) Classifier: The next algorithm we have selected for CVD classification problem is RF which is also known as ensemble learning technique. The RF technique provides the classification solution in an improved manner over the bagged trees since it de-correlating the trees and ultimately reduces the variance ate the time of averaging the trees. The steps for RF approach are mentioned below:

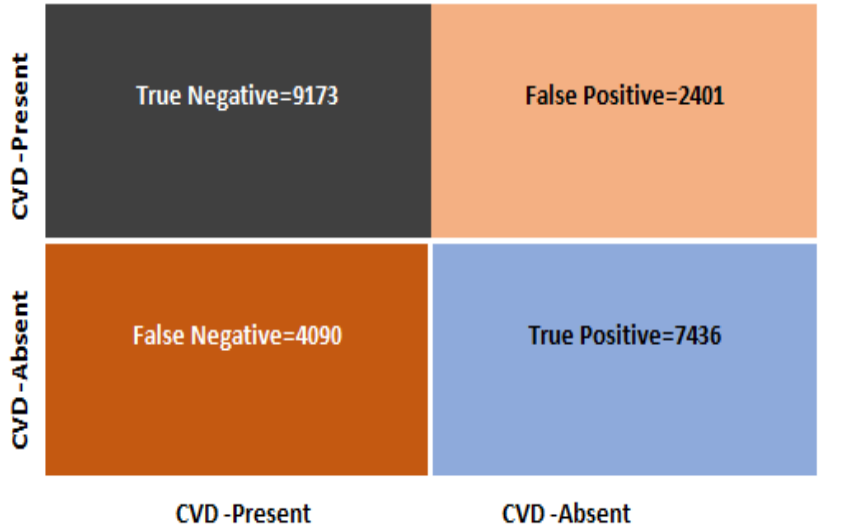

Fig. 2: Confusion matrix for SVM based classification with Rbf Kernal

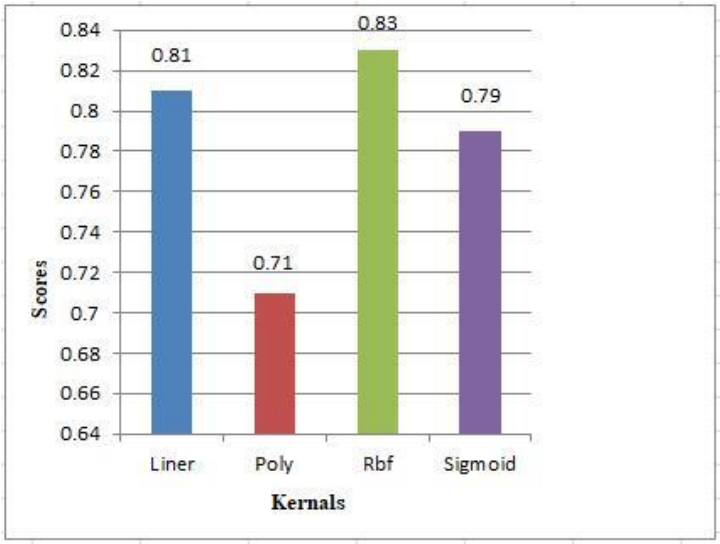

Fig. 3: SVM based classification with different Kernals

- Compute the Information Gain (IG), Chi-Square (CS) scores for each CVD attribute in the original dataset.

- Assign weights: Allocate wcs (weighted Chi square) as weights for importance assigned to Information gain, Chi-square for classification.

- Compute the Information Gain (IG), Chi-Square (CS) and CFS scores for each CVD attribute in the original dataset.

- Get the ranked CVD feature set based on weighted importance of IG, CS and CFS.
- Apply Random forest to ranked CVD feature set
- Validate the classification outcome.

- optimize the outcome till get best classification results

The performance of the RF based classifier can be observed at Fig. 4 in terms of confusion matrix and the other parameters are shown in Table 2. The random forest scores obtained from different number of estimators is shown in Fig. 5.

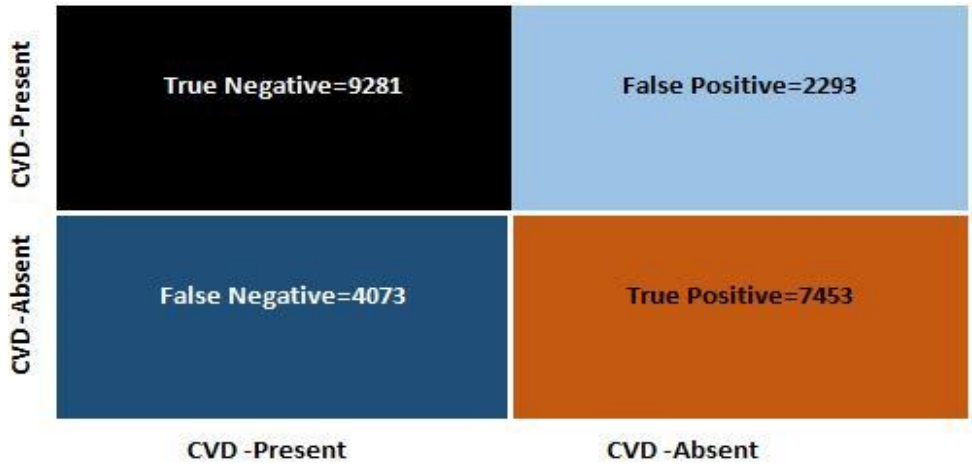

Fig. 4: Confusion matrix - RF based classifie

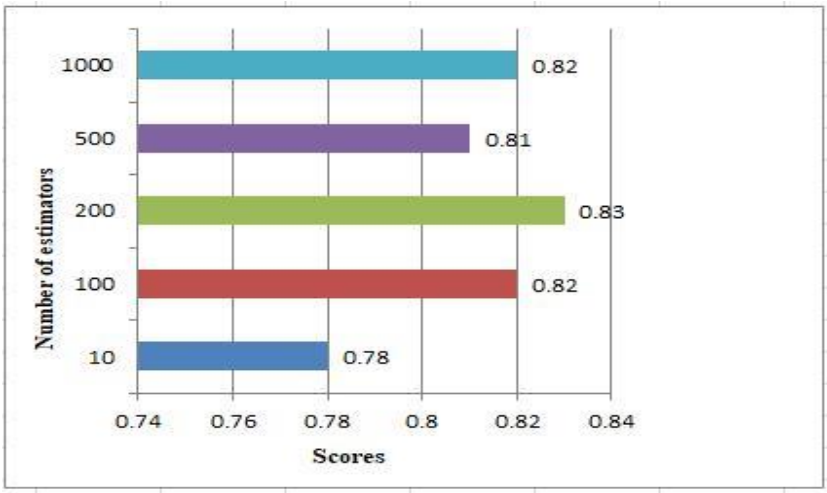

Fig. 5: Depicting random forest scores with different estimators

3 Results

Classifiers	Training Accuracy	Test Accuracy	F1 Score	Precision	Sensitivity
siviv Classifier	82.2%	81.9%	72.02%	73%	0.72
Random Forest Classifier	92.6%	92.4%	73.25%	74.0%	0.73

Table 2: Comparing the training and testing accuracy scores of the applied models

We have made use of the classification methods of machine learning to classify the cardiovascular diseased patients from the non-CVD patients. The classifiers are creating the base for diagnosing whether the person with particular attribute values can have cardiovascular disease by providing information regarding the patients at the early stage. It is observed from Table 2 that the results that the machine learning based predictors produce very accurate results for the identification of the patients having cardiovascular diseases. Our proposed system can predict whether the patient can have CVDs based on the attribute values on the basis of self-learning mechanism.

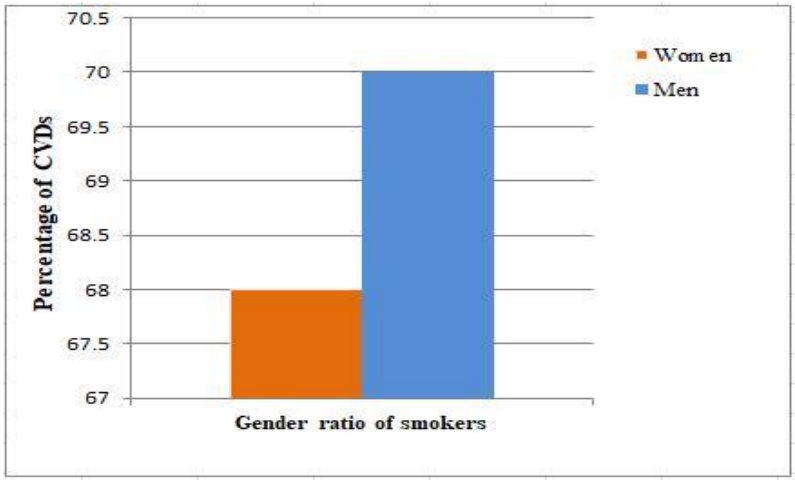

Fig. 6: Percentage of CVD patients *w.r.t.* male and female smokers

The other observations of the paper highlight the potential attributes that contribute in the occurrence of CVDs. We have elaborated the contributions of the attributes for CVDs by performing calculus formulae on CSV sample file.
The first attribute taken for study is 'smoking'. We have observed that out of the dataset of 70,000 patients, 21,000 men are smokers and 1600 females are smokers. 70% of men and 68% women who smoke are suffering from cardio-vascular diseases. The next parameter considered for study is cholesterol level. Out of 70,000 patients, 19000 patients have more cholesterol than the normal range and 4700 patients have very high cholesterol. Total 80% patients with very high cholesterol level have problems of CVDs and 40% patients with more than normal ranged cholesterol have problems of CVDs. The other important risk factor considered for the study is alcohol.

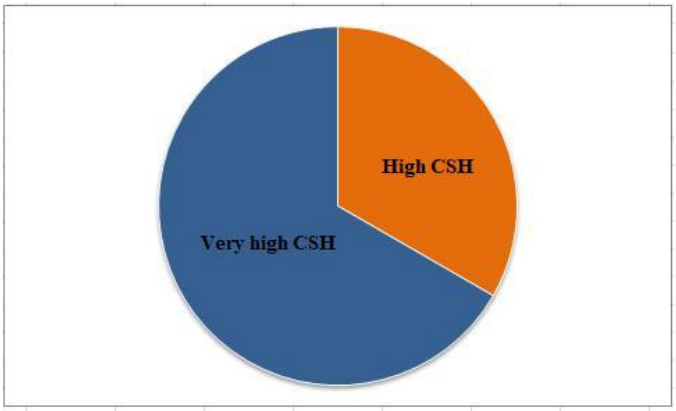

Fig. 7: Percentage of CVD patients w.r.t. cholesterol levels

Out of 70000, there are 42,250 men who consume alcohol and 12,375 women who consume alcohol. It is observed that 37,000 men out of the population of alcoholics are suffering from CVDs whereas 9000 women are suffering from CVDs. High BP is also a potential risk factor for CVDs. 93% of the population from the sample collected is suffering from Cardiovascular diseases who have problem of high blood pressure. Physical activity is very crucial for the heart health. As per our records, 69% patients over the age of 39 who do not perform any physical activity are classified as CVD diseased patients.

The next potential risk attribute for cardiovascular disease is Diabetes. Diabetes affects the heart muscle and can cause diastolic and systolic heart failure. From our sample data, 71% patients with diabetes type-2 and 41% patients with diabetes type-1 are suffering from cardiovascular diseases. A history of CVDs in one or more blood relatives of a person is one of the prime risk attribute for experiencing CVD issues. As per our sample data, 38% of the patients who are currently have complaints of CVDs have family history of heart diseases.

4 Discussions

In this paper we have collected the data of 70,000 patients from North-India region and analyze the impact of risk factors with respect to cardiovascular diseases. We have also attempted to apply classification algorithms for the identification of the patients suffering from cardiovascular diseases by an automated system to assist the physicians for identifying CVD patients at early stage. We have evaluated the performance of the techniques considered for the study on the basis of confusion matric, RoC curve, F1 score, recall and precision score. It is observed that automated machine learning techniques produce accurate results and can assist the doctors to diagnose the diseased patients in a fast manner. We have also studied the results considering each risk attribute and

observed the impact of each attribute on the patient diagnosed with CVD. We have also provided insights into some attributes which play as major risk factors for CVDs. In future, we can study that how group of attributes altogether impact the patient who suffers from cardiovascular diseases.

5 List of Abbreviations

Abbreviation	Description
ML	Machine learning
CVD	Cardiovascular disease
BP	Blood pressure
CSH	Cholesterol
ap hi	systolic blood pressure
ap low	diastolic blood pressure
PAD	Peripheral arterial disease

References

1. K. Buchan, M. Filannino, zlem Uzuner, Jour. of Biomedical Informatics,Elsevier 72, 23 (2017). DOI https://doi.org/10.1016/j.jbi.2017.06.019
2. M. Mitra, R. Samanta, Procedia Technology 10, 76 (2013). Https://doi.org/10.1016/j.protcy.2013.12.339.http://www.sciencedirect.com/science/article/pii/S2212017313004933. First Int.Conf. on Comp. Int.: Mod.Tech. and Appl. (CIMTA) 2013.
3. M.A. jabbar, B. Deekshatulu, P. Chandra, Procedia Technology 10, 85 (2013). DOI https://doi.org/10.1016/j.protcy.2013.12.340. First Int. Conf. on Comp. Int.: Modeling Tech. and App.(CIMTA) 2013
4. A. Gavhane, G. Kokkula, I. Pandya, K. Devadkar, in Second Int. Conf. on Elec., Comm. and Aero. Tech.(ICECA) (2018), pp. 1275-1278
5. S. Manikandan, in Int. Conf. on Energy, Comm., Data Anal. and Soft Comp. (ICECDS) (2017), pp. 817-820
6. S. Mohan, C. Thirumalai, G. Srivastava, IEEE Access 7, 81542 (2019)
7. R.G. Saboji, in Int. Conf. on Energy, Comm., Data Ana. and Soft Comp. (ICECDS) (2017), pp. 1780-1785
8. A.R. et al., in Int. Conf. on Int. Comp. and Cont. (I2C2) (2017), pp. 1{8
9. J. Thomas, R.T. Princy, in Int. Conf. on Circuit, Power and Comp. Tech.(ICCPCT) (2016), pp. 1-5
10. L.A. et al., IEEE Access 7, 34938 (2019)
11. T. Karaylan, . Kl, in Int. Conf. on Comp. Sci. and Eng. (UBMK) (2017), pp. 719{723
12. S.P.R. et al., in 11th Int. Conf. on Hum. Sys. Int.(HSI) (2018), pp. 233{239

13. S.S. et al., Procedia Computer Science 132, 1290 (2018). Int. Conf. on Comp. Int. and Data Sc.

14. C. Sowmiya, P. Sumitra, in IEEE Int. Conf. on Int. Tech. in Con., Opt. and Signal Proc. (INCOS) (2017), pp. 1-5

15. M. Gandhi, S.N. Singh, in Int. Conf. on Fut. Tren. on Comp. Anal. and Knowl. Mana.(ABLAZE) (2015), pp. 520-525

16. M. Sultana, A. Haider, M.S. Uddin, in 3rd Int. Conf. on Elec. Eng. and Info.Comm. Tech. (ICEEICT) (2016), pp. 1-5

17. K.G.D. et al., in Int. Conf. on Curr. Tren. to. Conv. Tech. (ICCTCT) (2018), pp. 1-7

18. Journal of Biomedical Informatics 58, S158 (2015).URL

    http://www.sciencedirect.com/science/article/pii/S153204641500194X. Proc. of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Chall. in NLP for Clinical Data

19. C.B.C. Latha, S.C. Jeeva, Informatics in Medicine Unlocked, Elsevier 16, 100 (2019)

20. D. Jain, V. Singh, Egy. Infor. Journal 19(3), 179 (2018)

21. N.I. Hasan, A. Bhattacharjee, Bio. Signal Processing and Control 52, 128 (2019). URL http://www.sciencedirect.com/science/article/pii/S1746809419301028

22. Proc. Comp. Sc. 120, 588 (2017). 9th Int. Conf. on Theory and App. of Soft Com. ICSCCW 2017.

23. E.E.T. et al., Comp. and Struc. Biotechnology Journal 15, 26 (2017). URL http://www.sciencedirect.com/science/article/pii/S2001037016300460

24. Purushottam, K. S., R. S., Procedia Computer Science 85, 962 (2016). URL http://www.sciencedirect.com/science/article/pii/S187705091630638X. Int. Conf. on Comp. Model. and Sec. (CMS 2016)

25.