International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Customized Prediction of Cardiovascular Disease Using Machine Learning: Analysis of Regional Risk Factors

Dr. Uthama Kumar*[1]. A. Yamini Sahukar P.[2]

Submitted: 29/12/2023 **Revised**: 05/02/2024 **Accepted**: 13/02/2024

Abstract: Cardio Vascular Disease (CVD) is the most silent killer the world has ever witnessed. The silence of the disease is the salient feature and the most deadly of all the characteristics of the disease. It is a combination of various collections of ailments of the blood vessels and heart. CVD cannot be considered as a single disease and it adds more complexity to diagnosis and prediction at an early stage. The paper uses standard data available in the public domain. Analysis and review of literature indicates not set of parameters are enough for prediction of CVD and these parameters vary from region to region. The paper provides a novel method which uses a combination of machine learning algorithms to calculate the probability of risk of individual using known parameters. The paper provides an algorithm for establishing set of best fit parameters, set of best fit Machine learning methods and also provides analysis of prediction using inbuilt tool box of Python vs the novel CVD index method. The paper considers customization of data and prediction methods for each region which provides better results.

Keywords: *CVD, Prediction of CVD, Machine learning algorithms, CVD Index.*

1. Introduction

It is estimated that nearly 18 million people met premature death (under age of 70) globally due to this epidemic and more worrying fact being 85% of them died due to heart attack also called as Coronary heart ailment and/or Stroke also called as Cerebrovascular ailment by Global Burden disease study). It is seen that the age standard CVD deaths in India is 272 for every 1,00,000 populations which is much higher when compared to 235 deaths for every 1,00,000 populations globally [11]. This signifies the amount of importance that needs to be attached to CVD in our country. The paper is organized into review of existing work, need for a customized index based on region, sex & minimum variables, deduction of CVD index and conclusion.

2. Review of Existing Work

There have been a number of studies on CVD in India. Considering the last two decades' milestones of Coronary heart disease analysis, there have been a few of them which serve as benchmarks to our study. The over a population of 1,235 in rural parts of India indicated that 7% of the rural mass had Coronary heart disease. The major risk factors indicated that 18.8% of them had hypertension, 4% had Diabetes and 21.9% were smokers [1]. A study year 2011 over a population of 14,886 in rural and urban parts of India indicated that 9.7% of the urban and 2.7% of the rural mass had Coronary heart disease. The major risk factors indicated that 10.6% of them had hypertension, 1.5% had Diabetes, 43.7% of them had high cholesterol and 18.1% were smokers. The study of Singh et al., in the year 1997 over a population of 3,575 in rural and urban parts of India indicated that 9.0% of the urban and 3.3% of the rural mass had Coronary heart disease. The major risk factors indicated that 23.4% of them had hypertension, 4.5% had Diabetes, 22.0% of them had high cholesterol and 19.7% were smokers. The study of Gupta et al., in the year 2002 over a population of 1,123 in urban parts of India indicated that 8.2% of the urban mass had Coronary heart disease. The major risk factors indicated that 36.9% of them had hypertension, 12.2% had Diabetes, 39.1% of them had high cholesterol and 23.9% were smokers [7].

The study over a population of 7,449 in rural and urban parts of India indicated the major risk factors indicated that 28.8% of them had hypertension, 14.8% had Diabetes, 54.1% of them had high cholesterol and 42% were smokers [2].

The study in the year 2011 by Cardiological society of India, conducted by its Kerala chapter on the topic "Coronary Artery Disease and its risk prevalence" over a population of 5,193 in rural and urban parts of India indicated that 15.7% of the population had Coronary heart disease. The major risk factors indicated that 39% of them had hypertension, 21% had Diabetes, 23% of them had high cholesterol and 31% were smokers [10].

As the analysis indicates there has been a constant rise in all the risk factors. Considering these studies, it is clear that not one risk factor could be considered as high risk factor. The

[1] *Asst. Professor, Dept. of Computer Science, St. Francis College, Bangalore, Karnataka, India.*
ORCID ID : 0009-0007-9459-6587
[2] *Assistant Professor, Dept of AI&ML, Bangalore Institute of Technology, Bangalore, Karnataka, India*
ORCID ID : 0009-0005-8535-4880
** Corresponding Author Email: uthamkumar.kumar897@gmail.com*

factors have changed from region to region and from time to time based on the life style habits of the population. The hypertension from various studies indicate that there is a variance from 10% to 39% in the past two decades, similarly there is a variance from 4% to 21% of Diabetes in the past two decades, variance from 22% to 54.1% of high cholesterol in the past two decades and a variation in smokers from 18.1% to 42% in the past two decades [6].

The interesting factor is that each of these studies indicates a different risk factor as the main culprit of CVD. Raman Kutty et al., indicates that main life style problem is “Smoking” leading to CVD. Chadha et.al. indicates that high clinical measure of bad cholesterol is a major cause of CVD and morbidity. Singh et at., indicates that high blood pressure (Hypertension) pays a spoil sport in a patient with CVD. Gupta et al., indicates that almost hyper tension and high cholesterol both form a combination of major cause of Coronary heart disease. Thankappa et.al., indicates that high cholesterol and smoking combination is a major combination cause of Coronary heart disease. Cardiological Society of India, Kerala chapter Coronary Artery Disease and its Risk indicates that a combination of all these is a high risk of Coronary heart disease. Each of these studies not only indicate the changing major risk factors over a period of time, but also indicate there could some more unexpected parameters that could add in or get deleted due to demographic and life style changes.

In the year 2016, estimation of risk of CVD specific to a population was attempted. The risk chart was considered to be primary prevention strategy of CVD. SCORE risk chart was first used for Germany to calculate the fatal CVD risk rate. The calculation was used for mortality data of 1998 and 1999. It is seen that there needs a calibration and recalibration of risk and risk factors is needed [8].

The German Health Interview and Examination Survey for Adults 2008-11 and official mortality rate from 2012 was used to provide a risk chart based on SCORE methodology. The competitive risk charts considered major risk factors od blood pressure (systolic), sex of the patient and bad cholesterol with an age group of five years. The recalibrated chart was mo re efficient than the older ones as it provided more accurate Score. The study of 3062 people between the age of 40 to 65 showed the risk 29% less. The authors conclude by indicating the need to constant recalibration of risk charts based on mortality and risk factors levels [11].

A study in 2016 provides the overview of number of prediction models for CVD so far. There were 363 predictive CVD models and 473 validations. A good number of these models trace to Europe (46%). Models that worked on last ten years data for both fatal and non-fatal CVD was 33%. Models working on more than ten years data were 58%. Some of the common risk factors listed were smoking, age and sex. There is a large difference in prediction, definitions, major factors, combination of these factors with a considerable amount of void in clinical and methods. Data is often missing in these models. It is seen that in 13% of the models prediction horizon is missing. In 25% of the models data needed for application of the model for calculation of individual risk data and other database is missing. 36% of the models were externally validated and only 19% were validated by independent investigators. Heterogeneity was the order of performance. The models had varied discrimination and calibrations were needed. The conclusion thought propagates that there are a number of excessive models existing, but are either not implementable due to lack of metadata or because they are not properly validated by third party for acceptance. Now that there is a mismatch in datasets and the need of the model is also a major reason for not having available information on calibration and proper use of models. It is proposed that these models should be properly put together and customized to suite individual group needs [2].

3. Machine Learning

Machine learning has been a new trend in the areas where there exists more complexity of data and complex grouping. This forms a perfect basis for the use of machine learning in CVD prediction. There has been a more interest growing in the researchers and the clinical professionals due to the accuracy and time sensitivity of the problem. Machine learning based algorithms are used to classify a healthy person from a possible CVD candidate. A number of methods in various papers improvements of decision tree, Bayesian classification, method derived from Chou’s pseudo amino acid composition, methods using support vector machine (SVM), using probability estimation rules, alternate classification rules of data mining, rotation forest (RF), artificial neural network are a path breakers in the development of CVD prediction model using the changing technology [3].

A risk prediction model is developed for all cause deaths and CVDs. The model developed had ten risk factors traditionally used for clinical practices. All these risk factors had good impact on the mortality rate. As CVDs are a major cause of concern in India, the model was targeted towards identification of high risk of mortality. The model claims identification of risk prediction in long term all cause and chronic ischemic CVD.

4. Calculation of CVD Index

The work infers from the research review of above that CVD criterion change from region to region based on socioeconomic and environmental conditions, Behavior and infections and Physiological pathways. The minimum risk factors identification for a particular region is also a major challenge, which is out of scope of this paper. The paper provides a frame work for fitting the existing data into the method and then refining it further based on the accuracy, false negatives and false positives. The method used 7 machine learning well known methods. The methods are Linear regression, Simple Vector Machine, Naive Bayes, Decision tree, Random forest, Ada Boost and Bagged tree. As the criterion varies the accuracy and false negatives and positives also vary. Hence we do not settle for one fixed method or set of methods. The method allows for random selection of methods for a specific region and evaluate from time to time for accuracy.

4.1. The algorithm for calculation of CVD Index is as follows:

Step 1: Consider Geographical Region based data training data

Step 2: Weight of each prediction method is set to 1.

Step 3: For each sample predict the risk factor using the methods Linear regression, Simple Vector Machine, Naive Bayes, Decision tree, Random forest, Ada Boost and Bagged tree.

Step 4: For each training set, if the prediction of a method is greater than 30% and less than 75%, then,

Step 6: For each of the methods in range of Step 5 increment the weight by 1 and for each of the methods not in the range of Step 5 decrement the weight by 1.

Step 7: On completion of training set, we select best 6 methods for a particular region for computation of CVD index.

Step 8: Calculate the CVD index based on the formula of weighted average using the accepted 6 methods chosen in Step 7

$$CVD\ Index = \frac{\sum(Prediction\ \%\ )(corrosponding\ weight)/\ \sum(corresponding\ Credit\ )}{\sum(Credits)}$$

Step 9: If CVD index is greater than 6 then it considered from inspection method of data the patient needs clinical assistance. If CVD index is greater than 5 but less than 6, life style, habits and more physical activity could reduce the risk and if it less than 5 then he may not be into CVD risk as of now.

5. Comparison of Results

Data available with Kaggle [5] and Cleveland Heart Disease (UCI Repository) [9] dataset is used for the following analysis.

The total number of patients data considered from Kaggle is 16,383. The calculation uses 12,287 patients' data (75% of total data) for training purpose and 4,096 patient's data is used for testing the prediction.

The total number of patients data considered from Cleveland Heart Disease (UCI Repository) dataset is 10,000. The calculation uses 7,500 patients' data (75% of total data) for training purpose and 2,500 patient's data is used for testing the prediction.

The following table 1 provides the average accuracy of each of the algorithms/methods along with false negative and positives for both the data sets. False negatives are number of patients who were physically dragonized by qualified physician for having CVD but diagnosed by machine learning algorithm as not having CVD. Similarly False positives are number of patients who were physically dragonized by qualified physician for not having CVD but diagnosed by machine learning algorithm as having CVD. The following is a result of using Python inbuilt library "Scikit-learn). This is a clear indication that no one method can be used for prediction as the criterion keep changing from region to region and time to time.

Table 1 Accuracy of each of the algorithms/methods when Scikit-learn tool box of Python is used

Algorithm	**Accuracy**	**False Negative** **(Out of 4096)**	**False Positive** **(Out of 4096)**
Logistic regression (LR)	0.73	708	404
SVM	0.73	798	314
Naive Bayes (NB)	0.57	1650	115
Decision tree (DT)	0.64	766	700
Random Forest (RF)	0.7	754	472
Ada Boost (AB)	0.74	720	345
Bagged tree (BT)	0.72	649	505

The following table 2 and table 3 provides the Accuracy using the proposed algorithm for men with CVD and women with CVD respectively. It can be seen that the accuracy of the given algorithm is much higher than the accuracy that is provided by conventional method used in table 1. The algorithm also takes into account the gender of the patient as well and as it is region wise data, the accuracy can be monitored and fine-tuned if the parameters based on which CVD is dependent changes in time, which is a expected feature of nature. Thus the method provided here is more flexible and more accurate when it comes to early prediction and also is easily maintainable and with supervised learning picking up inn its concepts and accuracy can easily be incorporated in this method so as to provide better results in near future as well.

Table 2 Accuracy using the proposed algorithm for men with CVD

LR	SVM	NB	DT	RF	BT
Kaggle dataset					
91.82	92.5	60.25	88.25	96.39	96.35
98.3	98.4	61.2	96.42	97.84	98.05
Cleveland Heart Disease (UCI Repository) dataset					
93.21	93.83	30.25	76.79	92.22	95.88
96.38	96.29	36.61	92.01	31.25	96.39

Table 3 Accuracy using the proposed algorithm for women with CVD

LR	SVM	NB	DT	RF	BT
Kaggle dataset					
92.64	93.04	60.1	85.52	95.83	96.55
97.8	97.95	63	94.78	97.87	98.06
Cleveland Heart Disease (UCI Repository) dataset					
93.12	94.74	71.31	81.06	90.93	89.16
98.1	98.69	76.73	92.31	91.5	96.6

6. Conclusion

The paper is a novel prediction procedure of CVD at an early stage so as to keep the patient out of risk, danger of hospitalization and other emergencies. The paper stresses on the need for regional data so that accuracy can be improved. The major risk factors also vary from region to region and hence there is a need to have a constant check on the variables need for prediction, these factors again vary from region to region and in a country like India where diversity is seen in abundance, it is all that more important to be cautious on both factors and prediction methods that are efficient for a particular region. The paper provides a comparison between the way data is used in the real world and novel method proposed to indicate better performance of CVD index method. The CVD index needs to be checked at various regions to analyze accuracy. The available data from two sources on prediction yields good results and encourages us to construct better digital datasets that can be used for prediction.

The use of above CVD index method provides the following results:

The accuracy of prediction improves enormously by the use of CVD index. On successful implementation, the method could serve as a major indicator in assessment CVD of risk.

References

[1] Anoop Misra, et., Consensus Dietary Guidelines for Healthy Living and Prevention of Obesity, the Metabolic Syndrome, Diabetes, and Related Disorders in Asian Indians Diabetes Technology & TherapeuticsVol. 13, No. 6 (2011), https://doi.org/10.1089/dia.2010.0198

[2] Brian C. S. Loh et. al., Deep learning for cardiac computer-aided diagnosis: benefits, issues & solutions, mHealth 2017

[3] Hajdu, A., Terdik, G., Tiba, A. et al. A stochastic approach to handle resource constraints as knapsack problems in ensemble pruning. Mach Learn (2021). https://doi.org/10.1007/s10994-021-06109-0

[4] https://archive.ics.uci.edu/ml/datasets/heart+disease

[5] https://www.kaggle.com/ronitf/heart-disease-uci

[6] M.N. Krishnan, Coronary heart disease and risk factors in India - On the brink of an epidemic?, Editorial, Indian heart journal 64 (2012), www.elsevier.com.

[7] Raman Kutty et al,., Availability and affordability of blood pressure-lowering medicines and the effect on blood pressure control in high-income, middle-income, and low-income countries: an analysis of the PURE study data, The Lancet Public Health, Volume 2, Issue 9, 2017, Pages e411-e419, ISSN 2468-2667, https://doi.org/10.1016/S2468-2667(17)30141-X. (https://www.sciencedirect.com/science/article/pii/S246826671730141X)

[8] Rücker V, Keil U, Fitzgerald AP, Malzahn U, Prugger C, Ertl G, et al. (2016) Predicting 10-Year Risk of Fatal Cardiovascular Disease in Germany: An Update Based on the SCORE-Deutschland Risk Charts. PLoS ONE 11(9): e0162188. https://doi.org/10.1371/journal.pone.0162188

[9] Sreeniwas Kumar, Nakul Sinha Med J Armed Forces India. 2020 Jan; 76(1): 1–3. Published online 2020

[10] Thankappan KR. Combating corona virus disease 2019 and comorbidities: The Kerala experience for the first 100 days. Int J Non-Commun Dis, [serial online] 2020 [cited 2021 Nov 19];5:36-42 [2], https://csi.org.in/

[11] Viktoria Rücker, et al., Predicting 10-Year Risk of Fatal Cardiovascular Disease in Germany: An Update Based on the SCORE-Deutschland Risk Charts. PLOS ONE | DOI:10.1371/journal.pone.0162188 September 9, 2016