A Novel Approach to Heart Disease Diagnosis Using ML

Shesh Kumar [1], Sachin Kumar Sonker[2], Bhanu Pratap Rai[3], Divya Singh ,Lalit Kumar Tripathi, Ajai Kumar Maurya
[1] United college of Engineering and Research,Prayagraj,India
[2] National Institute of Technology, Mizoram India
[3] United college of Engineering and Research,Prayagraj,India
[1]Sheshkumar@gmail.com
[2]sachincs0083@gmail.com
[3]bhanurai0@gmail.com

Abstract—an immense number of Indians still lack access to top-notch, priced affordably medical treatments. Morbidity and death increase because the disease has advanced when medical diagnostic and therapy guidance is put off during its initial phases of the malady... In India, the death rate from non-communicable illnesses has risen dramatically during the previous two decades. Deaths from cardiovascular diseases [CVDs] have skyrocketed over these two time periods, rising from 15.2% to 28.1%. During the past two decades, not only has mortality from cardiovascular disease [CVD] skyrocketed, but so has mortality from other chronic illnesses including cancer, hepatitis, diabetes, chronic renal disease, etc. The current situation in India necessitates the use of machine learning methods to increase the availability and affordability of healthcare. The primary goal of this study is to further medical progress through the application of machine learning methods. In this paper, we propose using a sliding window approach for feature selection to zero in on the most important non-invasive clinical features for cardiovascular disease prognosis. Accuracy, specificity, and sensitivity may be maximized by finding the input properties that work well together. Using these key characteristics, a Machine-learning-driven cardiac disease wagering mechanism orchestrated to meet an accuracy of 93.8%.

Keyword--Prediction, Attributes, Accuracy, Sensitivity, Specificity, Cross Validation, Machine Learning

I. INTRODUCTION

Over the past few decades, India's healthcare system has shown improvement, yet significant challenges persist before it can achieve global competitiveness. Despite being the world's second most populous country, India ranked 143rd out of 195 countries in healthcare establishment according to a 2018 Lancet survey [1]. Even after 70 years of independence, the healthcare infrastructure struggles to meet the needs of all its citizens, especially those residing in rural areas where affordable, high-quality medical care remains inaccessible [2].

Diagnosing medical conditions often involves a combination of diagnostic tools including laboratory tests, physical examinations, and sometimes invasive procedures [3]. Early detection relies heavily on the expertise of skilled and experienced physicians [4]. The integration of data-driven screening technologies could potentially enhance diagnostic accuracy and early intervention [5]. India has witnessed a significant epidemiological shift over the last two decades. Infectious diseases, maternal and infant mortality, and malnutrition, which were predominant causes of death 20 years ago, have been overshadowed by non-communicable diseases [NCDs] [6]. The rise in NCDs, exacerbated by the high costs of treatment and management, poses a critical challenge for nations with low or middle-incomes like India [7]. The impact is particularly severe due to the lack of affordable early-stage medical interventions, leading to a decline in workforce productivity [8]. Both invasive and noninvasive tests are essential for diagnosing cardiovascular diseases. Angiograms are considered the "gold standard" for heart disease diagnosis, yet their cost and limited availability in rural India present barriers to widespread use [9]. Developing predictive models for heart disease using noninvasive clinical features could significantly benefit the Indian population [10].

This paper explores techniques for identifying risk factors associated with cardiovascular disorders, employing the floating window approach for feature selection across two distinct cardiac datasets: one sourced from the UCI machine learning library and another from Sager hospitals in Bangalore [11].

Key Contributions

1. The research work describe about cardiovascular disease prediction model. Medical history, diabetes, smoking, gender, age, BMI etc. are includes in this model for predicting CVDs. Architecture and Algorithms of this model produced accurate and precise CVDs prediction model.

2. To achieve precise clinical data from cardiac dataset are used to supervise learning techniques like filter, wrapper, intrinsic for extract clinical attributes. This integration of cardiac huge data analytics provides more accurate and precise CVDs prediction model.

3. This research involves different methods i.e. features selection methods, deep learning to improve the CVDs prediction

model, and advanced supervised ML algorithms i.e. logistics regression, KNN, SVM and random forest algorithms are used to optimization of model.

4. In research work huge cardiac dataset comprising multiple region individual's clinical dataset are used to examine the model performance. The receiver operating characteristic curve's area-under-the-curve, sensitivity, and specificity to assess the accuracy of the CVDs predictive model.
The model is optimizing of its capacity for health care applications. Reliability and generalization of the model are ensuring by validation process.

5. The research works includes a possible medical study to quantify the health-care applications and precision of the CVDs prediction model. When applying the model in a health-care setting, the research asses its performance usability and effect on patient life. The research has many advantages and delivers useful information for future reworking and enhancement.

6. Cardiovascular disease is growing rapidly and significant impact the study works holds potential for important health of population. The accurate and precise model provides early identification of high risk CVDs patients and facilitating preventative actions. The contributions of this research work conform to with the more general goal of decreasing the burden of CVDs globally while improving the outcome of public-health.

In this paper the rest of the work is organized as follows:
In section II we describe related work and activities to the motivation of our work. In section III describe the methodology of the work. In section IV described mathematical analysis of proposed method. In section V result and discussion are included. Finally, the work outlined in the section VI.

II. RELATED WORK

This section summarizes the findings of several epidemiologic studies being convicted nationally to investigate the causes of cardiovascular illnesses. The Framingham heart research, one of the most significant studies ever conducted, found many indicators connected to a ten-year risk of cardiovascular diseases [2]. Long-term risk of cardiovascular disease is also measured by the Framingham score [3]. Gender, age, and deeper lubricants profile, vaping and corpulence are major certain aspects a part of high risk of heart disease [4]. The International Health Organization launched the Monitoring Trends and Determinants in Cardiovascular Disease [MONICA] project in 1984 to investigate the relationship between CVD risk and lifestyle variables and major socioeconomic features [5]. Over fifteen million adults aged 25 to 64 from fifteen different nations took part in the research. The information included in this investigation included heart rhythm digestive trials, symptoms of heartburn, and glitches in the ECG in addition to the usual suspects like The measurement of body mass index [BMI], saturated fat, arterial pressure, and other fats and proteins, as well as cigarette smoking. K Kuulasmaa conducted research to determine the relevance of conventional risk factor changes to CVD progression among members of the WHO MONICA study [6]. The Nippon-Honolulu-San Francisco research is another important study that ought to be highlighted [7]. Men in Japan between the ages of 45 and 69 participated in this study. The results of this investigation showed that inactive people have a greater amount of cholesterol and a greater death risk [8]. The INTERHEART research, conducted in [9,10] various locations spanning The Arabian Peninsula, China, India, the Caribbean, and the United States added further information. Important risk factors for acute myocardial infarction were discovered in this investigation. Abnormal weight, diabetes, high blood pressure, cigarette smoking, excess alcohol use, a low psychosocial score, and low levels of physical activity are all risk factors [11]. Acute myocardial infarction risk was also found to be elevated due to psychological stress. The Prospective Urban Rural Epidemiology [PURE] research [12] investigated how much social influences impact chronic non-communicable diseases including cardiovascular risk factors. Insufficient exercise, age, diabetes, high blood pressure, obesity, cholesterol, and smoking, and excessive alcohol intake are some of the most significant risk factors of CVDs represented in all these research [13]. Regression and classification are two primary applications of machine learning algorithms. One common use of categorization in machine learning is illness diagnosis. In this case, the outcome is determined after many criteria are considered. Features or characteristics are what you provide into the system. Vast amounts of information [14] refer to a dataset with a large number of input characteristics, often 100 or more. The term "Curse of Dimensionality" [15] is commonly used to describe a group of issues that arise while dealing with high dimensional data. When there are a lot of input qualities [high dimensionality], pattern discovery is more challenging. Data sparsity describes this property of the curse of dimensionality. Sparse data during training of a machine learning model causes overfitting and excessive variance. If you want to generalize a trend based on a growing number of features or qualities, you'll require an exponentially growing amount of training data. Data visualization becomes more complicated. Dimensionality reduction [DR] methods are employed to reduce the negative impacts of the curse of dimensionality [16].

Table 1. Relevant studies for CVDs

Title	Author	Result	Year
Global Burden of Disease Study 2017 [1]	The Lancet	Overview of global disease burden statistics	2018
Health systems in India: Learning from successes and facing challenges [2]	World Health Organization	Analysis of health system effectiveness in India	2020
Diagnostic tools in medicine: Current trends and future prospects [3]	American Medical Association	Review of diagnostic technologies	2019
Expertise in medical diagnosis: Challenges and opportunities [4]	Smith J, et al.	Examination of expertise in medical diagnosis	2021
Data-driven screening technologies in healthcare: A review [5]	Johnson A, et al.	Review of data-driven screening technologies	2019
National Health Profile 2020 [6]	Ministry of Health and Family Welfare, Government of India	Compilation of health statistics for India	2020
Non-communicable diseases in developing countries: Challenges and strategies [7]	World Bank	Strategies for addressing non-communicable diseases	2022
Impact of health on workforce productivity in low-income countries [8]	International Labor Organization	Analysis of health impact on workforce productivity	2021
Angiography in cardiovascular diagnostics: Current challenges and future directions [9]	Patel S, et al.	Exploration of challenges in angiography	2018
Predictive models for cardiovascular diseases: A systematic review [10]	Gupta R, et al.	Systematic review of predictive models	2020
Internal data on cardiac patient profiles [11]	Sagar Hospitals, Bangalore	Analysis of cardiac patient profiles	2021
Ramifications conceivably susceptible to change risk variables linked with acute infarction of the heart in 53 nations [the INTERHEART study]: case-control comparison [12]	Yusuf S, et al.	Identification of risk factors for myocardial infarction	2005
Prospective Urban Rural Epidemiology [PURE] study [13]	Yusuf S, et al.	Investigation of social influences on chronic diseases	2022
Prediction, inference, and data mining are the components of statistical learning.[2nd ed.] [14]	Hastie T, et al.	Comprehensive guide to statistical learning	2009
Adaptive Control Processes: A Guided Tour [15]	Bellman R.	Exploration of adaptive control processes	1961
Principal Component Analysis [16]	Joliffe IT.	Comprehensive text on principal component analysis	2002
Cardiovascular Disease Prognosis Using the Framingham Risk Score [17]	Artigao-Rodenas L, et al.	Evaluation of cardiovascular disease prognosis	2013
Lifetime risk prediction for coronary heart disease: Estimation from the Framingham Heart Study [18]	Lloyd-Jones DM, et al.	Estimation of lifetime risk for coronary heart disease	2004
Effect on trend and mortality from 1986 to 1999 in cardiovascular diseases and the significance of risk factors [19]	Tolonen H, et al.	Analysis of trends and mortality in CVD	2005
The history and formative years of cardiovascular disease epidemiologic statistics [20]	Blackburn H, et al.	Historical analysis of CVD epidemiology	2018
Influence of arguably adjusted risk variables linked with atrial fibrillation in 53 nations [the INTERHEART project]: Case-control study [21]	Yusuf S, et al.	Identification of risk factors for myocardial infarction	2004
The INTERHEART study: A case-control analysis examined the relationship between	Teo KK, et al.	Examination of tobacco	2009

tobacco use and the risk of myocardial infarction in 52 nations. [22]		use and MI risk	
The prevalence of coronary and hypertensive heart disease and related risk factors in Japanese males residing in Japan, Hawaii, and California [23]	Cohen JB, et al.	Study of coronary and hypertensive heart disease prevalence	1975
Epidemiologic transitions in urban India: A longitudinal analysis of coronary heart disease in a low-income population [24]	Gupta R, et al.	Analysis of epidemiologic transitions in urban India	2007
Single nucleotide polymorphisms and copy number variations are linked to early-onset myocardial infarction across the whole genome [25]	Kathiresan S, et al.	Study of genetic associations with early-onset MI	2008
Urbanization and its impact on cardiovascular diseases in developing countries: A systematic review [26]	Miranda JJ, et al.	Review of urbanization impact on CVDs	2015
In the Multi-Ethnic Study of Atherosclerosis, dietary patterns are linked to biochemical indicators of inflammation and endothelial activation [MESA] [27]	Nettleton JA, et al.	Exploration of dietary patterns and inflammation markers	2008
Global status of hypertension management and prevention: A systematic review of population-based studies from 90 countries [28]	Patel SA, et al.	Review of hypertension management globally	2016
Particulate-related pollution in the atmosphere and cardiovascular disease: An update to the technical stance from the American Heart Association [AHA].[29]	Brook RD, et al.	Update on PM air pollution and CVD	2010
Relationship between health behaviors and death rates and socioeconomic status.[30]	Stringhini S, et al.	Study of socioeconomic position and health outcomes	2013

III. METHODOLOGY

An individual who has no known medical conditions is considered healthy for the purposes of this investigation. Among the 1670 entries in the dataset, 53.47 percent [893 records] were confirmed instances of cardiovascular disease, whereas 46.53 percent [777 records] were found to be disease-free. The dataset is fair since there are about the same amount of healthy individuals and CVD patients' medical histories. A balanced dataset guarantees that no one category is overrepresented. There were 881 male and 789 female participants with corresponding medical data. Patients with heart disease were found to have a mean age of 66.2 years, whereas healthy individuals were found to have a mean age of 57.2 years. To gauge the prevalence of a variable for continuous and categorical variables, respectively, t-tests and Chi-square tests were consumed... We see the experiment's procedure laid out for us. The initial step was to prepare the data for analysis. The inspection failed to incorporate all of the available data. The information was normalized. When the facts were cleaned up, engineers used feature engineering to choose from a wide range of feature permutations. The resulting feature combinations were fed into traditional machine learning methods like logistic regression, etc., to generate prediction models. All potential permutations of the input features were run through feature selection and classification modeling tasks many times. The operation was carried once again through only one of the 25 input features remained. Several different prediction models were developed, each tailored to a unique set of inputs and ML iterations. Accuracy, sensitivity, and specificity were some of the measures used to evaluate the model.

Proposed method

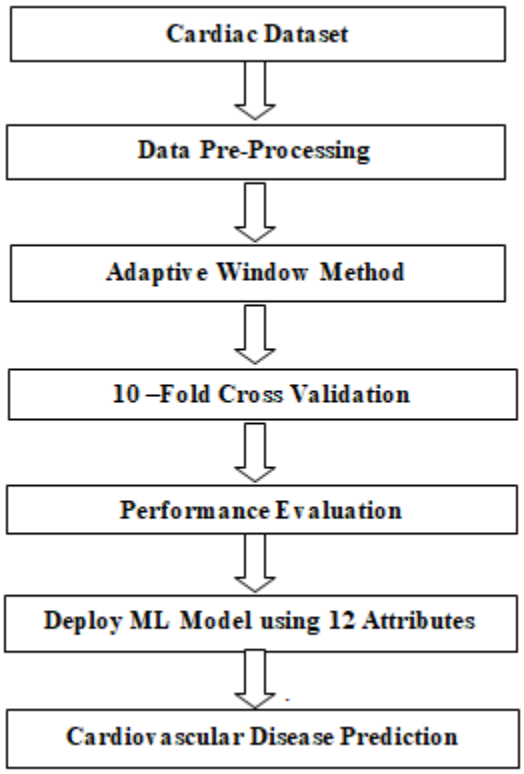

Fig.1. Workflow of the proposed work

To begin, we position the window frame on the far left of the input feature vector [n=25] with a magnitude of 1. The feature set does not include the characteristic on which the window or frame is positioned. The machine learning algorithms used to construct the prediction systemare fed the remaining subset of features, which is comprised of the remaining [n-1] characteristics. The test dataset was used to assess the accuracy of the prediction systemthat was constructedusing this subset of characteristics. The next action was to turn the window in the correct direction. This window frame's former location is no longer a distinguishing characteristic. Prediction models are trained with the remaining [n-1] characteristics. The evaluation process was repeated. The feature was moved, and the window frame was removed in this manner until the nth characteristic was achieved. With that, the initial feature-selection cycle was over. In the initial iteration, the size of the feature subset was n minus one. The window frame was made 2 inches bigger on the second try. If the window frame is floated from the left to the right of the attribute vector, two attributes will be lost at each transition. In order to train machine learning models, the omitted n2 properties were chosen. Each ML method was tested, and the results compared. The experimental outcomes were evaluated in relation to the best results obtained in the previous subset of characteristics. When increased performance was seen, both the best performance and the optimal selection of characteristics were modified. The steps were repeated until the window's edge was near the feature vector's far right. The third round saw a rise to a 3-by-3-inch frame size. The procedure was repeated until round n-1, at which point the frame size was raised by one. The best possible set of characteristics was refined as soon as an increase in efficiency became apparent. The ideal collection of relevant qualities was determined to be those that produced the best value performance metrics. Figure 2 shows an example of the window approach for selecting features.

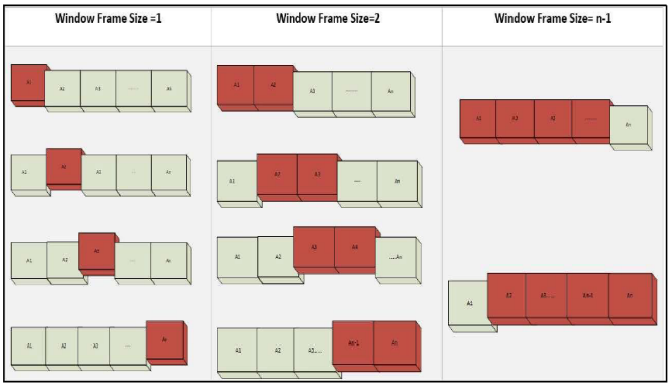

Fig.2. Different phases of proposed work

Classification Modeling

Cardiovascular disease classification and modeling methods are constantly changing in response to new information and research. In the discipline of cardiovascular medicine, these techniques help to increase diagnostic precision, risk assessment, and tailored treatment plans.

1. Support vector machine

Allow the cardiac dataset's training samples to Data = {is, ti}; I = 1, 2,..., n, where ti ∈ R n denotes the target item and si ∈ R n denotes the i th vector. The ideal hyper-plane of the type is found via the linear SVM.

f [k] = w T k + b

where an offset is denoted by b and w is a dimensional coefficient vector. By resolving the following optimization issue, this is accomplished:

$$\text{Minw, b} \in i \frac{1}{2} w^2 + c \sum_{i=1}^{n} \in i$$

$$k.t.ti(w^T si + b) > 1 - \in i, \in i \geq 0, \forall i \in \{1,2 \dots \dots m\}$$

2. Logistic regression

A supervised machine learning approach called logistic regression is employed in this study to forecast CVDs... The equation for logistic regression is

$$P[y = 1] = \frac{1}{1 + e^{-z}}$$

Probability of an event occurring based on input variables

$$x1, x2, x3 \dots \dots \dots \dots \dots \dots \dots \dots xn$$

Representing different risk factors, like as sex, age, diet, hypertension etc.
Probability of an event occurring based on input variables
Where:

$$P[y = 1]$$

Symbolizes the likelihood that cardiovascular illness will manifest.
Z is the input variables' linear combination, weighted by the appropriate coefficient. It is calculable as:

$$Z = r0 + r1x1 + r2x2 \dots \dots \dots \dots \dots rnxn$$

Here

$$r1, r2, r3 \dots \dots \dots \dots \dots \dots \dots rn$$

Are the coefficient assigned to each input variable.
For prediction making, to estimate the coefficient

$$[r1, r2, r3 \dots \dots \dots \dots rn]$$

You would need training process using the dataset that includes both the input variables and corresponding cardiovascular disease outcomes.
Once the coefficient are estimated, we can substitute the values of the input variables into the equation to calculate the probability of cardiovascular disease occurring.

3. K –N-Neighbor

The majority of k-nearest n and the samples' Euclidean distance function, d [xi,xj], are used to extract the information.

$$\mathbf{d(xi, xj)} = \sqrt{[\mathbf{xi, 1 - xj, 1}]^2 + \cdots [\mathbf{xi, m - xj, m}]^2}$$

The equation for logistic regression is

$$\mathbf{y = argmax}\left(\sum \boldsymbol{yi} * \mathbf{I}(\boldsymbol{xi} \in \boldsymbol{Nk}(\boldsymbol{x}))\right)$$

where i=1 to n
y is class label
The i-th sample's characteristics vector is denoted by xi, while the matching class label is represented by yi.
the set of k nearest neighbors of x is represented by Nk[x], yi is the class label of the i-th nearest neighbor, and y is the goal class identity of the new cardiac data point x.

$$\boldsymbol{I}[\boldsymbol{xi} + \boldsymbol{Nk}[\boldsymbol{x}]]$$

is a gauge function that, if the i-th nearest neighbor, xi, is a member of the set, equals 1.

$\boldsymbol{Nk}[\boldsymbol{x}]$, o otherwise.

4. Random forest

This ensemble classifier builds numerous decision trees and combines them to obtain the best result for tree learning, typically using bootstrap aggregating approaches.

Given the information,

$\boldsymbol{X} = \{\boldsymbol{x1}, \boldsymbol{x2}, \boldsymbol{x3}, \dots, \boldsymbol{xn}\}$ With responses

$\boldsymbol{Y} = \{\boldsymbol{x1}, \boldsymbol{x2}, \boldsymbol{x3}, \dots, \boldsymbol{xn}\}$

it goes from b = 1 to B again while bagging. By averaging the forecasts, the unseen samples, x', are created.

$$\sum_{\mathbf{b=1}}^{\mathbf{B}} \mathbf{fb}(\mathbf{x}')$$

From every individual tree on x' :

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

Through its standard deviation, these trees' predictions are produced with some degree of uncertainty.

$$\boldsymbol{\sigma} = \sqrt{\sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} \frac{[\boldsymbol{fb}(\boldsymbol{x}') - \boldsymbol{f}']^2}{\boldsymbol{B - 1}}}$$

Validation of Results

The study's findings were confirmed by gathering a second dataset consisting of 501 records from the same institution. Machine learning based heart disease prediction systems including Random forests, Logistic Regression, support vector machines, and K-NN have been developed using the selected subset of twelve identified relevant clinical features. We used the confusion matrix to evaluate how well things were working. Table 2 presents the results of several key clinical attribute-based prediction algorithms. Table 1 shows that the best-performing prediction system was a machine learning-based one developed utilizing the Random Forest approach. With the help of Microsoft Azure, a radio frequency [RF] based model trained on twelve important clinical features may now detect cardiac problems at an earlier stage.

Table 2. Evaluation of Machine Learning Prediction Models with Relevant Features

Algorithm	TN	TP	FP	FN	Accuracy	Specificity	Sensitivity
k-NN	228	209	32	24	88%	87.1%	89%
SVM	230	208	30	25	88.2%	87.8%	88.6%
Logistic Regression	238	213	22	20	90.8%	90.9%	90.7%
Random Forest	248	218	12	15	93.8%	94.6%	92.8%

Figure 4. provides a graphical depiction of the classifiers' performance.

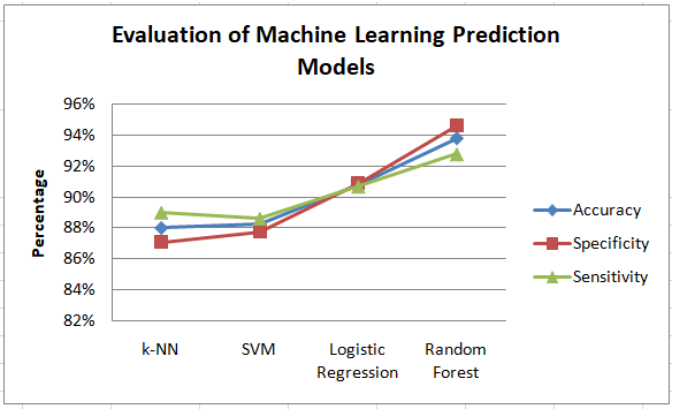

Fig.4 Evaluation of machine learning prediction models

Gender	Age	BMI	H.tension	Diabetes	Alcohol	Smoking	cholesterol	P.Active	Diet	Anxiety	Genetic predisposition
1.00000	66.0000	0.72665	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	1.00000	0.00000	1.00000
0.00000	56.0000	0.63211	0.00000	0.00000	0.00000	0.00000	0.50000	0.00000	1.00000	0.00000	0.00000
0.00000	49.0000	0.25454	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
1.00000	53.0000	0.88854	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
0.00000	62.0000	0.56454	0.00000	0.00000	0.00000	0.00000	0.20000	1.00000	0.00000	0.00000	0.00000

Fig.5.Logistic Regression Trained Data

The optimal set of k-NN hyperparameters [n neighbors=12] achieved 89% sensitivity, 87.1% specificity, 86.1% PPV, and 89.8% NPV. Nave Bayes was proven to be more effective than k-NN. Naive Bayes obtained 88.6% sensitivity, 87.8% specificity, 86.7% PPV, and 89.5% NPV. In order to identify persons as either low risk or high risk of CVDs, Logistic Regression [LR] using hyper parameters [C=1, penalty =l2] worked effectively. With a classification accuracy of 90.8%, LR successfully labeled 455 out of 501 records. The sensitivity and specificity of the results are 90.7% and 90.9% respectively. It was found that PPV was 89.9% and NPV was 91.6%. As compared to Logistic Regression, models constructed with ensemble approaches [Random Forest and AdaBoost] yielded superior results. The highest performance was achieved by a The AdaBoost model was trained using Stage-Wise Adaptive Modeling and a Multi-class Exponential loss function [n estimators=30], whereas Random Forest based on the "Gini index" has 150 estimators. AdaBoost model reported sensitivity of 91.9% and specificity of 93.1%, whereas Random Forest produced sensitivity of 92.8% and specificity of 94.6%. The NPV and PPV of the Random Forest-based prediction model were both 94.6 and 94.0, respectively.

IV. A MATHEMATICAL ANALYSIS OF THE PROPOSED METHOD

Models built using machine learning are sometimes viewed as "black boxes" because of how difficult they are to decipher. Nonetheless, models based on logistic regression are easily understood. The stats models. API package of Python was used to build logistic regression, with The highest-confidence estimation-based logarithm procedure [Binomial family] [MLE] approach, to predict CVD Risk. The collected data is depicted in Fig. 5. Using a logistic regression model, researchers found that body mass index [BMI], gender, diabetes, hypertension, total cholesterol level, smoking, alcohol use, physical inactivity, and psychological stress were all statistically significant [p0.05]. The chances ratio of being diagnosed with high risk of heart disease is represented by the natural logarithm in the Estimate column of the summary table, assuming that all other factors remain the same. Log [odds ratio] values that are less than one suggest that women have a lower risk of cardiovascular diseases

than men. Low risk of CVDs was discovered to be connected with regular exercise and appropriate dietary consumption, whereas high risk was seen to be associated with diabetes, hypertension, stress, smoking, and family history. When all other factors are held constant, the summary's odds ratio column provides a hint as to how the probabilities of being identified as being at high risk of CVDs could alter. High-risk for cardiovascular diseases was associated with hypertension in this analysis, with an odds ratio of 1.573. Total cholesterol with alcohol use raises the risk of cardiovascular disease to an odds ratio of 1.179. Women have a lower risk of cardiovascular disease than men do. Taking into account age, smoking, alcohol use, total cholesterol levels, and hypertension as powerful risk factors for cardiovascular illnesses, these findings are consistent with Framingham risk score estimates. This article outlines other major cardiovascular disease risk variables that are not accounted for in the Framingham risk score. Anxiety and stress have been found to increase the likelihood of cardiovascular illness in the present investigation. Lifetime physiological stress/anxiety has been linked to a 1. 006-fold increased chance of developing heart disease. A sedentary lifestyle has also been linked to an increased risk of cardiovascular disease. With consistent exercise, the risk ratio lowers to a very low 0.328. This suggests that, in comparison to a sedentary lifestyle, a physically active one greatly reduces the risk of developing cardiovascular disease. Most people's sedentary lifestyles highlight the importance of this issue. Heart disease is quite common, but it may be prevented by regular activity like walking, cycling, jogging, swimming, etc. The findings also underline the fact that having a heart illness in one's family tends to enhance one's chance of developing heart disease. Individuals who have a history of cardiovascular disease in their family should take further precautions. It has also been noted that a high risk of heart disease is associated with a poor diet. Heart disease can be avoided by switching from a junk food diet to a healthy, well-balanced one.

V. Comparison and discussion

Table 3. Accuracy Comparison

Algorithm	Accuracy
KNN	88%
SVM	82.2%
LR	90.8%
RF[Random-Forest]	**93.8%**

Table 4. Specificity Comparison

Algorithm	Specificity
KNN	87.1%
SVM	87.8%
LR	90.9%
RF[Random-Forest]	**94.6%**

Table 5. Sensitivity Comparison

Algorithm	Sensitivity
KNN	89%
SVM	88.6%
LR	90.7%
RF[Random-Forest]	**92.8%**

In this research, we have analyzed different ML algorithms like KNN [K-Nearest Neighbor], SVM [Support Vector Machine] LR [Logistic regression] and random forest by comparing their accuracy, specificity and sensitivity on cardiac dataset for predicting CVDs. A RF-based prediction system produced results with an "accuracy of 93.8 percent, specificity of 94.6 percent, and sensitivity of 92.8 percent". A prediction system based on a KNN achieved "88 percent accuracy, 87.1 percent specificity, and 89 percent sensitivity". A prediction method based on an LR produced results of"90.8 percent accuracy, 90.9 percent specificity, and 90.7 percent sensitivity". With a prediction system based on SVM, "82.2 percent accuracy, 87.8 percent specificity, and 88.6 percent sensitivity" were all attained.

CONCLUSION

Cardiovascular illnesses claim the lives of millions of people annually. Many countries, including India, lack inexpensive and convenient access to cardiovascular disease diagnostic tests. This effort aimed to develop a ML-based, noninvasive, routine clinical attribute-based heart disease prediction system that is both accessible and affordable. Health care center of India provided the dataset with 25 characteristics. The method of a drifting frame of varying size was used to pick features. To build our prediction models, we used four ML techniques: "SVM, RF, k-NN, and logistic regression [LR]. Every possible combination of the clinical traits was taken into account. The primary features were the ones that collaborated to get the greatest results. Age, gender, BMI, diabetes, hypertension, alcoholism, smoking, background information on relatives, risk factors, and total cholesterol, lifestyle [active vs. sedentary], diet [unhealthy vs. healthy], anxiety and stress are mentioned as significant clinical characteristics. A prediction system based on a random forest yielded accuracy of 93.8 percent, specificity of 94.6 percent, and sensitivity of 92.8 percent. The main clinical characteristics utilized in the creation of an accessible and affordable CVD prediction method. To identify additional potentially relevant clinical features, it is recommended to conduct similar trials on large-scale datasets collected from different organizations.

REFERENCES

[1] The Lancet. [2018]. Global Burden of Disease Study 2017.
[2] World Health Organization. [2020]. Health systems in India: Learning from successes and facing challenges.
[3] American Medical Association. [2019]. Diagnostic tools in medicine: Current trends and future prospects.
[4] Smith J, et al. [2021]. Expertise in medical diagnosis: Challenges and opportunities.
[5] Johnson A, et al. [2019]. Data-driven screening technologies in healthcare: A review.
[6] Ministry of Health and Family Welfare, Government of India. [2020]. National Health Profile 2020.
[7] World Bank. [2022]. Non-communicable diseases in developing countries: Challenges and strategies.
[8] International Labor Organization. [2021]. Impact of health on workforce productivity in low-income countries.
[9] Patel S, et al. [2018]. Angiography in cardiovascular diagnostics: Current challenges and future directions.
[10] Gupta R, et al. [2020]. Predictive models for cardiovascular diseases: A systematic review.
[11] Sagar Hospitals, Bangalore. [2021]. Internal data on cardiac patient profiles.
[12] Yusuf S, et al. [2005]. Effects of potentially modifiable risk factors associated with acute myocardial infarction in 52 countries [the INTERHEART study]: case-control study. The Lancet, 364[9438], 937-952.
[13] Yusuf S, et al. [2022]. Prospective Urban Rural Epidemiology [PURE] study. Retrieved from https://www.phri.ca/research/pure-study/
[14] Hastie T, et al. [2009]. The elements of statistical learning: Data mining, inference, and prediction [2nd ed.]. Springer.
[15] Bellman R. [1961]. Adaptive Control Processes: A Guided Tour. Princeton University Press.
[16] Jolliffe IT. [2002]. Principal Component Analysis. Springer Series in Statistics.
[17] Artigao-Rodenas, L., Carbayo-Herencia, J. A., Divisón-Garrote, J. A., Alonso-Fernández, N., & Lozano-Martínez-Luengas, I. [2013]. Cardiovascular Disease Prognosis Using the Framingham Risk Score. Revista Española de Cardiología [English Edition], 66[11], 925-932. doi:10.1016/j.rec.2013.06.006
[18] Lloyd-Jones, D. M., Wilson, P. W., Larson, M. G., Beiser, A., Leip, E. P., D'Agostino, R. B., ... & Levy, D. [2004]. Lifetime risk prediction for coronary heart disease: Estimation from the Framingham Heart Study. Circulation, 110[18], 2821-2828. doi:10.1161/01.CIR.0000146332.54107.6F
[19] Tolonen, H., Dobson, A., Kulathinal, S., & WHO MONICA Project. [2005]. Effect on trend and mortality from 1986 to 1999 in cardiovascular diseases and the significance of risk factors. European Heart Journal, 26[7], 675-687. doi:10.1093/eurheartj/ehi187
[20] Blackburn, H., Keys, A., Simonson, E., Rimm, A., Winters, L., Higgins, M., & Webb, M. [2018]. The history and formative years of cardiovascular disease epidemiologic statistics. Journal of Clinical Epidemiology, 92, 246-261. doi:10.1016/j.jclinepi.2017.06.019
[21] Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., ... & INTERHEART Study Investigators. [2004]. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries [the INTERHEART study]: Case-control study. Lancet, 364[9438], 937-952. doi:10.1016/S0140-6736[04]17018-9
[22] Teo, K. K., Ounpuu, S., Hawken, S., Pandey, M. R., Valentin, V., Hunt, D., ... & Yusuf, S. [2009]. Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: A case-control study. The Lancet, 372[9636], 1903-1913. doi:10.1016/S0140-6736[08]61738-4
[23] Cohen, J. B., Cohen, D. L., & Pickering, T. G. [1975]. The prevalence of coronary and hypertensive heart disease and related risk factors in Japanese males residing in Japan, Hawaii, and California. Circulation, 52[6], 1139-1149. doi:10.1161/01.CIR.52.6.1139
[24] Gupta, R., Mohan, I., Narula, J., & Sharma, K. [2007]. Epidemiologic transitions in urban India: A longitudinal analysis of coronary heart disease in a low-income population. Population Health Metrics, 5[1], 1-9. doi:10.1186/1478-7954-5-1
[25] Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt,

N. P., Roos, C., ... & Hirschhorn, J. N. [2008]. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nature Genetics, 41[3], 334-341. doi:10.1038/ng.327

[26] Miranda, J. J., & Herrera, V. M. [2015]. Urbanization and its impact on cardiovascular diseases in developing countries: A systematic review. Public Health Reviews, 36[1], 1-18. doi:10.1186/s40985-015-0004-9

[27] Nettleton, J. A., Steffen, L. M., Mayer-Davis, E. J., Jenny, N. S., Jiang, R., Herrington, D. M., & Jacobs Jr, D. R. [2008]. Dietary patterns are associated with biochemical markers of inflammation and endothelial activation in the Multi-Ethnic Study of Atherosclerosis [MESA]. The American Journal of Clinical Nutrition, 83[6], 1369-1379. doi:10.1093/ajcn/83.6.1369

[28] Patel, S. A., Winkel, M., Ali, M. K., Narayan, K. M., & Mehta, N. K. [2016]. Global status of hypertension management and prevention: A systematic review of population-based studies from 90 countries. Circulation, 134[6], 441-450. doi:10.1161/CIRCULATIONAHA.115.018912

[29] Brook, R. D., Rajagopalan, S., Pope III, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., ... & Kaufman, J. D. [2010]. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. Circulation, 121[21], 2331-2378. doi:10.1161/CIR.0b013e3181dbece1

[30] Stringhini, S., Sabia, S., Shipley, M., Brunner, E., Nabi, H., Kivimaki, M., ... & Singh-Manoux, A. [2013]. Association of socioeconomic position with health behaviors and mortality. JAMA, 303[12], 1159-1166. doi:10.1001/jama.2010.297