Journal of Advances and Scholarly Researches in Allied Education
Vol. 19, Issue No. 6, December-2022, ISSN 2230-7540

Identification of Significant Clinical Attributes for Developing Heart Disease Prediction System

Kunal Gupta[1]*, Dr. Faizanur Rahman[2]

[1] Research Scholar, Kalinga University

[2] Research Guide, Deparment of Computer Science, Kalinga University.

Abstract- Health care, also known as medical care or medical service, is a system that aims to improve health-related services in order to meet the clinical needs of individuals. Disease diagnosis is a process of establishing the type of disease from the symptoms and sign. The decision on a particular type of disease is reasoned from collected information from patient's history and examination. This paper describes a method for identification of significant attributes accountable for heart diseases. The floating window method for feature selection proposed in this chapter has been applied to two datasets namely, Cleveland heart disease dataset collected from UCI machine learning repository and a cardiac dataset collected from Sir Ganga Ram Hospital in Delhi. CVD prediction system developed using significant clinical attributes is quite affordable and easily accessible. It is proposed to perform similar experiments on voluminous datasets collected from various hospitals so that other potential significant clinical attributes can be identified.

Keywords- Healthcare, Clinical, Heart Diseases, Machine Learning

----------X----------

INTRODUCTION

In order to diagnose a disease, the medical practitioner relies on various laboratory tests, physical examination, and sometimes even on invasive tests. A highly qualified and experienced doctor with a deep in-depth domain knowledge is needed to diagnose the disease in early stages. Machine learning based diagnostic tools may be effectively used as a helping aid for medical personnel. Diagnosis of heart diseases is dependent on invasive as well as noninvasive medical tests. Angiography is considered to be a gold standard test for heart disease diagnosis. However, in India, these tests are neither affordable nor easily accessible for a major section of people living in remote areas. Heart disease prediction models built using noninvasive clinical attributes shall be highly beneficial to Indian population.

METHODOLOGY

Study Setting : Use of anonymized dataset ensured adherence to data privacy regulations.

Data Collection : Data was collected from Delhi Hospitals, India. A panel of cardiologists was consulted to understand all the risk attributes associated with heart diseases in India. Twenty- five such clinical attributes were identified. To build ML prediction model, a diagnostic dataset of 1670 records was gathered from a tertiary hospital in India. Twenty- five such clinical attributes were identified. To build ML prediction model, a diagnostic dataset of 1670 records was gathered from a tertiary hospital in India. This dataset contains anonymized information of 777(46.53%) healthy persons and 893(53.47%) heart patients. A few exclusion criteria were applied before collecting the data by a random selection. Later, a dataset of 501 records was also collected from the same hospital to validate the results of prediction system.

Exclusion criteria : Exclusion criteria applied in this study includes1) people using anti- depressants and medicines for chronic diseases like severe mental illness, atrial fibrillation, tuberculosis, asthma, cancer, and chronic kidney disease for a long time 2) Pregnant females were excluded from the study 3) people younger than 20 years of age 4) Patients on antipsychotic drugs, oral corticosteroids, and immune suppressants.

Descriptive Characteristics of Study Population

In this study, healthy person refers to a person who was not diagnosed with any disease. Of 1670 records in the dataset, 53.47% (893 records) were established cases of heart diseases while the rest 46.53 % (777 records) were identified to be healthy. Nearly equal number of records of healthy and CVD patients assures that dataset is balanced. Balanced dataset assures that it is not biased in favor of any class. There were 881 medical records of males and

789 records of females confirming equal participation. Average age of heart patients was observed to be 66.2 years and average age of healthy people was 57.2 years. Statistical significance was determined by calculating t-test and Chi square test for numeric and categorical attributes, respectively. Categorical attributes have been shown with count (%) and numerical input variables have been represented with average (with standard deviation) in Table 1.

Table 1: Descriptive Characteristics of Indian dataset

Clinical Attribute	Unit	Healthy (count=777)	Heart Diseases (count =893)	p-value
Age	Years (SD)	57.3(12.4)	66.2(11.2)	<0.001
Gender(female)	(%)	545(70.1)	244(27.3)	<0.001
Body Mass Index	Kg/cm^2	26.1(5.3)	28.3 (3.2)	0.161
Waist Circumference	cm	87.1(8.2)	99.1(5.4)	0.232
Total cholesterol levels	mg/dL (SD)	218.4(13.9)	267.7(14.1)	<0.001
HDL cholesterol	* mmol/L (SD)	1.46 (0.43	1.39 (0.41)	< 0.001
LDL cholesterol	mmol/L (SD)	3.40 (0.88)	3.45 (0.91)	< 0.001
Triglycerides	mmol/L (SD)	1.37 (0.63)	1.89 (0.75)	< 0.01
Hypertension (yes)	(%)	182(23.4)	614(68.7)	<0.001
Diabetes (yes)	(%)	318(40.9)	630(70.5)	<0.001
Heart Rate	bpm	79(9.3)	83(10.2)	0.031
Fasting Blood Sugar		100(9.3)	130(10.4)	0.211
HbA1c+	% (SD)	5.64 (1.64)	7.26 (1.61)	< 0.01
Systolic blood pressure	mm HG	127 (17.2)	153 (15.6)	< 0.001
Serum Fibrinogen	g/L (SD)	3.43 (1.35)	3.96 (1.12)	0.159
CRP	mg/L (SD)	7.33 (10.5)	11.2 (12.7)	< 0.001
Smoking	(%)	258(33.2)	570(63.8)	<0.001
Serum creatinine	umol/L	85.6 (14.0)	94.9 (12.3)	.234
Exercise (yes)	(%)	737(94.8)	412(46)	<0.001
Alcohol (yes)	(%)	305(39.2)	623(69.7)	<0.001
Stress (yes)	(%)	352(45.3)	568(63.6)	<0.001
gamma GT	IU/L (SD)	32.3 (13.6)	43.3 (23.7)	0.341
AST/ALT ratio	(SD)	1.01 (0.25)	1.06 (0.36)	0. 09
Healthy Diet (yes)	(%)	398(51.2)	496(55.5)	0.077
Family History of CVD (yes)	(%)	299(38.4)	592(66.2)	<0.001

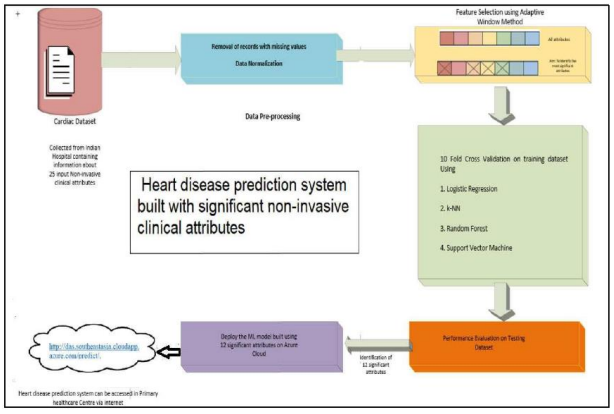

Figure 1 Workflow of the Study Carried Out On Indian Dataset

Data Preprocessing

Data preprocessing is an integral task for building a robust ML prediction system. The dataset was looked out for any missing values and outliers. None of the records had any missing values or outliers in the data.

The categorical attributes were encoded and converted into numeric. Since the attributes have different ranges, z- score standardization was carried out bring them on same scale.

$$z = \frac{actual\ value - mean\ value}{standard\ deviation}$$

Proposed method

Seventy five percent of the data records were used for training the ML module while the remaining twenty five percent were used for evaluating the performance of the system. To select the most significant attributes, various subsets of different combinations of attributes were chosen to be used with four algorithms Logistic regression, k-NN, support vector machine (SVM) and Random Forest. To determine the most significant attributes, it was decided to assess the performance of every possible combination of clinical attributes and each of four machine learning algorithms. The scheme has been represented in Figure 3.2. It is known that the total number of possible combinations attainable from a set of n attributes, is 2n. If the empty set is eliminated, the total possible combinations are 2n−1.

In this study, there were n=25 input attributes, hence, a total number of possible combinations is 2^{25} -1= 3,35,54,431. All these possible combinations of input attributes were tried and fed to ML techniques to build the prediction models.

Details of feature selection technique used in this study are represented here. The feature selection module employs a search scheme to identify the best possible subset of clinical attributes which are applied to each of four machine learning techniques. The feature selection unit utilizes an adaptive window frame which continually scans the input attribute vector.

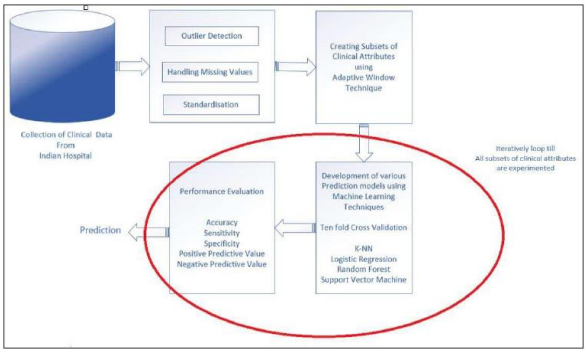

Figure 2 Proposed Methodology

At the start, the magnitude of window frame is set to 1 and, is placed at the left most side of the input feature vector (n=25). The attribute on which the window/frame is placed, is excluded from the feature subset. The left-over (n-1) attributes make up the subset of features which are then provided to the machine learning algorithms for building the prediction system. The performance of the prediction system built using this subset of features was evaluated on test dataset.

In the following step, the window was shifted towards the right direction. Now, the feature, where this window frame was positioned, was eliminated. Remaining (n-1) attributes are used to train the prediction models. Evaluation was again carried out. This entire process of shifting the window frame and elimination of the feature was carried out till the window frame reached the last (nth) attribute. This completed the first cycle of feature selection. Size of the feature subset was n-1 in the first cycle. Fig.2 represents the technique of identification of significant attributes using floating window frame method. The attribute shown in red color is removed while the remaining attributes are used to train the models.

In the second round, the size of the window frame was increased to 2. It means that two attributes were removed each time the window frame was floated from left to right side of attribute vector. The left-out n − 2 attributes were selected to train machine learning models. Performance was evaluated for each ML technique. The results of the experiment were compared with the finest performance attained on the preceding subgroup of attributes. The best performance as well as optimal subset of attributes were updated whenever the performance was found to have improved than the previous best performance. The process was reiterated till the window frame approached the extreme right side of the feature vector.

In the third round, the size of frame was increased to 3. The process was continued till n-1 th round where the frame size was increased to n-1. Whenever an improvement in performance was observed, the optimal subset of features was updated accordingly. Ultimately, the subgroup of attributes which generated the best values performance metrics, was identified as the optimal subset of significant attributes. Window technique for feature selection has been illustrated in Figure 3.

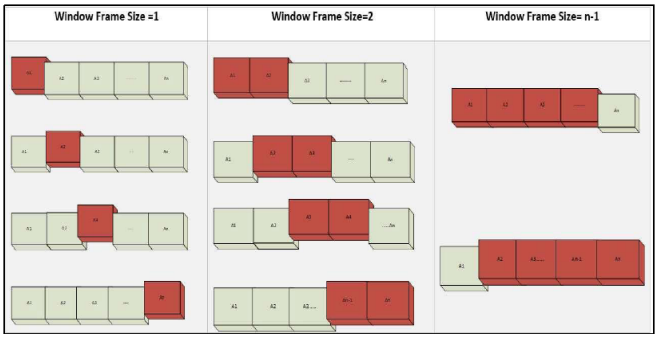

Figure 3 Various Stages of Proposed Algorithm

Prediction modelling using machine learning technique

For each round of feature selection, four machine learning algorithms were employed to develop the prediction models. These algorithms are k-NN, Logistic regression, Random Forest, and support vector machine. The performance of the models was validated using ten- fold cross validation.

Performance Evaluation Metrics

The performance of the prediction system was assessed in terms of accuracy, specificity and sensitivity. Confusion matrix was analyzed to determine the performance of heart disease prediction system. There are four main elements of a confusion matrix: true positives, false positives, true negatives, and false negatives.

RESULTS

The results of this study have been shown in this section. The performance of four machine learning techniques on all the combinations of clinical attributes were examined one by one. Performance metrics namely accuracy, specificity, sensitivity have been tabulated in Tables respectively. Among all the possible combinations of clinical attributes, the combination of features which accounted for the highest performance was identified.

Table 2 Highest Classification accuracy achieved by various ML methods

ML technique	Highest Classification Accuracy	Combination of attributes
Logistic Regression	86.2%	Age, Diabetes, TC, HDL, EX, FH, HT, HR, AL, SM, Gender, WC
k-NN	87%	Diabetes, HT, TC, SM, HD, TG, Gender, ST, AL, SF, BMI
Support Vector Machine	86.8%	TC, HDL, EX, FH, gender, BMI, FBS, CRP, CP, TG
Random Forest	90.1%	Age, Diabetes, LDL, SM, HD, BMI, ST, AL, SC, Gender

It is clear from Table 2 that Logistic Regression based system attained a maximum accuracy of accuracy of 86.2% when trained on input attributes like age, diabetes, total cholesterol, HDL, Exercise,

family history, hypertension, heart rate, alcohol, smoking, gender, and waist circumference.

SVM performed better than Logistic Regression attaining an accuracy of 86.8% when trained on total cholesterol, HDL, exercise, family history, gender, BMI, Fasting blood sugar, CRP, and chest pain. Of all techniques, Random Forest performed the best and attained a maximum accuracy of 90.1% when trained on Age, HbA1c, LDL, Smoking, HD, BMI, Stress, Alcohol, Serum creatinine and Gender.

Table 3 Highest Sensitivity achieved by various ML methods

ML technique	Highest Sensitivity	Combination of attributes
Logistic Regression	87.2%	Age, Diabetes, TG, EX, FH, ST, CP, HD, BMI, SF
k-NN	85.4%	Age, Gender, TC, HDL, LDL, FH, AL, SM, ST, CRP, HR, WC
SVM	83.5%	Gender, Diabetes, TC, HT, EX, FH, HD, WC, FBS, GGT, LDL
Random Forest	91%	Age, Gender, Diabetes, HT, EX, AL, SM, BMI, HDL, HR, SC

Highest sensitivity attained using SVM was 83.5% while that of k-NN was observed to be 85.4%. Logistic regression when fed with input attributes age, HbA1c, Triglycerides, exercise, family history, stress, chest pain, diet habits, BMI, and serum fibrinogen attained the highest sensitivity of 87.2%. The highest sensitivity attained by Random Forest was observed to be maximum at 91% when the combination of age, gender, HbA1c, hypertension, exercise, alcohol, smoking, BMI, HDL, heart rate, serum creatinine was fed as input features.

The highest specificity attained using k-NN and SVM were nearly 86% and 86.2% respectively. The highest specificity scored by Logistic regression was 88.7% when the input attributes were age, hypertension, HbA1c, diet habits, BMI, family history and stress/anxiety, AST/ALT ratio, total cholesterol, triglycerides, exercise, and smoking. The highest specificity attained using RF was 93%.

The input attributes which made RF attain the highest specificity were age, gender, HbA1c, diet habits, smoking, alcohol, stress, exercise, total cholesterol, and chest pain. It is clear from Tables 3.2-3.4 that Random Forest performed the best of all the machine learning techniques applied in this work.

Table 4 Highest specificity achieved by various ML methods.

ML technique	Highest specificity	Combination of attributes
Logistic Regression	88.7%	Age, HT, Diabetes, HD, BMI, FH, ST AST/ALT, TC, TG, EX, SM
k-NN	86.2%	Gender, TC, FH, SM, AL, BMI, FBS, LDL, CRP, FEV1, Diabetes, CP, WC
SVM	86%	Age, Gender, HT, FH, BMI, HDL, HR, EX, AL, LDL, FBS
Random Forest	93%	Age, Gender, Diabetes, HD, SM, AL, ST, EX, TC, CP, TG,

Table 5 Role of clinical attributes on performance

Attributes	Attribute Code	Occurrence Highest Accuracy	Occurrence Sensitivity	Occurrence Specificity	Total Frequency
Age	Age	2	3	3	8
Gender	Gender	4	3	3	10
Body Mass Index	BMI	3	2	3	8
Waist Circumference	WC	2	1	1	4
Cholesterol levels	TC	3	2	3	8
HDL cholesterol	HDL	2	2	1	5
LDL cholesterol	LDL	1	2	2	5
Triglycerides	TG	2	1	2	5
Hypertension	HT	2	2	2	6
Diabetes	Diabetes	3	3	3	9
Fasting Blood sugar	FBS	2	1	1	3
Heart Rate	HR	1	1	1	3
FEV1	FEV	0	0	1	1
gamma GT	GGT	0	1	0	1
C-reactive protein (CRP)	CRP	1	1	1	3
Serum fibrinogen	SF	1	1	0	2
Serum creatinine	SC	1	1	0	2
AST/ALT ratio	AST/ALT	0	0	1	1
Chest Pain	CP	1	1	2	4
Alcohol	AL	3	2	3	8
Smoking (last 5 years)	SM	2	2	2	6
Exercise (Weekly 3 Hours)	EX	2	3	3	8
Stress	ST	2	2	2	6
Family History CVD	FH	2	3	3	8
Healthy Diet	HD	2	2	2	6

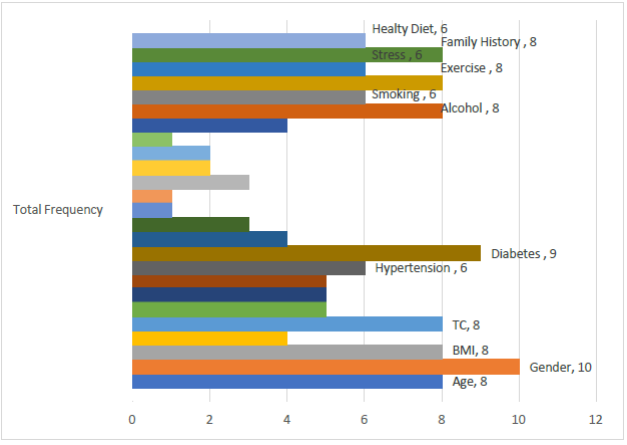

Figure 4 Frequency of occurrence of significant clinical attributes for CVD prediction

The significant noninvasive clinical attributes identified in this study for heart disease prediction are gender, age, body mass index, hypertension, Diabetes (HbA1c>7), alcohol consumption, family

history, total cholesterol, exercise, smoking, intake of healthy diet and stress/anxiety in life.

VALIDATION

To validate the result of this study, an independent dataset of 501 records was collected from the same hospital. The subgroup of twelve identified significant clinical attributes was shortlisted to develop machine learning based heart disease prediction systems using four techniques namely k-NN, logistic regression, support vector machines and random forest.

The performance was assessed using the confusion matrix. Table 6 discusses the performance of these prediction systems developed using the significant clinical attributes. It is clear from Table 9 that machine learning based prediction system built using Random Forest technique performed the best. This RF based model trained on twelve significant clinical attributes was deployed on Microsoft Azure for early diagnosis of heart diseases.

Table 6 Performance of ML based prediction models using significant attributes

Algorithm	TN	TP	FP	FN	Accuracy	Specificity	Sensitivity
k-NN	230	211	34	26	88%	87.1%	89%
SVM	232	210	32	27	88.2%	87.8%	88.6%
Logistic Regression	240	215	24	22	90.8%	90.9%	90.7%
Random Forest	250	220	14	17	93.8%	94.6%	92.8%

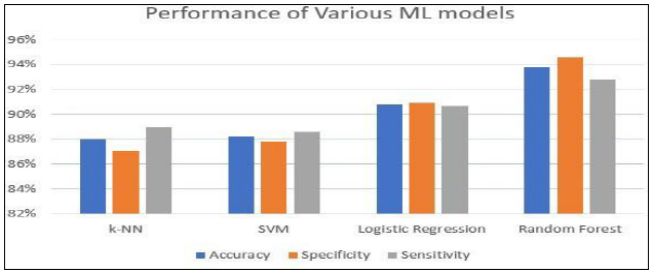

Figure 5 Performance of Various ML based heart disease prediction system

The best hyperparameters for k-NN (n_neighbors=12) resulted in a performance of sensitivity 89%, specificity 87.1%, PPV86.1%, NPV 89.8%. The performance of Naïve Bayes was found to better than k-NN. Sensitivity 88.6%, specificity 87.8%, PPV 86.7%, NPV 89.5% were achieved by Naïve Bayes.

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

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	-2.250	2.931	-2.19	0.034	
Age	0.035	0.006	10.31	0.087	1.035
Height	-0.004	0.005	-6.23	0.533	0.996
Weight	0.076	0.013	0.82	0.002*	1.078
Gender(female)	-0.238	0.007	-8.68	0.025*	0.788
Diabetes	0.029	0.312	0.07	0.0359*	1.029
Hypertension	0.453	0.004	4.08	0.001*	1.573
Total Cholesterol	0.165	0.041	1.08	0.003*	1.179
Smoke	0.093	0.155	0.47	0.006*	1.097
Alcohol	0.165	0.133	0.35	0.035*	1.179
Exercise	-1.113	0.004	-0.54	0.001*	0.328
Family History	0.003	0.002	3.42	0.054	1.003
Diet	-0.012	0.006	-6.23	0.533	0.988
Stress	0.006	0.232	0.623	0.003*	1.006

No. of Fisher Scoring Iterations-6

AIC 225.1

*Attributes are statistically significant (ρ < 0.05)

Figure 6 Mathematical interpretation of Prediction model

DISCUSSIONS

During the past few years, extensive studies have been done to develop ways for early diagnosis of CVDs. However, due to unaffordable and inaccessible diagnostic tests used in these studies, these research works cannot be utilized efficiently for low- and middle-income group countries like India in present scenario. Feature selection carried out using a drifting window frame of variable size in this study, helps in identification of significant clinical attributes which can be used to develop ML based heart disease prediction system. Twelve significant clinical attributes are: gender, age, body mass index, hypertension, diabetes (HbA1c) >7, alcohol consumption, family history, total cholesterol, sedentary lifestyle, healthy diet, smoking and stress/anxiety. Tests for hypertension, total cholesterol, diabetes (HbA1c) can be easily carried out in primary healthcare centers. The other attributes namely, age, family history, diet habits, smoking, alcohol consumption etc. can be easily enquired from the person. This ensures the cost effectiveness of the prediction system.

The prediction model trained on significant attributes using Random Forest performed the best with an accuracy of 93.8%. Low-cost, high-performance CVD prediction system created in this study is easily accessible via internet at http://das.southeastasia.cloudapp.azure.com/predict/. An authorized medical practitioner in primary healthcare center can fill in the input details to investigate the risk of CVD in an individual. Figure 7 shows a demo screenshot of the CVD prediction model. For ease of use, weight and height are fed as inputs (rather than BMI). BMI is internally calculated by the system. A value of HbA1c greater than 7 indicates that a person is diabetic. HbA1c>7 has been encoded as presence of diabetes in this prediction system.

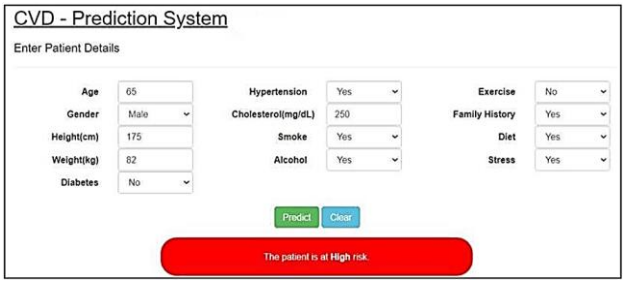

Figure 7 Screenshot of the CVD prediction system created using significant clinical attributes.

CONCLUSION

Millions of lives are lost every year due to heart diseases. CVD diagnostic tests like angiography etc. are neither affordable nor easily accessible in many countries like India. This study was aimed at development of machine learning based cost effective and easily accessible heart disease prediction system using noninvasive significant routine clinical attributes. A dataset of 25 attributes was collected from an Indian hospital. Feature selection was carried out using the technique of a drifting frame of variable size. Four ML algorithms namely logistic regression, k-NN, Support vector machine and Random Forest were used to develop prediction models. All the possible combinations of clinical attributes were analyzed. The combination of attributes which resulted in the best performance were considered significant. Significant clinical attributes identified in this study are: gender, age, body mass index, hypertension, diabetes, alcohol consumption, smoking, family history, total cholesterol, sedentary lifestyle, healthy diet, and stress/anxiety. Random forest-based prediction system attained the best performance with an accuracy of 93.8%., specificity of 94.6% and sensitivity of 92.8%.

REFERENCES

1. A. Davari Dolatabadi, S. E. Z. Khadem, and B. M. Asl, “Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM,” Comput. Methods Programs Biomed., vol. 138, pp. 117–126, Jan. 2017
2. J. L. Warford, *Environment, Health and Sustainable Development: the Role of Economic Instruments and Policies*, Discussion Paper: Director-General's Council on the Earth Summit Action Programme for Health and Environment. Geneva, World Health
3. K. Kourou, T. Exarchos, K. Exarchos, M. Karamouzis and D. Fotiadis, "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal, vol. 13, pp. 8-17, 2015. Available: 10.1016/j.csbj.2014.11.005 [Accessed 28 May 2021].
4. K. Priyanka and N. Kulennavar, “A survey on big data analytics in health care”, *International Journal of Computer Science and Information Technologies*, vol. 5, no. 4, pp. 5865–5868, 2014
5. Maini, E., Venkateswarlu, B., Gupta. Data Lake-An Optimum Solution for Storage and Analytics of Big Data in Cardiovascular Disease Prediction System. International Journal of Computational Engineering & Management. 2018, Vol.21, Issue 6, pp 33-39
6. Maini, E., Venkateswarlu, B., Maini, B., Marwaha D. Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India. Medical Journal Armed Forces India.2021; ISSN 0377 1237, https://doi.org/10.1016/j.mjafi.2020.10.013.
7. National Program for Prevention and Control of Cancer, Diabetes, CVD and Stroke(NPCDCS) | National Health Portal Of India", Nhp.gov.in, 2021. [Online]. Available: https://www.nhp.gov.in/national-programme-for-prevention-and-control-of-c_pg. [Accessed: 28- May- 2021
8. Tseng, CJ, Lu, CJ, Chang, CC, Chen, GD & Cheewakriangkrai, C 2017,’Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence’, Artificial Intelligence in Medicine, vol. 78, no. C, pp. 47-54.

Corresponding Author

Kunal Gupta*

Research Scholar, Kalinga University