Prediction of Cardiovascular Disease Using Feature Selection Techniques

Priya Singh, Gyanendra Kumar Pal, and Sanjeev Gangwar

***Abstract*—Cardiovascular diseases is one of the harmful diseases and many people suffered from this disease across the world. In the field of the Healthcare Industry, on-time and efficient prediction of cardiovascular diseases plays a prominent role in healthcare. Currently, the Medicare industry is "Data-rich" yet "Insight poor". The aim of this research work is to develop an efficient and accurate system to inspect cardiovascular diseases and the system is based on data mining techniques that can help to remedy such a situation. The system is developed based on classification algorithms like Random Forest, Logistic Regression, Naive Bayes' and Support Vector Machine while feature selection algorithm has been used like Pearson Correlation and Chi-Square in order to increase the accuracy and reduce the execution time of classification systems. With these results, it is found that Logistic Regression achieved the highest accuracy of 84% as compared to the others.**

***Index Terms*—Cardiovascular diseases (CVDs), Chi-square, data mining, logistic regression, Naive Bayes', Pearson correlation, random forest, Support Vector Machine (SVM).**

I. Introduction

Among all fatal diseases, cardiovascular diseases are treated as the most popular and critical health issue and globally many people suffered from these diseases. Cardiovascular Diseases have common symptoms such as dyspnea, chest pain, weakness or coldness in your legs or arms etc. In this rapid-moving world, people want to live an expensive life so they work like a machine in order to earn a lot of money that leads to an unhealthy diet, obesity, high blood pressure, high cholesterol, depression and smoking [1]. These are some factors that increase the possibility of heart disease among numerous people. The report from the World Health Organization shows us a large number of people that die every year due to cardiovascular disease all over the world. An approximated 17.9 million people died from heart diseases in 2016, that represents 31% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. Over 75% of CVD deaths take place in low- and middle-income countries where high blood pressure happens to be amongst the most important risk factors for CVDs. In 2016 India reported 63% of total deaths due to Non-Communicable Diseases (NCDs), of which 27% were attributed to CVDs. CVDs also responsible for 45% of deaths in the age group of 40 to 69. Individuals at risk of CVD may demonstrate raised blood pressure, glucose, and lipids as well as overweight and obesity. Identifying those at the highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths [2].

Data mining is the process of extracting previously hidden patterns and trends in databases and applying that information to construct predictive models. The application of data mining has been used in various disciplines such as web mining, customer relationship management, mobile computing, marketing, engineering and medicine analysis. Data Mining uses a combination of statistical analysis, machine learning and database technology in order to extract hidden patterns and relationships from the large database.

Nowadays, datamining plays a vital role in the field of the healthcare industry. It provides a meaningful base for critical decisions. Researchers apply various data mining techniques for the identification of CVDs is respect to the accuracy and execution time.

Traditionally, diagnosis of CVDs done by the analysis of the medical history of the patient, physical examination report and analysis of concerning symptoms by a physician. But the results obtained from this diagnosis technique are not accurate in identifying the patient of heart disease. Moreover, it is expensive and computationally difficult to analyze.

In this research work, four different machine learning techniques such as Random Forest, Logistic Regression, Naive Bayes' and SVM are used in order to predict heart disease.

The Proposed model forms a basic procedure for carrying out the CVDs predication using any machine learning techniques. The performances of various machine learning classifiers for CVDs detection have been checked on selected features. The Pearson correlation and Chi-square feature selection algorithm have been used for extracting an optimal set of features. The model's performance measuring metrics include accuracy, precision, recall and F-1 score.

The continuing part of this research paper consists of Section II as Related Works, Section III as Dataset Description, Section IV as Methodology, Section V as Data Mining Techniques, Section VI as Experimental Result and Discussion and Section VII as Conclusion.

II. Related Works

H. Beley *et al.* [3] have done research work in which three data mining algorithms like Random Forest, Decision Tree and Naive Bayes' are used in order to develop a prediction system. After the classification and performance evaluation, the Random Forest is considered as the best with 81% precision as compared to other algorithms.

Boshra Bahrami *et al.* [4] describes three different classification techniques such as Decision Tree, KNN and Naive Bayes in the diagnosis of heart disease. It is observed that classification by Decision Tree performs the best

Manuscript received February 11, 2022; revised May 25, 2022.

Priya Singh, Gyanendra Kumar Pal, and Sanjeev Gangwar are with VBS Purvanchal University, India (e-mail: sspriyasingh1994@gmail.com, gangwar.sanjeev@gmail.com).

contrast to the other two techniques.

Chaurasia and Pal *et al.* [5] led concentrate on the forecast of coronary failure risk levels from the coronary illness data set with information mining method like Naïve Bayes, J48 decision tree and Bagging approaches and CART, ID3 and Decision Table. The result shows that Bagging procedures execution is more exact than Bayesian order and J48.

Shantakumar B.Patil *et al.* [6] have proposed an intelligent and effective heart disease prediction system based on Multi-Layer Perceptron with Back-propagation in addition to MAFIA algorithm used for the mining of frequency patterns of the heart disease from a dataset.

M.hanumathappa *et al.* [7] presented attributes selection method to predict heart disease. Thirteen attributes are reduced to Six attributes. They concluded that SVM (97.9%,89.4%) has highest accuracy out of other two simple logistic (69.2%,71.6%) and Multilayer Perceptron (74.3%,79.1%) techniques in two states.

Sellappan Palaniappan*et al.* [8] discussed data mining techniques to predict heart disease based on Decision Tree, Naive Bayes' and Neural Network and develop a model refer as Intelligent Heart Disease Prediction System (IHDPS).

Deepika N. *et al.* [9] have proposed Pruning Classification Association Rule (PCAR) to predict heart disease that derives from the Apriori algorithm. A frequent itemset is obtained from itemsets by removing minimum frequency items with minimum frequency itemsets.

Salha M.Alzahani*et al.* [10] developed heart disease classification system using data mining techniques such as Decision Tree, Artificial Neural Network and SVM. The observation was that SVM has the highest accuracy with 84.12% as compared to the other two techniques.

W. J. Frawley *et al.*[11] have carried out a research work in which a 10-fold cross-validation method is used to find out the best heart disease prediction model based on performance.

O.W.Samuel *et al.* [12] have built an intelligent and effective heart attack prediction system based on Artificial Neural Network and Fuzzy Analytic Hierarchy Process (AHP). The Performance of the proposed system is 91.10% in terms of accuracy.

M. Gudadhe *et al*. [13] have proposed a heart disease diagnosis system based on Multi-Layer Perceptron and Support Vector Machine (SVM) data mining classification technique. The accuracy of the proposed system is 80.41%.

Chaurasia and Pal *et al.* [14] proposed utilizing information mining ways to deal with distinguish coronary illness. The WEKA information mining instrument is utilized which contains a bunch of AI calculations for the purpose of mining.

Naive Bayes, J48 and bagging are utilized for this viewpoint. The UCI AI research center gives an informational collection on coronary illness that incorporates 76 ascribes. Just 11 ascribes are utilized for expectation. Naive bayes offer 82.31% precision. J48 gives 84.35% exactness. 85.03% of the exactness is acquired by bagging gives a superior grouping rate on this informational index.

III. Dataset Descriptions

In this study, the Statlog Heart Disease dataset [15] of the UCI repository is considered for testing purposes. It furnishes an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. The dataset contains 270 instances and 13 attributes with no missing values. The target label has two classes to describe the absence of CVDs and the presence of CVDs. (See Table I).

TABLE I: Dataset Description of CVDs

S.No.	Attribute Name	Attribute Code	Description	Type	Range
1.	Age	age	Patient's age in completed years.	Numeric	29-65
2.	Sex	sex	Patient's Gender	Nominal	Male=0, Female=1
3.	Chest Pain	Cp	Types of Chest Pain	Nominal	Typical Angina=1, Atypical Angina=2, Non-Anginal Pain=3, Asymptomatic=4
4.	Resting Blood Pressure	trestbps	Level of blood pressure at resting mode(in mm/Hg at the time of admitting in the hospital)	Numeric	92-200
5.	Serum Cholesterol	SerumCho	Serum Cholesterol in mg/dl	Numeric	126-564
6.	Fasting Blood Sugar	fbs	Blood sugar levels on fasting> 120mg/dl	Nominal	Yes=1, No=0
7.	Resting Electrograp-hic	restecg	Results of electrocardiogram while at rest	Nominal	Normal state=0,Abnormality inST-T wave=1,any probability or certainity of LV hypertrophy by Estes' criteria=2
8.	Maximum Heart Rate	Thalach	Maximum rate of the heart	Numeric	82-185
9.	Exercise Induced Angina	Exang	Angina induced by exercise	Nominal	Yes=1, No=0
10.	Oldpeak	oldpeak	Exercise-induced ST depression in comparison with the state of rest	Numeric	71-202
11.	PeakSlope	peakslope	ST segment measured in terms of the slope during peak exercise	Numeric	Unslopping=1, Flat=2, Downsloping=3
12.	Number of vessels	numVes-sels	Fluoroscopy coloured major vessels	Numeric	0-3
13.	Thallium Scan	thal	Status of the heart	Nominal	Normal=3, Fixed defect=6, Reversible defect=7
14.	Target	target	Detection of CVDs	Nominal	Yes=2, No=1

A. Class Distribution

target:
1
2
dtype: int64

The class level distribution '1' depicts False positive while '2' depicts True positive in the target variable of the dataset. In the given dataset, the ratio between CVDs and Non CVDs is 5:4 which is shown with the help of bar-graph.

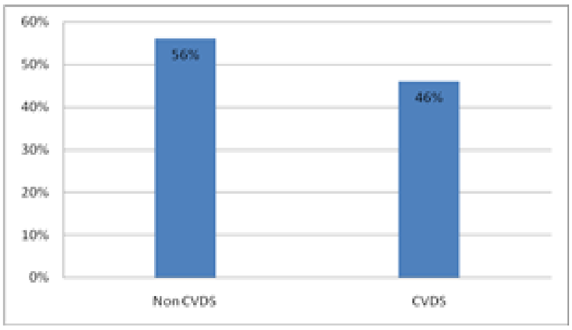

Fig. 1. Representation of Non CVDs vs CVDs.

All the experiment has been performed in an Anaconda Distribution environment while Python platform version: 3.8.5 have been used to analyze medical datasets and evaluate the performance of machine learning techniques applied to these sets.

B. Box and Whisker Plot

Box and Whisker Plot is an advantageous approach to outwardly showing the information conveyance through their quartiles [16].

The lines it is known as the whisker, which is utilized to demonstrate fluctuation outside the upper and lower quartiles to expand equal from the containers. Exceptions are at times plotted as individual spots that are in-accordance with stubbles. Box Plots can be drawn either upward or on a level plane.

The five capacities include characterizing a box and whisker plots are:

Lower value: The littlest worth in the informational index

First quartile: The worth underneath which the lower 25% of the information are contained

Median value: The center number in a scope of numbers

Third quartile: The worth above which the upper 25% of the information are contained

Upper value: The biggest worth in the informational index.

Fig. 1 addressed box and whisker plot of features of coronary illness dataset.

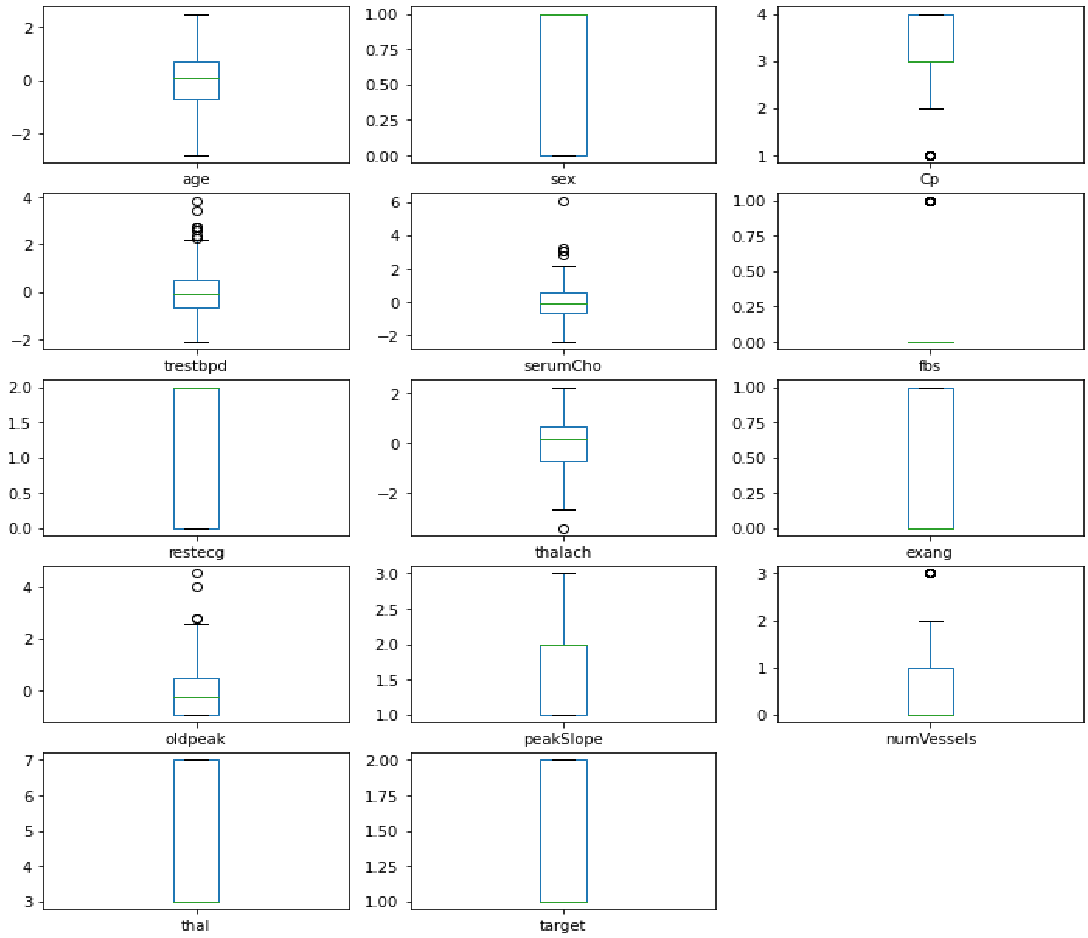

Fig. 2. Representation of Box and Whisker plotting of CVDs attributes.

C. Histogram

A histogram is a graphical portrayal that sorts out a gathering of main informative elements into indicated ranges [17]. Comparative in appearance to a reference chart, the histogram gathers an information series into an effectively deciphered visual by taking numerous elements and gathering them into sensible ranges or receptacles.

A histogram is a reference diagram like portrayal of information that cans a scope of results into segments along the *x*-pivot.

The *y*-pivot addresses the number includes or level of events in the information for every section and can be utilized to envision information disseminations. Fig. 2 address the histogram of the component appropriation of coronary illness dataset (See Fig. 3).

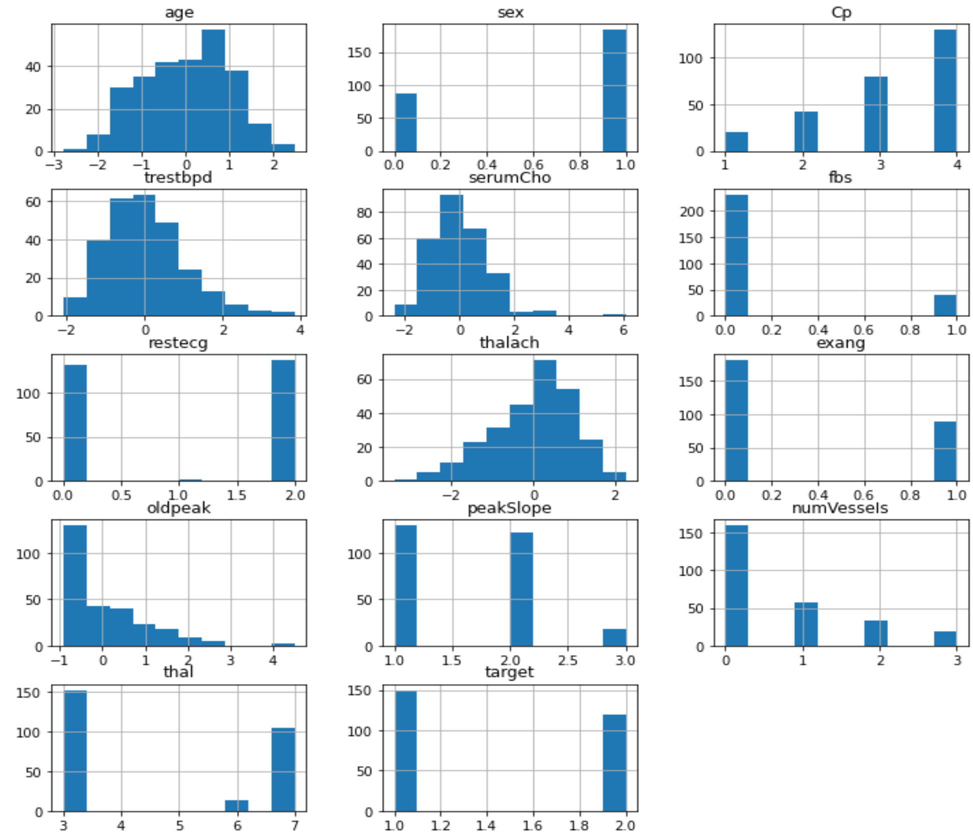

Fig. 3. Representation of Histogram plotting of CVDs attributes.

IV. Methodology

Pseudo-code of Proposed Cardiovascular Disease Diagnosis System.

1. Begin
2. The pre-processing of Cardiovascular Disease Dataset using pre-processing methods.
3. Feature selection using Pearson Correlation and Chi-Square methods.
4. Trains the classifiers (Logistic Regression, Random Forest, Naive Bayes', Support Vector Machine) using training dataset.
5. Validate using testing dataset.
6. Computes performance evaluation metrics.
7. End.

V. Feature Selection Techniques

A. Pearson Correlation Technique

Pearson Correlation defines correlation between sets of data is a measure of how well they are related. The values range between -1.0 and 1.0. A correlation of -1.0 represents a perfect negative correlation, while 1.0 represents a perfect positive correlation. Formula for Pearson's Correlation Coefficient is,

$$\rho_{(x,y)} = \frac{Cov(X,Y)}{\sigma X \sigma Y} \quad (1)$$

where, X is input feature,

Y is output feature,

σX is the standard deviation of X,

σY is the standard deviation of Y.

B. Chi-Square Technique

A chi-square technique is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.

The formula for Chi-Square is

$$X^2 = \sum \frac{(O-E)^2}{E} \quad (2)$$

where, O= Observed value(*s*)

E= Expected value(*s*)

VI. Data Mining Techniques

Four different data mining techniques are briefly described as follows:

A. Classification Using Random Forest

A Random Forest is a supervised machine learning technique that is used to solve classification and regression based problems. A random forest algorithm consists of a large number of decision trees in the forest. It is used for prediction which is based on an outcome of an average of the output taking from various trees.

Steps involved in Random Forest Algorithm:

1. In Random Forest, ‘m’ number of random samples are taken from the dataset having ‘*n*’ number of records.
2. Build the individual Decision Trees for each sample.
3. Each Decision Tree will produce an output.
4. Final output is examined based on majority voting for classification.

B. Classification Using Naive Bayes’ Method

Naive Bayes' is a probabilistic machine learning technique based on Bayes' theorem that is used for the classification task [20]. It assumes that attributes are independent of each other.

Bayes’ Theorem can be simulated as follows:

$$P(A/B)=[P(B/A)P(A)]/P(B) \quad (3)$$

With the help of Bayes' theorem, we are able to find the probability of *A* happening, given that *B* has occurred. Here, *B* is the evidence and *A* is the Hypothesis. That is the presence of one attribute does not affect the other. Hence, it is known as naive.

Steps involved in Naive Bayes’ Algorithm:

1) Firstly, transform the given dataset into frequency tables.

2) Create Likelihood table by obtaining the probabilities of given features.

3) Finally, use Bayes’ theorem to compute the posterior probability.

C. Classification Using Logistic Regression

Logistic regression is a statistical analysis technique used to detect an element based on prior observations of a set. It is used to classify data points of a set into two groups (binary Classification) by estimating the probability of each data value of the data set. Sigmoid function is used to map the predicted values to probabilities. This function has the ability to map any real value into another value within a range of 0 to 1 that shaped like the letter “*S*”.

$$S(x)=\frac{1}{1+e^{-x}} \quad (4)$$

D. Classification Using Support Vector Machines

SVMs are a set of supervised machine learning techniques that are used for classification, regression and outliers prediction. However, it is primarily used for classification problems.

The linear SVM finds the optimal hyperplane of the form $f(x)=w^T x+b$, where w is a dimensional coefficient vector and *b* is an offset. This is done by solving the subsequent optimization problem.

$$\min_{w,b}\frac{1}{2}w^2 \quad (5)$$

Subject to $\quad yi(wx+b)-1\geq 0, i=1\ldots..m$

VII. Experimental Result & Discussion

In this research work, we have used various feature selection methods namely Pearson correlation and Chi-Square method and applied to various machine learning algorithms for better detection of cardiovascular diseases (CVDs).

TABLE II: Performance of Classifier on Full Feature Set

Predictive Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	79.62	83	80	81
Random Forest	81.48	81	87	84
Logistic Regression	79.62	81	83	82
Naive Bayes’	79.62	83	80	81

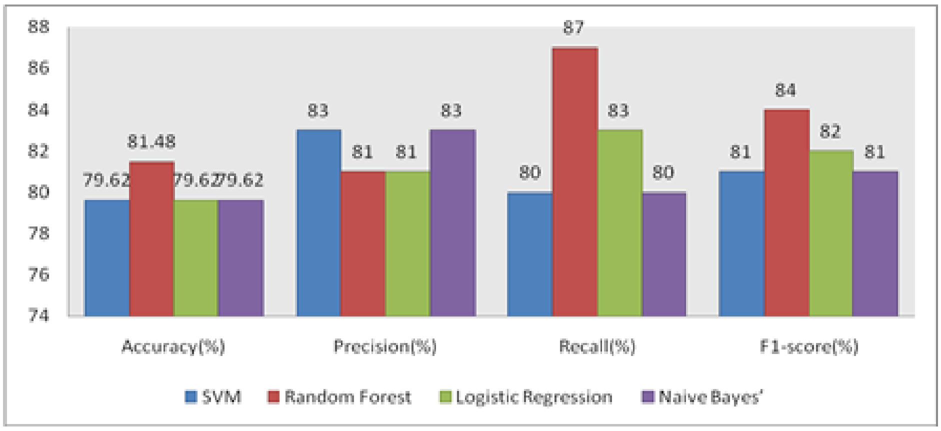

Fig. 4. Graphical representation of performance of classifiers on a full feature set.

The above Table II and Fig. 4, clearly show that the classifier random forest has good performance that obtained 81.48% accuracy,81% precision,87% recall and 84% F1-score as compared with the other classifiers on the full feature set.

TABLE III: Performance of Classifiers on Feature Selected by Chi-Square

Predictive Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	81.85	77	80	79
Random Forest	83.33	81	87	84
Logistic Regression	82.59	84	87	85
Naive Bayes’	84.07	77	77	77

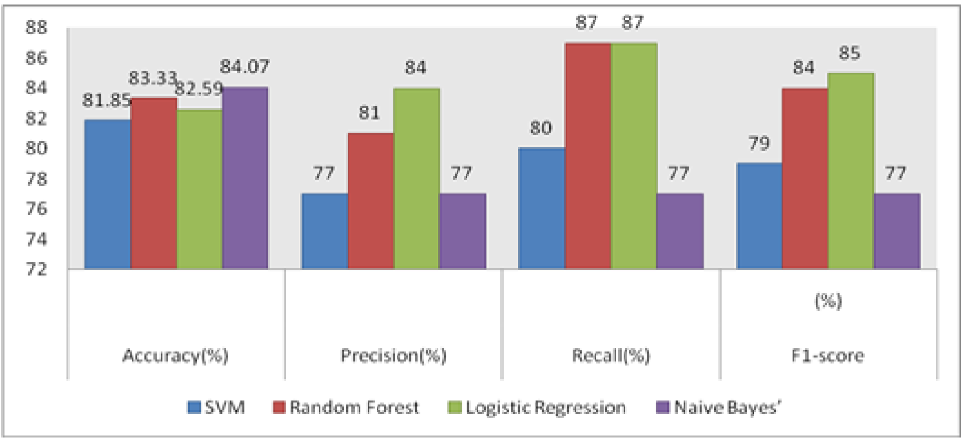

Fig. 5. Graphical representation of performance of classifiers on features selected by Chi-square.

TABLE IV: PERFORMANCE OF CLASSIFIERS ON FEATURES SELECTED BY PEARSON CORRELATION

Predictive Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	84.07	78	83	81
Random Forest	83.33	84	87	85
Logistic Regression	84.07	80	80	80
Naive Bayes'	83.33	68	63	66

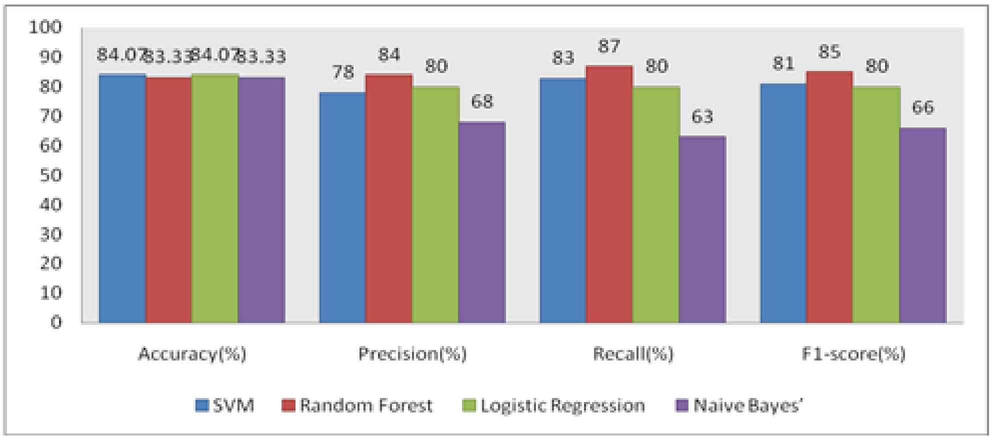

Fig. 6. Graphical representation of performance of classifiers on features selected by Pearson Correlation.

The above Table III, Table IV and Fig. 5, Fig. 6 clearly shows that the classifier logistic regression has good performance that obtained 84.07% accuracy, 80% precision, 80% recall and 80% F1-score as compared with the other classifiers on an optimal set of feature extracting from Chi-Square and Pearson Correlation algorithm.

Based on the above Table II, Table III and Table IV, we are able to find out the best classifier as Logistic Regression with an accuracy 84% as compared to other SVM, Random Forest and Naive Bayes'. The above tables show that the values for the accuracy, precision, recall and F1-score of the Logistic Regression are higher.

TABLE V: REPRESENTATION OF PREVIOUS YEAR PAPER ACCURACY SCORE

Authors	Year	Data mining Techniques	Accuracy
Sitar-Taut [16]	2009	Naive Bayes', Decision Tree	60.40%
Rajkumar and Reena [17]	2010	Naive Bayes', Decision Tree, KNN	52.33%
M. Gudadhe [13]	2010	Multi-layer Perceptron, SVM	80.41%
H. Benjamin [3]	2018	Random Forest, Decision Tree, Naive Bayes'	81%

We have found about 15 years old previous paper [V] and find maximum accuracy near about 81%. In studies this research work, we have tried to test with different features selection method applied on different machine learning algorithms. Finally, we have found that Logistic Regression provide the best result 84% accuracy.

VIII. CONCLUSION

The overall objective of this research paper is to predict whether a person suffering from Cardiovascular Disease or not by using data mining techniques. In this work, Pearson Correlation and Chi-Square feature selection methods are used for extracting the best features that are applied on various data mining techniques. Also, the UCI data repository is used for performing the comparative analysis of four algorithms namely Random Forest, Logistic Regression, Naive Bayes' and SVM. From the research works, it has been experimentally proven that Logistic Regression provides the best results as compared to others.

The future work of this research paper is to improve the accuracy of the classification algorithm by applying other feature selection methods in order to find the optimal subset of the feature that is significant for the detection of Cardiovascular Diseases (CVDs).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Priya Singh and Sanjeev Gangwar conducted the research work, developed the methodology of work and wrote the paper. Sanjeev Gangwar and Gyanendra Kumar Pal supervised and provided guidance in the preparation of manuscript and edited the manuscript. All authors agree with the final version of the manuscript.

REFERENCES

[1] J. P. Li, Amin, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol.8, pp. 107562-107580, 2020.

[2] WHO. (2016). Health topics – Cardiovascular disease. [Online]. Available: https://www.who.int/india/health-topics/cardiovascular-diseases

[3] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, "Heart disease prediction using data mining techniques," in *Proc. 2017 International Conference on Intelligent Computing and Control (I2C2)*, 2018.

[4] B. Bahrami and M. H, Shirvani, "Prediction and diagnosis of heart disease by data mining techniques," *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, vol. 2, issue 2, pp. 164-168, 2015.

[5] V. Chaurasia and P. Saurabh, "Early prediction of heart diseases using data mining techniques," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208-217, 2013.

[6] S. B. Patil and Y. S. Kumaraswam, "Intelligent and effective heart attack prediction system using data mining and artificial neural network," *European Journal of scientific Research*, vol. 31, no. 4, pp. 642-656, 2009.

[7] M. Hanumathappa and K. Uma, "Heart disease prediction using classification techniques with feature selection method," *Adarsh Journal of Information Technology*, vol. 5, issue 2, pp. 22-29, 2016.

[8] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *International Journal of Computer Science and Network Security*, vol. 8, no.8, pp. 343-350, 2008.

[9] N. Deepika, "Association rules for classification of heart attack patients," *IJAEST*, vol. 11, no. 2, pp. 253-257, 2015.

[10] S. M. Alzahani, "An overview of data mining techniques applied for heart disease diagnosis and prediction," *Lecture Notes on Information Theory*, vol. 2, no. 4, pp. 310-315, 2014.

[11] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview," *AI Magazine*, vol. 13, no. 3, pp. 57-70, 1992.

[12] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and fuzzy AHP for heart failure risk prediction system," *Expert Syst. Appl.*, vol. 68, pp. 163–172, 2017.

[13] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in *Proc. 2010 International Conference on Computer and Communication Technology (ICCCT)*, 2010, pp. 741-745.

[14] V. Chaurasia and P. Saurabh, "Data mining approach to detect heart diseases," *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, vol. 2, pp. 56-66, 2014.

[15] Statlog (Heart) Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/statlog+(heart)

[16] V. A. Sitar-Taut, D. Zdrenghea, D. Pop, and D. A. Sitar-Taut, "Using machine learning algorithms in cardiovascular disease risk evaluation," *Journal of Applied Computer Science & Mathematics*, vol. 5, pp. 29-32, 2009.

[17] A. Rajkumar and G. S. Reena, "Diagonsis of heart disease using data mining algorithm," *Global Journal of Computer Science and Technology*, vol. 10, pp. 38-43, 2010.

[18] A. Edwards, R. Thomas, R. Williams, A. L. Ellner, P. Brown, and G. Elwyn, "Presenting risk information to people with diabetes: evaluating effects and preferences for different formats by a web-based randomised controlled trial," *Patient Education and Counseling*, vol. 63, no. 3, pp. 336-349, 2006.

Priya Singh is currently pursuing M. Tech. in computer science & engineering at Uma Nath Singh Institute of Engineering & Technology, VBS Purvanchal University, Jaunpur. She has done B.Tech. in computer science & engineering at Uma Nath Singh Institute of Engineering & Technology, VBS Purvanchal University, Jaunpur. She has also done MBA in finance & marketing from VBS Purvanchal University.

Her technical skills include C, C#, SQL and python. Her areas of interests are data mining, machine learning, operating system and database management system.

Gyanendra Kumar Pal is currently working as an assistant professor at the Department of Information Technology, Uma Nath Singh Institute of Engineering & Technology, VBS Purvanchal University, Jaunpur. His main research areas are software engineering, data mining, cryptography and cyber law. He has teaching experience of more than ten years at the undergraduate and postgraduate program in institutions.

Sanjeev Gangwar is currently an assistant professor at the Department of Computer Applications, Uma Nath Singh Institute of Engineering & Technology, VBS Purvanchal University Jaunpur (U. P.). He has teaching experience of more than fifteen years at the undergraduate and postgraduate program in various institutions. His areas of interests are mobile ad hoc networks, wireless sensor network, computer network, operating system and data mining.