International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Non-Invasive Detection of Coronary Heart Disease Using Machine Learning

Divya Sunil Nerkar

Department of Computer Science Engineering SSVPS Bapusaheb Shivajirao Deore College of Engineering Dhule

ABSTRACT

Cardiovascular illness can be accurately diagnosed by the use of coronary arteriography (CAG) coronary heart disease can be detected with an invasive method; however, this is not suited for the yearly physical examination. The heart disease diagnosis through conventional medical history particularly has not been considered reliable in many views. In classifying the fit people having heart disease, non-invasive methods like cloud-based data using machine learning techniques are reliable and efficient. These techniques for the prediction can aid the medical field. Apart from dietary control and a healthy lifestyle, the right time for diagnosing is also very important. Age, cholesterol, sex, high blood pressure, smoking, obesity, family history, physical inactivity, poor diet, diabetes, alcohol intake, and hereditary, are risk causes for heart disease. There are different categories of heart diseases such as coronary heart disease, angina pectoris, congestive heart failure, cardiomyopathy, congenital heart disease, arrhythmias, and myocarditis. In this study, we used the UCI (University of California, Irvine) machine learning repository evaluated on the Cleveland heart disease dataset. We applied various feature selection methodologies to the pre-processed dataset and by applying the machine learning classification algorithms we will be predicting coronary heart disease at an early stage which will help the patient to get treatment on time.

Keywords: ***Coronary heart disease, Annona Test, machine learning, feature selecti***

1. INTRODUCTION

The most vital and essential part (organ) of the human body is the Heart. Many diseases are linked to the heart so the analysis of prediction of the heart must be accurate. To resolve this, a virtual study in this field is obligatory. Normally these diseases are predicted at the end stage and this is the main reason for the death of heart patients due to deficiency of correctness because of this there is a requisite to identify proficient algorithms for disease prediction [1]. One of the proficient capable and effective technologies is Machine Learning which is established specifically for training and testing with the support of python and python libraries. Method acquires training directly from data and skill, based on this training, testing should be done on various types of needs as per requisite algorithms. For Testing and Training, Machine learning can be used as an effective technique. It belongs to AI (Artificial Intelligence). AI has one of its branches as machine learning. The work which is done by humans using human intelligence that work can be done using machine learning technology. To enable human intelligence features in ML, The ML technology is equipped with a different process to make use of data. As ML describes, it absorbs ordinary phenomena. By using python libraries with the algorithms of machine learning, this prediction has to be done [2]. In this detection, elements of biological are used such as chest pain (cp), sex, blood pressure(bp), and cholesterol (chol). By using these elements six algorithms of ML such as SVM, Decision Tree, Random Forest, Naïve Bayes, and Logistic Regression are applied for the prediction of analysis and conclude which technique is best based on the confusion matrix [3]. Cardiovascular disease has long been regarded as the most dangerous and lethal of all human diseases. Cardiovascular diseases and their high death rates are putting the health care systems of the globe in jeopardy. When it comes to cardiovascular disease, men are more prone than women to suffer from it in middle and old age. Though children with identical health conditions exist, [4] [5] and [6] are examples of this.According to the World Health Organization, heart disease is to blame for one-third of all fatalities globally (WHO). More than a third of all fatalities around the globe would be caused XA by cardiovascular disease (CVD) by 2022, according to the World Health Organization. 85 per cent of these persons died from heart attacks or strokes. [7] [8]

. According to the European Society of Cardiology (ESC), cardiovascular disease affects more than 26 million people worldwide, with an additional 3.6 million cases identified each year. 50% of individuals diagnosed with the cardiovascular disease die within two years, and 3% of all health care spending is devoted to treating it. [9]. To effectively predict heart disease, you'll need a slew of different tests. The incompetence of the medical team might lead to inaccurate predictions. [10]. It might be difficult to get a diagnosis in the early stages. [11]. surgically treating cardiac disease can be difficult, especially in less developed countries where medical professionals and diagnostic equipment are lacking. [12]. In certain cases, cardiac failure can be prevented by precisely diagnosing patients' risk of heart failure. [13]. Data-trained algorithms can help detect diseases. [14]. Heart disease datasets are available online to the general public to compare prediction methods. Machine learning and artificial intelligence may be used to develop the best prediction model possible using data from massive datasets.In this work, we analyzed the CVD disease dataset and applied various classification algorithms to predict the CVD more accurate. We also applied the feature selection methods to reduce the number of features which further gives the benefits of analyzing CVD faster. The remainder of this paper is organized as follows. In section II, methodologies have been mentioned and the preprocessing methods of the data are introduced. In section III, the Experimental results are presented, followed by the conclusion in Section IV

2. METHODOLOGHY

In this research, we will be proposing a machine learning-based decision support system for the prediction of heart disease. This proposed system consists of three stages: data collection, data pre-processing, and model construction. In the pre-processing stage, feature selection is done, and class balancing is done. In the dataset, 561 instances are belonging to class 0, and 629 instances belong to class 1. Here no missing values are in the dataset & all features are in int, float data type. After that split dataset into train & test a ratio of 70:30 and apply all classification models on the training dataset & calculate the accuracy of all models. After that feature selection technique which is the ANOVA test applies to the dataset & finds out the top 7 datasets which are highly co-related. After getting the top 7 features, fit the data on classification models & calculate accuracy.A classification model is to be built using these features with the help of classification algorithms using Python Machine Learning technologies

3. SYSTEM ARCHITECTURE

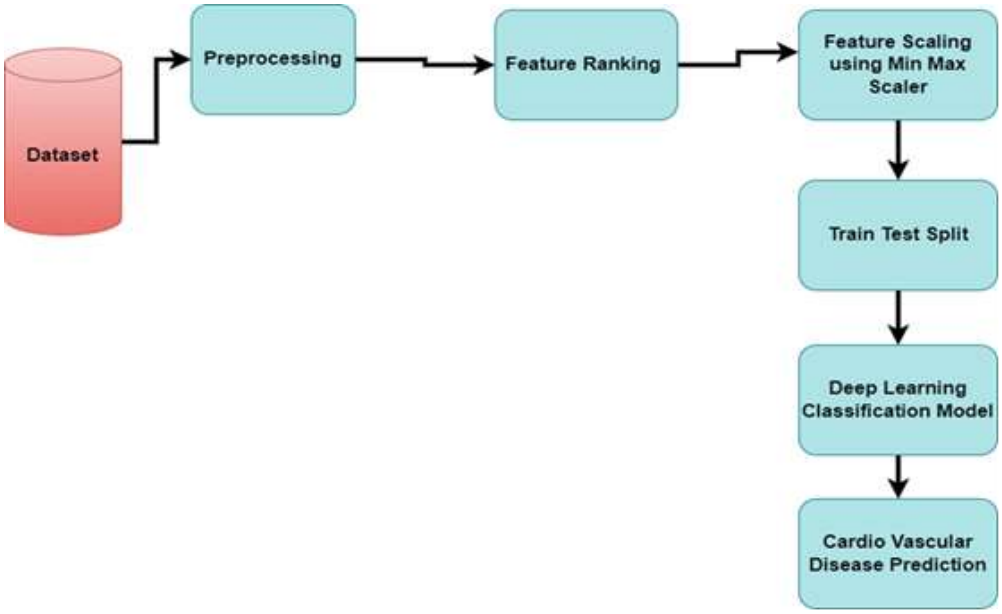

Figure 1: System Architecture

Data Collection:

The studies were carried out using data from the UCI library on Cleveland Heart Disease (CHD). Categorical characteristics comprise eight of the dataset's 14 features, while numeric features account for the other six. There are some examples of the dataset's properties and descriptions in Table 1.[1].

Table 1: Cleveland Heart Diseases Dataset Features[1]

Feature name	Feature Code	Description
Age	AG	Age between 29 and 77
Sex	SX	Male: 1, female: 0
Type of chest pain	CP	Typical angina: 1, atypical angina: 2 non-angina pain: 3, asymptomatic: 4
Resting blood pressure	RBP	Between 94 mm Hg and 200 mm Hg
Cholesterol	SCHOL	Between 126 mg/dl and 564 mg/dl
Fasting blood sugar	FABS	FBSR > 120 mg/dl (true:1, false: 0)
Resting electrocardiographic results	REAR	Normal: 0, ST-T wave abnormality: 1, Hypertrophy: 2)
Maximum heart rate achieved	HR	Between 71 and 202
Exercise-induced angina	EIAG	YES:1, NO:0

ST depression induced by exercise relative to rest	STD	Up sloping: 1, Flat: 2, downsloping: 3
Target	TARG	Heart disease present: 1, heart disease absent: 0

This dataset includes information on patients ranging in age from 29 to 77. Heart disease might manifest as chest discomfort. Typical angina, atypical angina, non-angina pain, and asymptotic are all forms of chest pain. Feature resting blood pressure (RBP) is the patient's resting blood pressure. The patient's cholesterol level is shown by SCHOL. FABS is used to calculate fasting blood sugar levels. Sugar levels over 120 mg/dl are recorded as 1, otherwise as 0. MHR stores the patient's maximal heart rate, which is obtained via RECR. If you have exercise-induced angina, your EIGA score will be 1, else it will be 0. Upslope, downslope, and flat are all potential values of the STD for the depression generated by exercise. The slope of maximum exertion (SPE) The number of main arteries and veins that can be seen during fluoroscopy is recorded in the NMVCF. A person's TARG characteristic lets you know whether they have a cardiac condition. As seen in Figure 1, the proposed hybrid method for the prediction of cardiovascular disease is either disease or not. Five potential values 0 for no heart disease, and 1 to 4 for various stages of illness are available in this feature. To detect illness, levels 1–4 are combined into a single indicator.

Experiment Analysis & Results:

After pre-processing of data, we apply train_test_split to the dataset. In that, we kept 70% data for training and 30% data for testing purposes. And then we apply machine learning classification models with hyperparameters. While applying the model, parameters set as the criterion is 'entropy' and set max depth. so these give the best performance as compared to other criteria & max depth value. After that, we got the accuracy of all models & compared them by graphs.

The above graph(fig.2) shows the accuracy of machine learning models with the best parameters and here Random Forest Classifier model gives 92.43% accuracy with all Independent features.

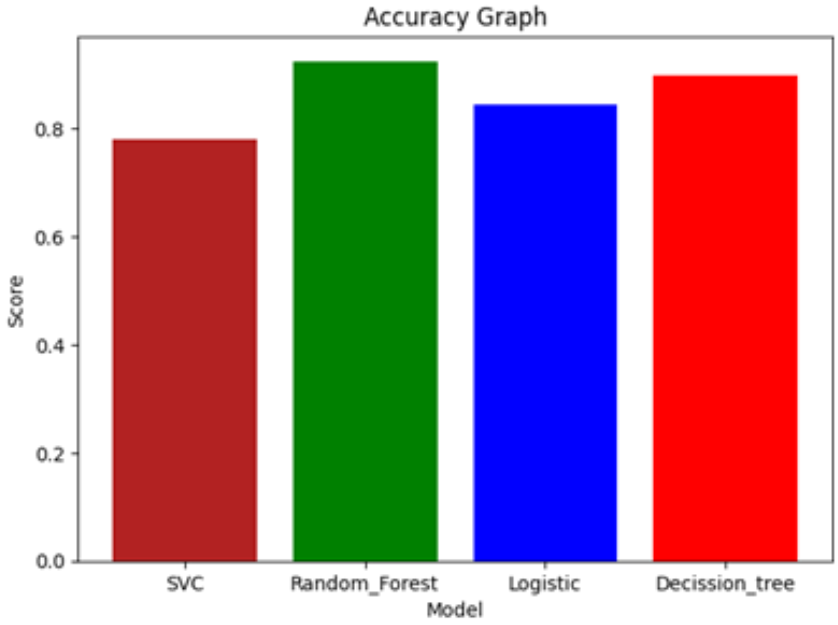

Figure 1:Accuracy Graph

After that, we drop some features that are less co-relate with the dependent feature. Here we used the ANOVA test for feature selection. From 11 features, we select the top 7 features which are highly co-relate to dependent features.

Now we have got the top 7 features which are highly correlated & again apply train_test_split for these top features. After splitting data, fit training data to the above machine learning models & get the accuracy of each model.

The below graph(fig.3) shows the accuracy of machine learning models with parameters and here Decision tree classifier model also gives 92% accuracy with 7 Independent features. It requires less execution time as compared to 11 features and gives nearer accuracy.

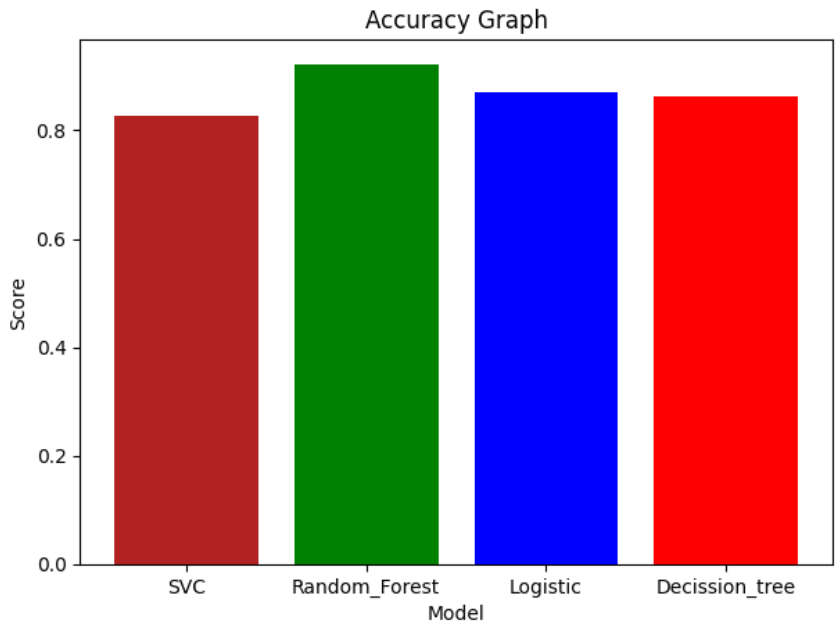

Figure 1Accuracy graph for ML model with feature selection technique

We also apply the cross-validation method with and without the feature selection technique. Here we use k=15 folds because it gives the best accuracy as compared to k=5 & k=10 folds. And we got less accuracy for both cross-validations with & without feature selection technique & it requires more execution time of code due to 15 folds validation.

The following results are the classification report & performance matrix score of one of the best performing model Random Forest Classifiers with top 7 features using hyperparameters.

Accuracy Score is :- 0.9215686274509803
Classification Report is :-

	precision	recall	f1-score	support
0	0.93	0.90	0.91	165
1	0.91	0.94	0.93	192
accuracy			0.92	357
macro avg	0.92	0.92	0.92	357
weighted avg	0.92	0.92	0.92	357

Figure 2 Classification report for Random forest model with feature selection

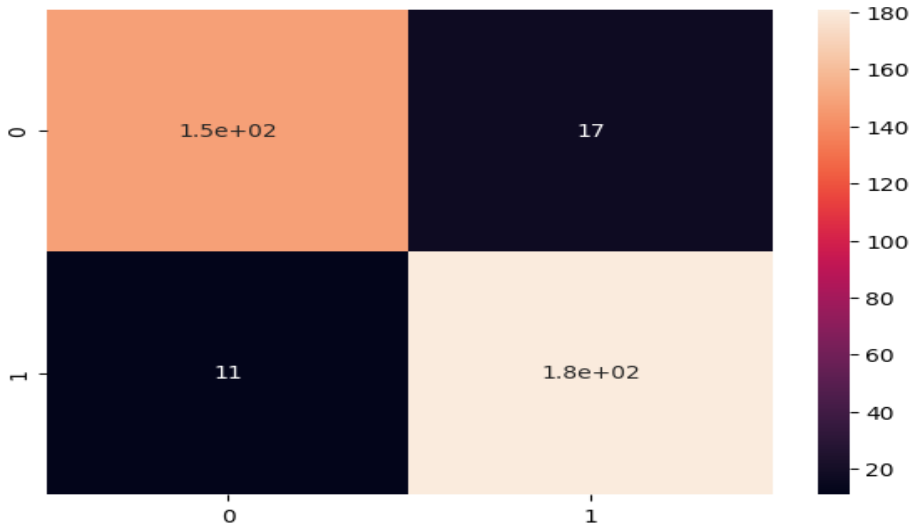

Figure 3 Confusion matrix for Random Forest Model

Table 2 shows the accuracy of all performing machine learning models with Hyperparameters and Cross-validation. Here also shows the time of execution of the code of each experiment.

Table 2: Performance sheet for all Machine Learning Models

Accuracy(%)					
Sr.No.	**Machine Learning Models**	**Hyper Tunning with all features**	**Feature Selection - 7Features(ANOVA Test) For Hyperparameters**	**Cross Validation(15) with all features**	**Feature Selection -7 features(ANOVA Test) After Cross Validation**
1	**SVC**	78.15	83	69.14	75.5
2	**Random Forest Classifier**	92.43	92	89.31	86.54
3	**Logistic Regression**	84.59	87	81.02	81.01
4	**Decision Tree Classifier**	89.91	86	85.58	82.27
	Execution Time(seconds)	5.31	2.4	8.4	6.37

5. CONCLUSION

The heart acts a major role in the corporeal organism. The disease of the heart wants more perfection and exactness for diagnosis and analysis. In real-time heart diseases may not be detected in the early stage. This needs further analysis. In the proposed work, an accurate and early heart disease prediction is presented by using data set of heart diseases. The presented methodology requires various ML algorithms. The analysis is carried out based on the Confusion matrix and comparing accuracy among them and get Random forest is the finest algorithm. Thus the efficacy of the presented work has been verified. This technique may be used as a support for the early and accurate prediction of heart disease. There are many more ML algorithms that can be used for the finest exploration and earlier prediction of heart diseases for the upcoming possibility. This needs further diagnosis

6. References

[1]. E.Taylor,P.s.Ezekiel, F.B.Deedam. (2019). “A Model to Detect Heart Disease using Machine Learning algorithm” International Journal of Computer Science and engineering.vol-7,issue-11

[2]. R. Goel and A. Jain. (2018) "The Implementation of Image Enhancement Techniques on Color n Gray Scale IMAGEs," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 204-209, doi: 10.1109/PDGC.2018.8745782

[3]. Archana Singh, Rakesh k. (2020). ”Heart disease Prediction Using Machine Learning Algorithms” International Conferences On Electrical and Electronics Engineering(ICE3)

[4]. P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, “A decision support system for heart disease prediction based upon machine learning,” J. Reliab. Intell. Environ., vol. 7, no. 3, pp. 263–275, 2021, doi: 10.1007/s40860-021-00133-6.

[5]. F. Ali et al., “A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion,” Inf. Fusion, vol. 63, no. April, pp. 208–222, 2020, doi: 10.1016/j.inffus.2020.06.008.

[6]. G. Saranya and A. Pravin, “Hybrid Global Sensitivity Analysis Based Optimal Attribute Selection Using Classification Techniques by Machine Learning Algorithm,” Wirel. Pers. Commun., no. 0123456789, 2021, doi: 10.1007/s11277-021-08796-3.

[7]. G. Magesh and P. Swarnalatha, “Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction,” Evol. Intell., vol. 14, no. 2, pp. 583–593, 2021, doi: 10.1007/s12065-019-00336-0.

[8]. A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and W. A. Wan Ahmad, “A novel approach for heart disease prediction using strength scores with significant predictors,” BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–16, 2021, doi: 10.1186/s12911-021-01527-5.

[9]. “UCI Machine Learning Repository: Heart Disease Data Set.” https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed Dec. 26, 2021).

[10]. P. Rani, R. Kumar, A. Jain, and R. Lamba, “Taxonomy of Machine Learning Algorithms and Its Applications,” J. Comput. Theor. Nanosci., vol. 17, no. 6, pp. 2508–2513, Sep. 2020, doi: 10.1166/JCTN.2020.8922.

[11]. U. N. Dulhare, “Prediction system for heart disease using Naive Bayes and particle swarm optimization,” undefined, vol. 29, no. 12, pp. 2646–2649, 2018, doi: 10.4066/BIOMEDICALRESEARCH.29-18-620.

[12]. M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, “Prediction of heart disease using random forest and feature subset selection,” Adv. Intell. Syst. Comput., vol. 424, pp. 187–196, 2016, doi: 10.1007/978-3-319-28031-8_16.

[13]. “Machine Learning Basics with the K-Nearest Neighbors Algorithm | by Onel Harrison | Towards Data Science.”

https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (accessed Jan. 13, 2022).

[14]. "XGBoost Algorithm | XGBoost In Machine Learning." https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ (accessed Jan. 13, 2022).