Cardiac Disease Prediction using Supervised Machine Learning Techniques.

Chiradeep Gupta1, Athina Saha2, N V Subba Reddy3, U Dinesh Acharya1
1MTech, Computer Science and Information Security.
1,2,3Department of Computer Science, Manipal Institute of Technology, MAHE, Manipal.
4chiradeep.gupta@learner.manipal.edu 5Athina.saha@learner.manipal.edu
2nvs.reddy@manipal.edu, 3dinesh.acharya@learner.manipal.edu

Abstract. Diagnosis of cardiac disease requires being more accurate, precise, and reliable. The number of death cases due to cardiac attacks is increasing exponentially day by day. Thus, practical approaches for earlier diagnosis of cardiac or heart disease are done to achieve prompt management of the disease. Various supervised machine-learning techniques like K-Nearest Neighbour, Decision Tree, Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM) model are used for predicting cardiac disease using a dataset that was collected from the repository of the University of California, Irvine (UCI). The results depict that Logistic Regression was better than all other supervised classifiers in terms of the performance metrics. The model is also less risky since the number of false negatives is low as compared to other models as per the confusion matrix of all the models. In addition, ensemble techniques can be approached for the accuracy improvement of the classifier. Jupyter notebook is the best tool, for the implementation of Python Programming having many types of libraries, header files, for accurate and precise work.

1. Introduction
The essential pumping organ of the human body requires proper care. Unforeseen development may inculte confinement of random disease [5]. With the help of proper datasets from the validated repository, and utilization of supervised techniques would facilitate monitoring otherwise prediction to cease any unfortunate occurrence. One such technique possibly market-favoured is the machine learning techniques[1]. Machine learning is based on testing and training the data where the system takes data from experience, and once the model has been trained, the predictions are made on the test dataset[2]. Supervised learning like Logistic regression, Naïve Bayes, Decision Tree, SVM Model, K-NN and Random can be defined as learning in the presence of a teacher [3]. Machine-learning techniques are useful for vast data to be incorporated into the development of robust predictive analytics, often without restrictions of standard modelling techniques [4].

For the proposed work in this paper, the dataset is split in the ratio of 7:3(train: test) where the ‘train’ is the teacher for prediction on the ‘test’ [4]. In recent times, those techniques aid in making medical aid software for the early diagnosis. The risk of a fatality can be reduced at the primary stage by identifying any heart-related illness [11]. For the implemented project as discussed further in the paper, the biological parameters used are – age, cholesterol and blood pressure, chest pain type, sex, exercise-induced angina, and others. On their basis, the comparison has been performed in the terms of accuracy and other performance metrics of the algorithms used.

2. Related Works
The proposed work would be considering the related works discussed in [2], [3], [5], [6], [7]. The techniques that which be used for approaching a mechanism to detect-otherwise-predict the cardiac disease can be the machine learning techniques [1]. The machine learning techniques feasible to be approached to achieve the proposed work are namely the supervised classifiers: K-Nearest Neighbour, Decision Tree, SVM Model and Logistic Regression, using datasets from the UCI repository [2].
Zhenya and Zhang [2] proposed a cost-sensitive ensemble method to improve the efficiency of diagnosis, which reduces the misclassification cost. Five varied classifiers: random forest, logistic regression, SVM Model, extreme learning machine, and K-Nearest Neighbour have been used and T-test investigated the ensemble performance. According to ten-fold cross-validation, the method proposed did achieve the finest performance. C. Beulah et al. [3] had examined the accuracy of predicting cardiac disease with help of ensemble of classifiers and majority voting. Bagging improved the accuracy by 6.92% whereas 5.94% by boosting, 7.26% when weak classifiers were assembled with majority voting, and 6.93% when stacking was used. Archana Singh et al.[5] calculated the accuracies of the various ML techniques to classify whether a person has heart disease such as K-nearest neighbour, decision tree, linear regression, and SVM Model on the dataset from the UCI repository. Based on this dataset, K-Nearest Neighbour was found to be the best one when the algorithms were analysed.

Kohali et al. [6] worked on prediction of heart disease using logistic regression, diabetes prediction using SVM Model, and forecasting of breast cancer using the Adaboot classifier. They deduced that the logistic regression produced 87.1% accuracy, SVM 85.71% accuracy, and 98.57% accuracy for the Adaboot classifier. Thus, these models are good for predicting diseases. Cardiac Disease severity can be labelled based on various techniques such as the K-nearest neighbour method, decision tree, generic algorithm as well as Naïve Bayes. Mohan et al. [7] has combined two different approaches to a single unit approach, which is termed as hybrid approach with an accuracy of 88.4%. Himanshu et al. [8] made a proposal that the performance of the Naïve Bayes with respect to low variance and high biasness is better than the K-Nearest Neighbour, which also suffers from the overfitting problem. Use of low variance and high biasness for small datasets benefits by taking less time for training and testing the model. An increase in the dataset gives asymptotic errors and low biasness. The Decision Tree also suffers from the problem of overfitting but can be solved by some overfitting removable techniques.

3. Research Methodology
Cardiac Diseases are the disease referring or relating to the pumping organ of a human body – the heart. A well-known medical fact is that such disease describes a range of conditions that could affect the heart.

| Sl.No. | Attribute | Description | dType |
|-------|-----------|-------------|-------|
| 1     | age       | Age of the patient: 29 to 77 years | int64 |
| 2     | sex       | Male : 0, Female : 1 | int64 |
| 3     | cp        | Chest Pain Type | int64 |
| 4     | texthsps  | Resting Blood Pressure(in mmHg) Range: 94-200 | int64 |
| 5     | chol      | Serum Cholesterol(in mg/del) Range: 126-564 | int64 |
| 6     | fbs       | Fasting Blood Sugar Range: < and > 120mg/dl True : 1, False : 0 | int64 |
| 7     | resteog   | Resting electrocardiographic result | int64 |
| 8     | thalach   | Maximum Heart Rate 71 to 202 | int64 |
| 9     | exang     | Exercise Included-Angina Yes: 1, No: 0 | int64 |
| 10    | oldpeak   | ST depression due exercise w.r.t. rest: 0 to 0.2 | float64 |
| 11    | slope     | The slope of the peak exercise ST segment: 0 to 1 | int64 |
| 12    | ca        | Number of major vessels: 0 to 3 | int64 |
| 13    | thal      | Normal; value: 3 | int64 |
| 14    | target    | Cardiac Disease Yes: 1, No: 0 | int64 |
3.1 Data Collection
Collection of the data from the UCI repository dataset [17] is the initial step for the commencement of the research work. Numerous researchers and the UCI authority verify this dataset.

3.2 Attribute Selection
Attributes are properties of the dataset, which are used for determining whether the person has the disease. Heart rate, chest pain types, sex, age, and exercised induced angina of the person and many more features are shown in table 1.

3.3 Data Exploration
This step is used for visually exploring the dataset to understand the various biological parameters used in the dataset. This dataset has the record of 303 patients, out of which 165 patients have heart disease. The ratio of records of female and male patients used in the dataset is 32:69. Similarly, there are records of patients with Types 0, 1 and 2-chest pain in the ratio of 143:50:87:23. The patients with blood sugar for the range up to 120 mg/dl are in the ratio of 15:86 and the patients with Exercised Induced Angina and without it are in the ratio 3:68 as depicted in figure – 1, 2. In figure 3, 54 is the age in the considered dataset having the highest frequency of heart disease among 29 – 77 age group while figure 4 plots a scattered graph for visualising maximum heart rate versus age, cholesterol and resting blood pressure (rbp) respectively.

![Figure 1. Pie-Chart Visualization of the dataset.](image1)

![Figure 2. Bar graph visualization of Dataset.](image2)
3.4 Data Pre-processing

Pre-processing of the dataset would eliminate the missing values or any duplicates in it to make the predictions of the various machine-learning algorithms more accurate. Standard Scaler would serve as the best technique that has been used in the proposed work to standardize the data into common scales without which the result could have been lower than the one achieved as discussed further. The objective of using it came as a conclusion of checking and finding no null – values in the dataset.

4. Supervised Learning Classifiers

4.1 Logistic regression

Logistic Regression is a classifier that predicts binary outcomes for a given set of non – dependent variables. Thus, the nature of the target or the dependent variable is dichotomous [16]. The assumptions, which should hold for logistic regression, are –

- The target variable should be categorical in nature.
- Little or no multi–collinearity among the independent variables.
- The link function and the independent variable should have a linear relation in the model.

4.2 Naïve Bayes

Naïve Bayes was earlier used for text classification. It can deal with datasets of large dimensions. Prerequisite is the Bayes Theorem with the assumption – independent nature of the attributes among the same class [12]. It is referred to as a distribution methodology that includes an expectation of predictors and independence [10].

4.3 Decision Tree Classifier

It is a tree structure for deducing decisions at the nodes and reach to some outcome at the leaf nodes. They are used for classification where each path is a set of decisions leading to one class. Root to the leaf node represents a set of rules, which leads to a classification [10]. It is a flowchart: branching node is a test on a feature; branch indicates the inferences made from the tests and the leaf node providing information about a single class label. It is of two types: classifier tree (categorical decision) and regression tree (continuous outcome) [12]. In Figure 5 dataset is visualized using Decision Tree, with maximum depth as 4 and minimum sample leaf as 7.

Figure 5. Decision Tree Visualization.
4.4 SVM Model
SVM abbreviates for Support Vector Machine model. It is formally cited by the split of the decision boundary to separate the classes of data points on either side of the boundary[10]. This classifier is –

- Point in the one-dimensional space,
- Line in two-dimensional space,
- Plane in the three-dimensional space, and
- Hyperplane in the four-dimensional space[12]

4.5 K – Nearest Neighbour
It is the convenient, easy to control and supervised ML classifier availed by following the steps below [10].

i. Stack of data
ii. Assign ‘k’, the value of the number of picked neighbours.
iii. Calculating the distance between current and queried instances.
iv. Instance and distance index count, together with an organized collection.
v. Calculation of Euclidean distance \((X,Y) = \sqrt{\sum (X_i - Y_i)^2}\)
vi. Tags for the initial ‘k’ number of entries.
vii. Return median (k labels) on relapse; return mode (k labels) on classification [10].

This is to be noted that k-values should never be considered as 1. In the implementation part of the paper, the considered k-value is 7 after analysing the different lower k values and the corresponding obtained accuracies. As an extension to this, plotting a graph for k-values in the range (2 to 20) yielded a better accuracy result for the k-value of 14, as shown in figure 6. Accordingly, the other parameters would yield better performance for this ‘k’ value in comparison to the earlier one, as discussed in further sections.

![Figure 6. A Graph to determine ‘k’ value for better accuracy/test score.](image)

4.6 Random Forest
Random Forest classifier learning model is a supervised classifier used for classification as well as regression [12]. As the name entails, every single decision tree makes one class-prediction and the majority votes’ class develops into the models’ forecast. Steps followed in this algorithm –

i. Choose faces for all aspects, where the number of faces = K and features = M such that, \(K < M\).
ii. Among those attributes, ‘D’ node is measured.
iii. Now, divide that node, obtained in the previous step, into child nodes thus giving the prime split.
iv. Construct the forest using replication of steps for n number of times for ‘n’ decision trees [10].

In the implementation part of the paper, the considered value for the number of trees (n-estimator) inside the classifier is 40 after analysing the different lower values of the same and the corresponding obtained accuracies.
5. Results and Discussion

Dataset is split into 7:3 ratio of ‘train’ and ‘test’ dataset, respectively. Fitting the model in the training dataset is followed by evaluating it in the test dataset, with the objective of estimating the performance on the new data rather than the already trained data.

Followed by this, confusion Matrix, and the various performance metrics of the models such as accuracy, recall, specificity, F1-Score, and precision are used to compare, contrast, and determine the best and accurate model for predicting cardiac disease, amongst the six different supervised machine-learning classifiers. Further, the obtained results are compared with that of the existing works.

5.1 Confusion Matrix

Confusion Matrix is the most effective tool to analyse cardiac disease prediction in this field of study. It is deployed to perceive the behaviour of the different classifiers.

Such a matrix provides the information about the way the classifier has performed of matching the correctly predicted examples corresponding to the incorrectly predicted examples. Therefore, a confusion matrix is the tabular representation of the model estimated and the actual values of the dataset. It is used by machine learning classification-based problems for measuring the performance of the models. It consists of four separate commixtures of the predicted – actual values namely True Negative, True Positive and False Positive and False Negative as illustrated clearly in figure 7. These four elements are the building blocks that help to design the confusion matrix.

| Predicted Values | Actual Values |
|------------------|--------------|
| 0                | 0            | True Negative (TN) | False Positive (FP) |
| 0                | 1            | False Negative (FN) | True Positive (TP) |

Figure 7. Confusion Matrix.

True Positive: Model does the correct prediction of the positive class (patient has the disease) whereas True Negative: correct prediction of negative class (patient does not have the disease). False Positive: Incorrect prediction of positive class whereas incorrect prediction of the negative class is coined as False Negative. In medical testing, too many false negatives are very risky, since reports will not show that the person is having heart disease when the person is having heart disease.

The above-cited technical zone of the project can be tabulated, where the confusion matrix of the implemented Machine–Learning Algorithms can be compared as in table 2. The implemented classifiers have done in total 91 predictions, i.e., 91 records.
Table 2. Confusion Matrix Comparison of Implemented Classifiers.

| Confusion Matrix | Description |
|------------------|-------------|
| **Logistic Regression** | 91 records: 54 times “Yes” and 37 times “No” Prediction |
| | True Negative – The model has predicted 35 times that the patient is not having heart disease accurately. |
| | True Positive – The model has predicted 49 times that the patient is having the disease accurately. |
| | False Negative – Twice The model has predicted that the patient is not having the disease when the patient is having the disease. |
| | False Positive – The model has predicted 5 times that the patient is having disease while the patient is not having the disease. |

| **SVM Model** | 91 records: 53 times “Yes” and 38 times “No” Prediction |
| | True Negative – The model has predicted 35 times that the patient is not having heart disease accurately. |
| | True Positive – The model has predicted 48 times that the patient is having the disease accurately. |
| | False Negative – Thrice The model has predicted that the patient is not having the disease when the patient is having the disease. |
| | False Positive – The model has predicted 5 times that the patient is having disease while the patient is not having the disease. |

| **Naïve Bayes** | 91 records: 50 times “Yes” and 41 times “No” Prediction |
| | True Negative – The model has predicted 34 times that the patient is not having heart disease accurately. |
| | True Positive – The model has predicted 45 times that the patient is having the disease accurately. |
| | False Negative – The model has predicted 7 times that the patient is not having the disease when the patient is having the disease. |
| | False Positive – The model has predicted 6 times that the patient is having disease while the patient is not having the disease. |

| **Decision Tree** | 91 records: 50 times “Yes” and 41 times “No” Prediction |
| | True Negative – The model has predicted 34 times that the patient is not having heart disease accurately. |
| | True Positive – The model has predicted 44 times that the patient is having the disease accurately. |
| | False Negative – The model has predicted 7 times that the patient is not having the disease when the patient is having the disease. |
| | False Positive – The model has predicted 6 times that the patient is having disease while the patient is not having the disease. |
5.1 Confusion Matrix

**K-Nearest Neighbour with k = 7.**

- **True Negative**: The model has predicted 33 times that the patient is not having heart disease accurately.
- **True Positive**: The model has predicted 46 times that the patient is having the disease accurately.
- **False Negative**: The model has predicted 5 times that the patient is not having the disease when the patient is having the disease.
- **False Positive**: The model has predicted 7 times that the patient is having the disease while the patient is not having the disease.

Thus, this model with **k = 7** is less risky than with **k = 14** since False Negative(k : 7) > False Negative(k : 14).

**K-Nearest Neighbour with k = 14.**

- **True Negative**: The model has predicted 34 times that the patient is not having heart disease accurately.
- **True Positive**: The model has predicted 47 times that the patient is having the disease accurately.
- **False Negative**: The model has predicted 4 times that the patient is not having the disease when the patient is having the disease.
- **False Positive**: The model has predicted 6 times that the patient is having the disease while the patient is not having the disease.

Thus, this model with **k = 14** is less risky than with **k = 7** since False Negative(k : 7) > False Negative(k : 14).

**Random Forest**

91 records: 53 times “Yes” and 40 times “No” Prediction

- **True Negative**: The model has predicted 35 times that the patient is not having heart disease accurately.
- **True Positive**: The model has predicted 46 times that the patient is having the disease accurately.
- **False Negative**: The model has predicted 5 times that the patient is not having the disease when the patient is having the disease.
- **False Positive**: The model has predicted 5 times that the patient is having the disease while the patient is not having the disease.

Inference from the above comparison: Implementing the Logistic Regression classifier would give the situation of least risk of predicting wrongly that the patient does not have heart disease when the patient is suffering from the disease.

### 5.2 Accuracy

Accuracy is the ratio of true predictions done by a classifier to the sum of all predictions by that model. Thus, it is formulated as below:

\[
\text{Accuracy} = \frac{\text{(Correct Predictions)}}{\text{(Total Predictions)}}
\]  

Now, Correct Predictions is the sum of true positives and true negatives whereas wrongly made predictions is the (false positives + false negatives). Thus, equation (1) can rewrite as:
Accuracy = \frac{(TP + TN)}{(TP+FP+FN+TN)} \quad (2)

It is a good measure for balanced variable classes but should not be used when its classes in the data are a majority. The above-cited technical zone of the project can be consolidated in the form of the Table, where the accuracy of the classifiers is compared in table 3.

**Table 3.** Comparing accuracies of the classifiers.

| Classifier             | Accuracy Level |
|------------------------|----------------|
| Logistic Regression    | $92.30\%$      |
| SVM Model              | $91.20\%$      |
| Naïve Bayes            | $87.91\%$      |
| Decision Tree          | $85.71\%$      |
| K-Nearest Neighbour ($k = 7$) | $86.81\%$ |
|                        | ($k = 14$) $90.11\%$ |
| Random Forest          | $85.71\%$      |

5.3 Sensitivity / Recall

Sensitivity indicates the proportions of cardiac patients diagnosed (by the model) as with cardiac disease. It is also termed recall.

Thus, Recall or Sensitivity for a model is computed as:

\[
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)
\]

For ailment treatment purposes, it is the ability of the test to correctly detect ill patients. Thus, it is the proportion of those testing positive among those having it. 100% recall identifies all with the disease.

5.4 Specificity

Specificity indicates the proportions of patients not having the disease been forecasted by the model to the category of non-cardiac disease. It is the exact opposite of Sensitivity. Thus, the Specificity for a classification problem is given as:

\[
\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \quad (4)
\]

5.5 Precision.

Precision provides information about the proportion of those classified by the model as with the disease, had heart disease. Thus, Precision for a machine-learning model is determined as below.

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (5)
\]

5.6 F1 Score

F1 Score is defined as the harmonic mean of sensitivity (or recall) and precision assigning a single number. F1 Score calculation is expressed as mentioned in equation (6).

\[
\text{F1 Score} = \frac{(2\cdot\text{Precision} + \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (6)
\]

The above-cited technical zone of the project can be consolidated in the form of a table, where the performance metrics – recall, specificity, precision, and F1 score of the implemented models are enlisted and compared in the table below. Also, the proposed work is collated with the existing approaches on the same dataset in table 5.
Table 4. Performance Metric Comparison of the Classifiers.

| Classifier          | Recall / Sensitivity | Specificity | Precision | F1 Score |
|---------------------|----------------------|-------------|-----------|----------|
| Logistic Regression | 0.9608 ± 96.08%      | 0.875 ± 87.5% | 0.9074 ± 90.74% | 0.933 ± 93.34% |
| SVM Model           | 0.9412 ± 94.12%      | 0.875 ± 87.5% | 0.9057 ± 90.57% | 0.9231 ± 92.31% |
| Naïve Bayes Decision Tree | 0.8824 ± 88.24% | 0.875 ± 87.5% | 0.90 ± 90% | 0.9334 ± 93.34% |
| K – Nearest Neighbour | 0.8628 ± 86.28% | 0.85 ± 85% | 0.88 ± 88% | 0.8713 ± 87.13% |
| Random Forest       | 0.8627 ± 86.27%     | 0.85 ± 85% | 0.88 ± 88% | 0.8713 ± 87.13% |

Table 5. Comparing Proposed Model with Existing Works.

| Author               | Approach to predict the cardiac disease                                    | Accuracy       |
|----------------------|---------------------------------------------------------------------------|----------------|
| Proposed Model       | Logistic Regression, SVM Model, Naïve Bayes, K-Nearest Neighbour, Decision Tree, Random Forest | 92.30% Logistic Regression |
| Zhenya and Zhang et al.(2020)[2] | Ensemble Classifier                                                      | 92.15%         |
| Mohan et al.(2019)[7] | Hybrid Random Forest with a linear model(HRFLM)                            | 88.4%          |
| Kohali et al.(2018)[6] | Logistic Regression(heart disease prediction)                             | 87.1% Logistic Regression |
| Archana Singh et al.(2020)[5] | Linear Regression, Decision Tree, SVM Model, k-Nearest Neighbour classifiers | 87%            |
| C. Beulah et al.(2019)[3] | Majorite voting among Bagging, Boosting, Stacking                         | 85.48%         |

6. Conclusion and Future Scope
The results of the proposed work depict that Logistic Regression is better than the other supervised classifiers in terms of the discussed performance metrics – accuracy, precision, sensitivity (or recall), specificity and F1 score. The model gives the results with the highest accuracy of 92.30%. The classifier is also less risky since the number of false negatives is low as compared to other models as per the confusion matrix of all the models. In future, detailed and further algorithm analysis can be done such as determining and comparing the value of estimators for best accuracy using the Random Forest Classifier, analysing the implemented linear kernel in SVM with other SVM Classifier Kernels for distinct and comprehensible visualization of its working for better prediction.

References
[1] Kendale S, Kulkarni P, Rosenberg A D and Wang J 2018 Supervise machine learning predictive analytics for prediction of postinduction hypotension Anaesthesiology 129 675-88.
[2] Zhenya Q and Zhang Z 2021 A hybrid cost-sensitive ensemble for heart disease prediction BMC Med Inform Decis Mak 21 73
[3] Latha C B C and Jeeva S C 2019 Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques Inf. in Med. Unl. 16 100203
[4] Tripoliti E E, Papadopoulos T G, Karanasiou G S, Naka K K and Fotiadis D I Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques Comp. and struct. biotech. 15 26-46.
[5] Singh A and Kumar R 2020 February Heart disease prediction using machine learning algorithms International Conf. on Electrical and Electronics Engineering 452-7
[6] Kohli P S and Arora S 2018 Application of machine learning in diseases prediction 4th International Conf. on Computing Communication And Automation
[7] Mohan S K, Thirumalai C and Srivastava G 2019 Effective heart disease prediction using hybrid machine learning technique IEEE Access 7 81542-54
[8] Sharma H and Rizvi M A 2017 August Prediction of heart disease using machine learning algorithms: A Survey Recent and Innovation Trends in Computing and Communication S IJRITCC
[9] Patel J, Khaked A A, Patel J and Patel J 2020 Heart disease prediction using machine learning Proceedings of 2nd Int. Conf. on Computing, Communications and Cyber Security 653-65
[10] Pattnaik P P, Padhy S R, Mishra B S P, Mishra S and Mallick P K 2021 Heart disease prediction using machine learning technique Advances in Electronics, Communication and Computing 709 193-201.
[11] Sharma S, Yadav S and Gupta M 2020 Heart disease prediction using machine learning techniques 2nd Int. Conf. on Advances in Computing, Communication Control and Networking pp 177-181.
[12] Katarya R and Srinivas P 2020 Predicting heart disease at early stages using machine learning: a survey Int. Conf. on Electronics and Sustainable Communication Systems pp 302-305
[13] Sujatha P and Mahalakshmi K 2020 Performance evaluation of supervised machine learning algorithms in prediction of heart disease IEEE Int. Conf. for Innovation in Technology pp 1-7.
[14] Sharma V, Yadav S and Gupta M 2020 Heart disease prediction using machine learning techniques 2nd Int. Conf. on Advanced in Computing, Communication Control and Networking pp 177-181.
[15] https://www.tutorialspoint.com/machine_learning_with_python
[16] Heart disease dataset. https://archive.ics.uci.edu/ml/datasets/Heart+Disease