**Heart Disease Prediction**

**An Engineering Project in Community Service**

**Phase – I Report**

***Submitted by***

**Team Members List**

**Sl. No. Register Number Name**

1. **20BCE10760 Uzer Khan**
2. **20BAI10240 Aditya Kumar Verma**
3. **20BCE10881 Nimish Sarathe**
4. **20BCE10506 Bharti Gattani**
5. **20BCE10720 Anoop Parashar**
6. **20BCE10023 Ayush Mishra**
7. **20BCE10773 Nimish Singh Chouhan**
8. **20BAI10350 Raj Dama**

***in partial fulfillment of the requirements for the degree of***

***Bachelor of Engineering and Technology***

****

**VIT Bhopal University**

**Bhopal**

**Madhyapradhesh**

**<<March,2023>>**

****

**Bonafide Certificate**

Certified that this project report titled **“Heart Attack Prediction”** is the bonafide work of “20BCE10760 Uzer khan, 20BAI10240 Aditya Kumar Verma, 20BCE10881 Nimish Sarathe, 20BCE10506 Bharti Gattani, 20BCG10023 Ayush Mishra, 20BCE10773 Nimish Chouhan, 20BC10726 Anoop Parashar, 20BAI10350 Raj Dama ”who carried out the project work under my supervision.

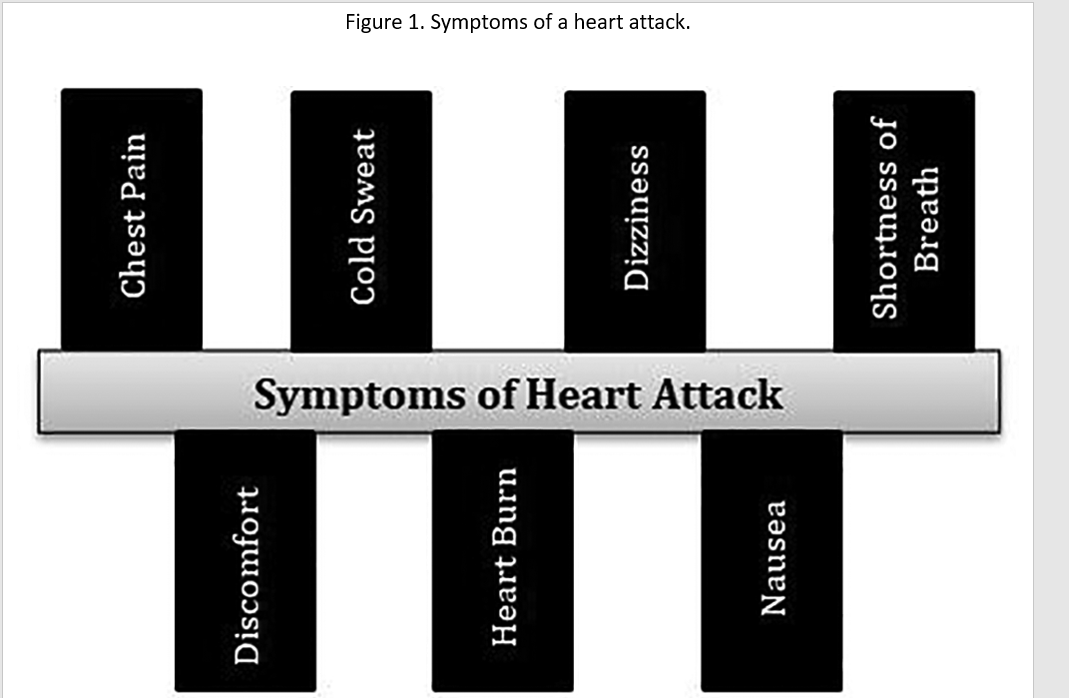
This project report (Phase II) is submitted for the Project Viva-Voce examination held on 10 March 2023.

**Supervisor**

# INTRODUCTION (- Include page numbers at bottom of every page)

A heart attack which is analogous to acute myocardial infarction (AMI) is one of the most serious diseases in the segment of cardiovascular disease. It occurs due to the interruption of blood circulation to muscle of the heart which damages the heart the muscle. Diagnosing heart disease is also a crucial task. The symptoms, physical examination, and understanding of the different signs of this disease are required to diagnose heart disease. Different factors including cholesterol, genetic heart disease, high blood pressure, low physical activity, obesity, and smoking can be reasons for the occurrence of heart disease. The major reason for heart attacks is the stoppage of blood to the coronary arteries. The red blood cells (RBC) start getting low when blood flow is reduced; due to this the human body stops getting necessary oxygen and loses consciousness. The early diagnosis through symptoms and signs can help prevent patients of heart attacks if the prediction is accurate enough. [Figure 1](https://f1000research.com/articles/11-1126#f1) shows different symptoms of a heart attack. The work presented takes 13 features/attributes as input having number values. It has been stated that little modifications in lifestyle including quitting smoking/alcohol/tobacco, having healthy food habits, and routine exercises can help in the prevention of heart attacks. Any person living a healthy lifestyle with early treatment after diagnosis can greatly increase the positive results. However, it is difficult to identify the high risk of heart disease where different risks like diabetes, high blood pressure, and cholesterol problems are present. It is a system that is made by the use of machine learning algorithms for guessing the possible diseases based on the patient’s symptoms. The growth of technology has been improving our lives so far. It provides many tools that can save millions of lives, and machine learning is one of them. Machine Learning is used to develop systems that can help us predict so many diseases based on symptoms. It can suggest the doctors, probability of the possible diseases. And diagnosis can be done based on suggestion, thus cost could be reduced.

We are living in the age of technology and nowadays humans can say that almost anything is possible with the help of technology. Today we have so many tools and methods to access information from any region of this world and Information at this age is so important that without information we would not survive. We have tools that can give us or suggest relevant information at our fingertips and the internet is one of those tools. Today billions of search queries are performed daily and sometimes there given results are relevant and sometimes they are not. In those search queries, thousands of searches are related to medical advice. People often want to know if they have any serious diseases based on their signs and symptoms. But there are no tools available to give them proper information. This research tries to give them tools so that possible disease prediction information can be provided to the end-user.



## 1.1 Motivation

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using three classification algorithms namely Naïve Bayes, Decision Tree, and Random Forest are used at different levels of evaluations. Although these are commonly used machine learning algorithms, the heart disease prediction is a vital task involving highest possible accuracy. Hence, the three algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better understanding and help them identify a solution to identify the best method for predicting the heart diseases.

## Objective

### **Heart disease in the context of machine learning**

Previous works have declared that prediction can be improved with the application of feature selection and proper engineering. An experiment with different machine learning approaches and models by tuning various hyper-parameters has been performed and improved the performance with optimized accuracy.[1](https://f1000research.com/articles/11-1126#ref1) Neural networks performed well when compared to other machine learning classifiers i.e., Naïve Bayes, J48, CART, Grading, and SVM with nearly 79% accuracy.

Other researchers worked on the reduction of cardiovascular features and extracted nonlinear features with discriminant analysis.[2](https://f1000research.com/articles/11-1126#ref2) Fisher was utilized for the experiment’s purpose to tackle overfitting problems and to improve the training speed. Results stated that 100% accuracy has been shown for the detection of coronary disease. [Table 1](https://f1000research.com/articles/11-1126#T1) represents the summary of literature survey done for the work

# 3. Existing Work / Literature Review ( 1- 2 Pages)

There have been numerous studies done related to predicting the disease using different machine learning techniques and algorithms which can be used by medical institutions. This paper reviews some of those studies done in research papers using the techniques and results used by them.

MIN CHEN et al, proposed a disease prediction system in his paper where he used machine learning algorithms. In the prediction of disease, he used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, and Decision Tree. This proposed system had an accuracy of 94.8%.

Sayali Ambekar et al, recommended Disease Risk Prediction and used a convolution neural network to perform the task. In this paper machine learning techniques like CNN-UDRP algorithm, Naive Bayes, and KNN algorithm are used. The system uses structured data to be trained and its accuracy reaches 82% and achieved by using Naïve Bayes. the accuracy of Diabetes is 97.02% and Liver disorder is 96.13.

Dhiraj Dahiwade et al,

[4] designed a model for prediction of the disease using approaches of machine learning and used techniques like KNN and CNN. This paper suggests disease prediction i.e., based on patient’s symptoms. The accuracy of KNN is 95% and the accuracy of CNN is 98%.

Lambodar Jena et al,

[5] focused on risk prediction for chronic diseases by taking advantage of distributed machine learning classifiers and used techniques like Naive Bayes and Multilayer Perceptron. This paper tries to predict Chronic-Kidney-Disease and the accuracy of Naïve Bayes and Multilayer Perceptron is 95% and 99.7% respectively.

Dhomse Kanchan B. et al,

[6] studied special disease prediction utilizing principal component analysis using machine learning algorithms involving techniques like Naive Bayes classification, Decision Tree, and Support Vector Machine. The accuracy of this system is 34.89% for Diabetes and 53% for heart disease.

Pahulpreet Singh Kohli et al,

[7] suggested disease prediction by using applications and methods of machine learning and used techniques like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest and Adaptive Boosting. This paper focuses on predicting Heart disease, Breast cancer, and Diabetes. The highest accuracies are obtained using Logistic Regression that is 95.71% for Breast cancer, 84.42% for Diabetes, and 87.12% for heart disease.

Deeraj Shetty et al,

[8] studied the uses of data mining for diabetes disease prediction by using Naïve Bayes and KNN algorithms. This system predicts diabetes and accuracy obtained by KNN are better than Naïve Bayes.

Rashmi G Saboji et al,

[9] tried to find a scalable solution that can predict heart disease utilizing classification mining and used Random Forest Algorithm. This system presents a comparison against Naïve-Bayes classifier but Random Forest gives more accurate results with accuracy 98%.

Rati Shukla et al,

[10] suggested prediction and detection for breast cancer by utilizing machine learning techniques like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network, and KNN. In this system, the Support Vector Machine gives more accurate results than all other algorithms.

Senthilkumar Mohan et al,

[11] focused on hybrid techniques in machine learning that can be used for effectively predicting heart disease and used algorithms like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN. The accuracy of this system is 88.47%.

Anjan Nikhil Repaka et al,

[12] designed and implemented a prediction model for heart disease using naive Bayesian. Any user can use this system using any smartphone device and get the prediction results. The accuracy of this system is 89.77%.

Aakash Chauhan et al,

[13] proposed a disease prediction model for heart disease by utilizing evolutionary rule.

learning. Association Rule is used in this proposed system. This system is not very efficient because it has an accuracy of 53%.

Aditi Gavhane et al, [14] suggested prediction for heart disease that utilizes Machine Learning. Multi-Layer Perceptron model is used in this system. This system predicts heart disease based on basic symptoms like age, sex, pulse rate, etc. The accuracy of this suggested system is 91%.

# Proposed Work ( 1- 2 Pages) –

* + There are following steps involved in our proposed methodology:
    - First, I will collect the datasets of symptoms and them
  + functional problem in the body.
    - Then I will collect the information that will associate the symptoms to possible diseases thus related disease information will be collected.
    - Then I will get the symptoms as input from the patient and process it by Logistic Regression.
    - After that Logistic Regression predicts the diseases that may be possible for those acquired symptoms.
    - Then the system will show the diagnosis in the form of max possible disease and min possible disease.



1. **System Design / Architecture-**

The architecture of DPS includes multiple following fields:

Input

We are taking input from the user of the disease prediction system as a symptoms list.

**Get Data**

In this field, the user will provide data about their symptoms.

**Data Acquisition and Processing**

In this field, the input is provided for processing. Data acquisition and processing perform two operations, first is the acquiring the data and then second is the processing of the data and extracting information based on that acquired data.

**Get Symptoms of the Body**

In this field symptoms of the body are gathered and analysed. So that this information can be used by the algorithm to predict the possible diseases.

**Get Functional Problem of Body**

In this field, functional problems of the body that is associated with the symptoms are gathered. So that it is analysed to get the possible disease.

**Dataset Disease (symptoms, functions)**

In this field, we have a predefined dataset of diseases that involves symptoms and functions that are caused by the disease. This dataset is further used to match the data that has been obtained from the user and if matched properly then the system will suggest the possible diseases.

**Train Data ()**

In this field training of the system is performed. Our disease prediction system is trained using the SVM (support vector machine) algorithm. Here we are using the SVM algorithm to solve a problem related to regression.

**Previses (Dataset\_Disease)**

In this field Dataset of the diseases is provided as parameter and processing are performed based on this dataset.

**Predict\_Logistic\_Regression (symptoms, function)**

In this field, the prediction is performed using the MLR algorithm. In MLR, multiple independent variables are used to perform the prediction of the disease. Symptoms and their functions in the user’s body are involved in the prediction.

**Possible\_Disease (symptoms, function)**

In this field symptoms and functions are passed as a parameter and possible diseases are calculated based on these parameters.

**Data Processing**

This field contains the above five data processing fields and is the main part of our disease prediction system. It has all the necessary fields for processing the data.

**Output**

After Data Acquisition and Processing, possible diseases are generated as output.

**Max Possible Disease**

This field contains the maximum possible disease as output.

**Min Possible Disease**

This field contains the minimum possible diseases as output.



**b) Working Principle**

**1. Logistic regression**: One of the very popular algorithms is considered as logistic regression which is a supervised learning model. It performs categorical predictions which can be ‘true’ or ‘false’. This model provides probabilistic values instead of exact ones. This algorithm works on both continuous and discrete values. A simple S-Shaped curve can elaborate the logistic regression very precisely.

**2. Support vector machines (SVM):** It is a supervised learning model which works on the concept of decision boundary or hyper plane. The aim of the algorithm is to maximize the margin of the hyper planes which helps in minimizing the misclassification problem. Model chooses extreme points to create the decision boundary which are called as support vectors.

**ALGORITHM FOR DISEASE PREDICTION SYSTEM-**

As the name suggests, in our disease prediction system, we are using the Support Vector Machine (SVM) for classification and Logistic Regression (MLR) for predicting the result. MLR is a form of regression algorithm where multiple independent values are involved, meaning that we try to predict a value based on two or more variables.

Simple Linear Regression, where a single Independent/Predictor(X) variable is used to model the response variable (Y). But there may be many situations in which the reply variable is affected by multiple forecaster variables; for such cases, we use the MLR algorithm.

The algorithm that we are using in our proposed system is given below and by using this algorithm we are getting the accuracy of up to 87%.

i. Take input of symptoms in p[ ] and their function in t[ ].

ii. Declaration:

S[n][m] - m set of symptoms of n disease

F[n][m] - m set of functions of n disease

iii. , and it has a minimum cardinality.

iv. Set Ssvm[ ] = new set of possible disease symptoms

v. Set Fsvm[ ] = new set of possible disease functions

vi. For each x, y in Ssvm[ ] and Fsvm[ ]

vii. If p[ ] in Ssvm[ ] and t[ ] in Fsvm[ ]

viii. Pdisease[ ] = Ssvm[ ], Fsvm[ ], Priority++

ix. endIf

x. endFor

xi. Possible disease Pdisease[0]

Explanation of the above-written algorithm:

• Input is taken from the user in the form of symptom and stored in p[ ] and function is stored in t[ ].

• In the second step, the declaration is done and S[n][m] stores m set of symptoms of n disease and F[n][m] stores m set of functions of n disease.

• , it has a minimum cardinality.

• Set Ssvm[ ] is a new set of possible disease symptoms

• Set Fsvm[ ] = new set of possible disease functions

• In this step, we are using a for loop to check each value of Ssvm[ ] and Fsvm[ ]

• And if given input lies within Ssvm[ ] and Fsvm[ ].

• Then we are storing that disease in Pdisease[ ] and increasing the priority of that disease.

• If statement end.

• For loop end.

• The possible disease is given as output Pdisease[0].

1. **Expected Results-**

The analysis and identification of the best classification algorithm in this research work is done and the results are provided here. For the validation of the results, several range of experiments are carried out using Cross validation and Percentage split methods which are described in the sections given below.

In this work, the evaluation of the performance metrices are being done with four machine learning classifiers i.e., SVM and Logistic regression.

We will have a live dashboard so that the prediction of the disease can be predicted

And can be seen visualy using a React.JS hosted website.

# CONCLUSION ( Phase:2) ( 1 Page)

In our research, we have used a support vector machine and Logistic regression algorithm to predict diseases. And we have also tested multiple algorithms like the k-nearest neighbor, convolution neural network, decision tree, etc. Despite testing these algorithms I have found that the support vector machine and Logistic regression combination gives higher accuracy than other algorithms.

The purpose of this research was to provide medical diagnosis information based on symptoms to normal people, fresher doctors, medical students, and anyone who wants to know about a set of symptoms and associated diseases.

In this research, we have found that possible disease prediction can go up to 87% for some diseases and minimum 68% for some diseases but these results are obtained using the minimum amount of data set but if we can feed the system humongous amount of data set then this disease prediction system can give accuracy up to 95%. Obtaining a humongous amount of data set related to diseases and their symptoms is very time consuming and it cannot be done within one or two years it requires multiple years to collect those data sets and train the system using those data searches. This system can be used by Ph.D. scholars to do further research.

With the help of a disease prediction system, it was possible to diagnose people based on symptoms. Disease prediction system provides only possible outcomes it does not guarantee that it will predict the disease correctly. But it has significantly higher accuracy for predicting possible diseases. In our research, we have analyzed the accuracy of this system for 5 different diseases and our accuracy can go up to 87%.

Tech Stack Used : **HTML ,CSS , JAVASCRIPT, REACT.JS, PYTHON(SCIKIT LEARN,NUMPY,PANDAS), GOOGLE COLLAB & VS CODE**

# 

# Reference:

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities” IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[2] **S**ayali Ambekar, Rashmi Phalnikar, “Disease Risk Prediction by Using Convolutional Neural Network” IEEE, 978-1-5386-5257-2/18, 2018.

[3] Naganna Chetty, Kunwar Singh Vaisla and Nagamma Patil, “An Improved Method for Disease Prediction using Fuzzy Approach” IEEE, DOI 10.1109/ICACCE.2015.67, pp. 569-572, 2015.

[4] Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, “Designing Disease Prediction Model Using Machine Learning Approach” IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4, pp. 1211-1215, 2019.

[5] Lambodar Jena and Ramakrushna Swain, “Chronic Disease Risk Prediction using Distributed Machine

Learning Classifiers” IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017.

[6] Dhomse Kanchan B. and Mahale Kishor M., “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis” IEEE, 978-1-5090-0467-6/16, pp. 5-10, 2016.

[7] Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Disease Prediction” IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, ” Diabetes Disease Prediction Using Data Mining” IEEE, 978-1-5090-3294-5/17, 2017.

[9] Rashmi G Saboji and Prem Kumar Ramesh,“A Scalable Solution for Heart Disease Prediction using Classification Mining Technique” IEEE, 978-1-5386-1887-5/17, pp. 1780-1785, 2017.

[10] Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.

[11] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access, DOI 10.1109/ACCESS.2019.2923707, pp. 81542-81554, 2019.

[12] Anjan Nikhil Repaka, Sai Deepak Ravikanti and Ramya G Franklin,”Design And Implementing Heart Disease Prediction Using Naives Bayesian” IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8, pp. 292-297, 2019.

[13] Aakash Chauhan, Purushottam Sharma, Vikas Deep and Aditya Jain, “Heart Disease Prediction using Evolutionary Rule Learning” CICT 2018.

[14] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, “Prediction of Heart Disease Using Machine Learning” IEEE Xplore ISBN: 978-1-5386-0965-1, pp. 1275-1278, 2018.

[15] Ankita Dewan and Meghna Sharma, “Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification” IEEE, 978-9-3805-4416-8/15, pp. 704-706, 2015.