IMPROVING THE ACCURACY AND COMPARING MACHINE LEARNING TECHNIQUES IN HEART DISEASE PREDICTION

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE OF B.TECH PROJECT-I (CO401)

OF

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted By:
Aanchal Kumari Shah (2K20/CO/004)
Bibek Khadka (2K20/CO/126)
Rohit Kumar Shah (2K20/CO/374)

Under the supervision of
Dr. Rakesh Kumar Yadav

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

DECEMBER, 2023

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLERATION

We, Aanchal Kumari Shah (2k20/CO/004), Bibek Khadka (2k20/CO/126), Rohit Kumar Shah (2k20/CO/374) hereby declare that the project Dissertation titled "**Improving the Accuracy and Comparing Machine Learning Techniques in Heart Disease Prediction**" which is submitted by us to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor Of Technology, is original and not copied from any source without any proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi — Name of Student

Date: **Aanchal Kumari Shah**

Bibek Khadka

Rohit Kumar Shah

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "IMPROVING THE ACCURACY AND COMPARING MACHINE LEARNING TECHNIQUES IN HEART DISEASE PREDICTION " which is submitted by Aanchal Kumari Shah (2K20/CO/004), Bibek Khadka (2K20/CO/126), Rohit Kumar Shah (2K20/CO/374), Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology/Bachelor of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

NAME OF SUPERVISOR

Dr. Rakesh Kumar Yadav

Assistant Professor,
Department of Computer Science and Engineering,
Delhi Technological University,
Delhi, India

ABSTRACT

In medicine, diagnosing heart illness is extremely difficult. It is difficult because in order to make decisions, physicians must review a large amount of clinical and pathological data. Because of this intricacy, medical professionals and academics are now more motivated to develop precise and effective methods for predicting cardiac disease. Time is a major element, thus early detection of heart disease is essential. It's critical to detect heart disease early because it's one of the main causes of mortality across the country. With appropriate training and testing, machine learning has emerged as a dependable medical tool in recent years, greatly aiding in the prediction of diseases. This study's primary objective is to investigate several models for heart disease prediction and identify significant characteristics associated with it by utilizing the Support Vector Machine and Logistic Regression Algorithm method. In terms of accuracy, Support Vector Machine Algorithm performs better than other machine learning algorithms like logistic regression. Our objective is to use different machine learning Algorithm to ascertain a person's risk of heart disease as well as improving the accuracy.

ACKNOWLEDGEMENT

We would like to offer our gratitude and thanks to each one of the individuals who gave us the help to finish this task. We would like to express our extraordinary gratitude to my mentor, Dr. Rakesh Kumar Yadav, for their guidance and support throughout this project. Dr. Rakesh Kumar Yadav was always available to answer my questions and provide me with feedback. He has been the guiding light and inspired us to proceed with this project. We also want to thank all of our department faculty members for their assistance, ongoing encouragement, and unwavering support in making this project a success. Last but not least we appreciate our parent's efforts in assisting us with our initiative during these trying times.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

CHAPTER 1

INTRODUCTION

The heart, a muscular organ that pumps blood into the body, is the primary part of the body's circulatory system, which also includes the lungs. Capillaries, arteries, and veins make up the cardiovascular system's blood vessel network. These blood arteries carry blood throughout the body. Heart problems, often known as cardiovascular diseases(CVD), are brought on by deviations from the heart's regular blood flow. Heart issues are the primary cause of death worldwide. According to estimates from the World Health Organisation(WHO), heart attacks and strokes are responsible for 17.5 million deaths globally. More than 75% of cardiovascular disease-related deaths take place in middle-class and lower-class countries. Moreover, 80% of deaths from CVD are caused by heart attacks and strokes. So, early identification of cardiac abnormalities and the creation of technology to predict heart disorders can prevent many deaths and help medical practitioners design effective treatment regimens that reduce the mortality rate from cardiovascular diseases. Because of the advancement of sophisticated healthcare systems, a large amount of patient data is now accessible, and this data may be used to generate prediction models for cardiovascular diseases. The healthcare industries of today generate massive volumes of data regarding patients, illnesses, and other relevant subjects.

Thus, in order to create a heart disease prediction system, a machine learning technique is applied in this project and verified on public heart disease prediction datasets. Investigating the mysterious dataset patterns in the clinical domain's data sets may be greatly aided by medical data mining. These patterns can be used to diagnose medical conditions. Nonetheless, the raw medical data that are now accessible are widely dispersed, substantial, and varied in kind. It is necessary to gather this data in an organized manner. After gathered, this data may be used to create a medical information system.

This project uses classification algorithms to analyze the heart disease forecasts. In healthcare data, these imperceptible patterns can be used to diagnose medical conditions. The most common cause of death for victims in nations likes the US and India was heart disease. With the help of categorization algorithms, we are forecasting heart illness in this project. To investigate various heart-based issues, machine learning method used as classification algorithms like Support Vector Machine(SVM) algorithm, Logistic Regression algorithm & Artificial Neural Network(ANN) are employed.

CHAPTER 2

LITERATURE SURVEY

Machine learning techniques are employed for the analysis and prediction of medical data information resources. In medicine, making the diagnosis of heart disease is an important and time-consuming process. All conditions that impact the heart are included under the umbrella term "heart disease Better accuracy is achieved in data categorization thanks to the use of the Supervised Machine Learning algorithm. Here, the dataset for heart disease is trained and the disease is predicted using the SVM Algorithm, Logistic Regression Algorithm and ANN Algorithm training method. The outcomes demonstrated that a well-designed prediction algorithm and medication prescription may accurately anticipate a heart disease.

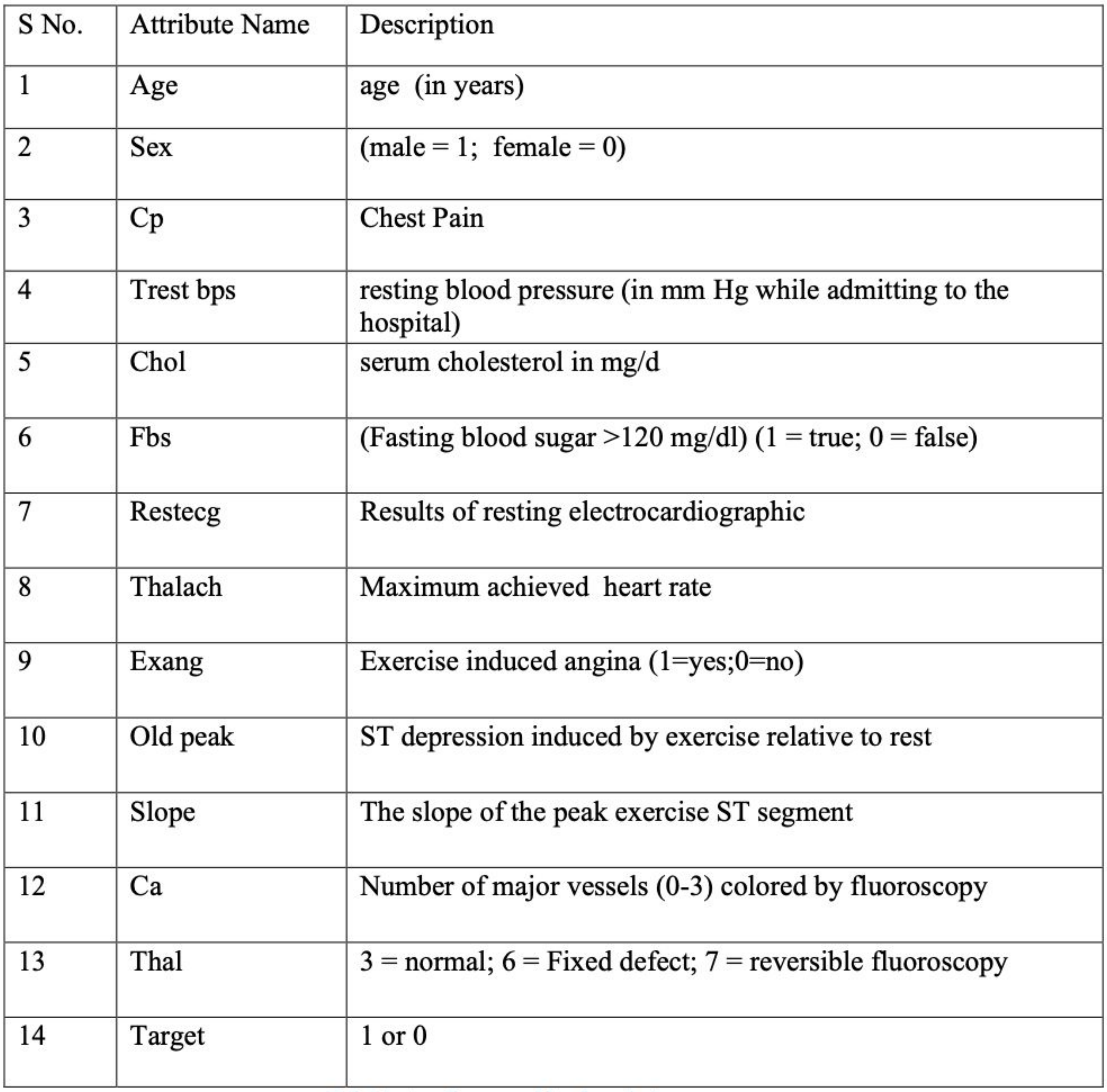
In recent years, a number of experiments and research projects have been conducted in conjunction with the expanding field of medical science and machine learning, leading to the publication of important publications on the heart disease.

Singh, Archana, and Rakesh Kumar. "Improving the Accuracy and Comparing ML Techniques in Heart Disease Prediction" using Support Vector Machine, KNN, Linear Regression, Decision Tree algorithms. They utilised the Kaggle dataset, and before using classification methods, the data is preprocessed. The accuracy of the system is about 83%[1]. Qadri, Azam Mehmood; Raza, Ali; Munir, Kashif; Almutairi, Mubarak S. " Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning" using Support Vector Machine, KNN, Linear Regression, Decision Tree algorithms algorithms for prediction of heart disease. They have also used Kaggle data set. Decision tree algorithm gives an 82% accuracy than other algorithms. Mohan, Senthilkumar; Thirumalai, Chandrasegar; Srivastava, Gautam " Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques." Using K-Nearest Neighbour Algorithms, Decision Tree Algorithm, Genetic Algorithm, Navie Bayes Algorithm that are used especially in the prediction of Heart Disease. They produced an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model[3]. Li, Jian Ping; Haq, Amin Ul; Din, Salah Ud; Khan, Jalaluddin; Khan, Asif; Saboor, Abdus. "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare" Using Support Vector Machine Algorithm, Logistic Regression Algorithm, Artificial Neural Network Algorithm, K-Nearest Neighbour Algorithm, Navie Bayes Algorithm and Decision Tree Algorithm for prediction of heart disease. When compared to other methods, the accuracy of SVM using the suggested feature selection approach is 92.37%, which is excellent.

Data Source:

A large quantity of data on patients and their illnesses has been gathered by Kaggle databases. A set of records with health-related characteristics was acquired from the Kaggle Heart Disease database. The dataset is used to extract the patterns important to the diagnosis of heart attack. The training dataset and testing dataset each received an equal portion of the records. Each and every attribute has a numerical value. We are focusing on a smaller collection of characteristics – just 14 in all.

The following table shows the list of attributes on which we are working.



S No.	Attribute Name	Description
1	Age	age (in years)
2	Sex	(male = 1; female = 0)
3	Cp	Chest Pain
4	Trest bps	resting blood pressure (in mm Hg while admitting to the hospital)
5	Chol	serum cholesterol in mg/d
6	Fbs	(Fasting blood sugar >120 mg/dl) (1 = true; 0 = false)
7	Restecg	Results of resting electrocardiographic
8	Thalach	Maximum achieved heart rate
9	Exang	Exercise induced angina (1=yes;0=no)
10	Old peak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	3 = normal; 6 = Fixed defect; 7 = reversible fluoroscopy
14	Target	1 or 0

Tab 01: Attributes and its description

CHAPTER 3

AIM & SCOPE OF PRESENT INVESTIGATION

3.1 Existing System:

Clinical decisions are often made using a physician's experience and intuition rather than the wealth of information hidden in a database. This practice reduces the quality of care provided to patients by causing unwanted biases, errors, and excessive medical costs. A medical misdiagnosis can occur in a variety of ways. A misdiagnosis of a major disease can have extremely detrimental implications, regardless of whether the medical staff or a doctor is at blame. According to data from the National Patient Safety Foundation, 42% of medical patients believe they have had a medical mistake or missed diagnosis. Sometimes, other considerations, including the price of medical tests, medications, and surgeries, recklessly take precedence over patient safety. Medical misdiagnosis poses a severe threat to the healthcare industry. If they persist, people will come to fear receiving medical care at a hospital. By educating the public and bringing legal action against the negligent medical professionals, we can eradicate medical misdiagnosis.

Disadvantages:

- Early phases are unpredictably impossible to predict.
- It takes time to use acquired data practically under the current system.
- Any errors made by the physician or hospital employees in foreseeing what would happen might result in deadly events.
- Before beginning treatment, a very costly and time-consuming procedure must be carried out to determine the patient's future risk of developing heart disease.

3.2 PROPOSED SYSTEM:

This section presents an overview of the suggested system and demonstrates all of the parts, methods, and resources that were employed in the system's development. An effective software tool is required to train large datasets and compare several machine learning algorithms in order to create a heart disease prediction system that is both clever and easy to use.

3.3 FEASIBILITY STUDY:

An initial investigation into the possibility of a project succeeding before actual work on it starts is called a feasibility study. It is an examination of several potential solutions to an issue and a suggestion for the best one.

3.3.1 Economic Feasibility:

It is described as the procedure for weighing the advantages and disadvantages of a project's development. A suggested solution needs to be a wise investment for the company and be technically and operationally viable. Users would profit immensely from the suggested system as they will be able to distinguish between fake and real news, and they will be aware of the majority of fake and true news that has been published in recent years. High system setup and extra software are not required for this suggested system. As a result, the suggested solution is financially feasible.

3.3.2 Technical Feasibility:

This refers to whether the suggested system can be constructed taking into consideration technical concerns such as the capacity, responsiveness, extensibility, and availability of the required technology. Python will be used for the project's construction. The Google Collaboratory Notebook is an easy-to-learn and efficient tool for expert programmers to utilize in a distributed online environment. Since the growing organization has all the resources necessary to construct the system, it is theoretically possible to implement the proposed system.

3.3.3 Operational Feasibility:

The process of determining how effectively the proposed system addresses business issues or controls business opportunities is known as operational feasibility. The system doesn't require any further complex training because it is self-explanatory. The system comes with built-in classes and methods that are necessary to generate the output. Even a beginner user may manage the program with ease. A user has to spend 14 minutes or fewer getting taught altogether. Since the software used to create this program is widely accessible and reasonably priced, it is a good choice. As a result, the suggested system may be implemented.

CHAPTER 4

METHODS AND ALGORITHMS USED

4.1 Introduction to Requirement Specification:

Witold Pedrycz and James F. Peters, "Software Engineering" Java by Ka, headfirst. A software product, program, or groups of programs that carry out a certain set of tasks in a target environment are described in a Software Requirements Specification (SRS) (IEEE Std. 830-1993).

a. Purposes: The software requirements specification (SRS) outlines the goals and target audience for the document.

b. Scope: The SRS's scope specifies the software product that must be created, as well as its capabilities, application, and pertinent items. Passive Aggressive Algorithm, which uses the test and training data sets, is suggested for implementation.

c. Definitions, Acronyms and Abbreviations Software requirements specification: It is an explanation of a certain software product, application, or group of applications that carry out a certain task in the intended environment.

d. References: IEEE Recommended Practice for Software Requirements by Bert Bates and Sierra, IEEE Std. 830-1993.

e. Summary: The SRS includes information on user attributes, product functionalities, DFDs, and process details. If any, the non-functional criteria are also included.

f. General overview: This section of the SRS describes the primary functions related to the product. The attributes of an individual utilizing this product are mentioned. This section's assumptions are the outcome of discussions with the project's stakeholders.

4.2 REQUIREMENT ANALYSIS

The process of building software begins with the Software Requirement Specification (SRS). It became clear as the system got more complicated that it was difficult to understand the overall objective of the system. Thus, the requirement phase became necessary. The software project is initiated by the client needs. The SRS is the tool used to convert customer concepts (the input) into a formal document (the requirement phase's output). The focus of requirement specification is on describing the analysis that has been conducted, including representation, languages and tools for specifications, and ensuring that the specifications are taken into consideration during the activity.

The requirement phase comes to an end with the preparation of the certified SRS document. Making the SRS document is the main goal of this phase. The Software Requirement Specification is meant to help clients and developers communicate with one other. The software requirement specification is the instrument used to precisely describe customer and user requirements. That is the initial stage of software creation. A strong SRS should satisfy all of the system's stakeholders.

4.2.1 Operational Requirements:

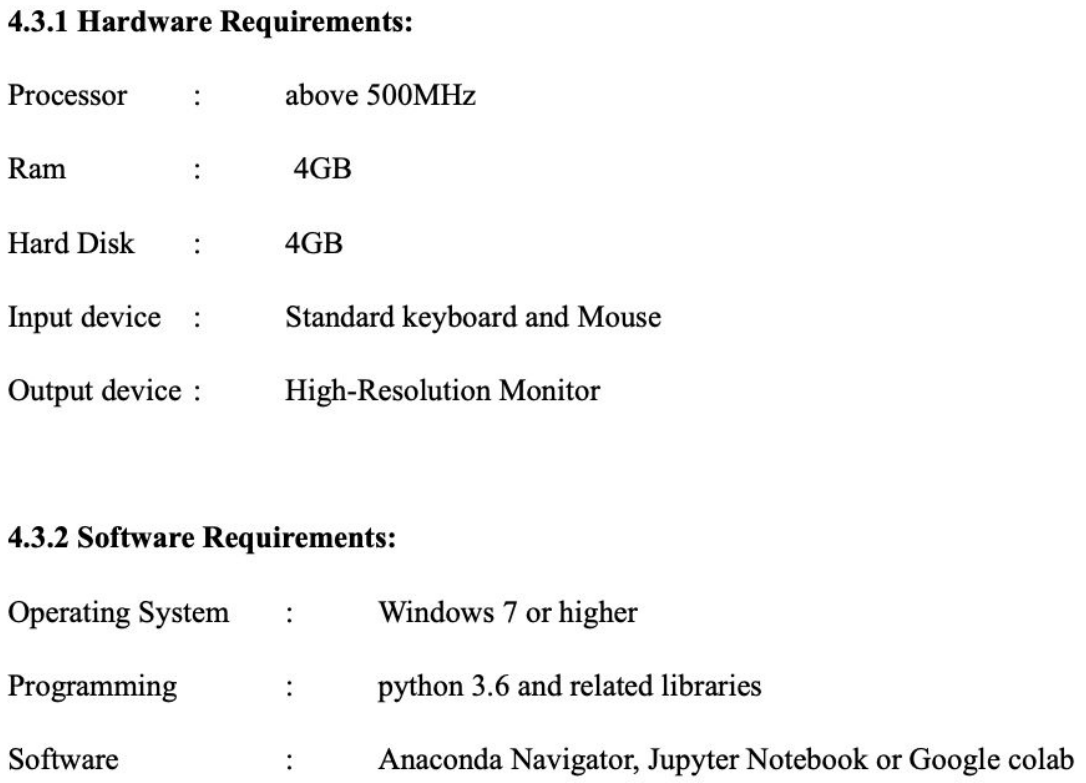
a. Economic: Since no hardware interface, etc., is needed, the produced product is inexpensive. Environmental statements of fact and assumptions that define the expectations of the system, in terms of mission objectives, environment, constraints and measures of effectiveness and suitability. The eight main roles of systems engineering are performed by the customers, with particular attention paid to the operator as the prime customer.

b. Health and Safety: There's a chance the programme poses a safety risk. If so, there are problems with its degree of integrity. Even while the software is a component of a system that is crucial to safety, it could not be.

   - Writing "perfect" code in a language is meaningless if the system software and hardware (in the broadest sense) are unreliable. A computer system should not execute software with a lower integrity level while also supporting software with a higher integrity level.

   - Systems need to be kept apart if their safety standards differ. If not, all systems operating in the same environment must adhere to the strictest possible integrity standards.

4.3 SYSTEM REQUIREMENTS:



4.3.1 Hardware Requirements:

Processor	:	above 500MHz
Ram	:	4GB
Hard Disk	:	4GB
Input device	:	Standard keyboard and Mouse
Output device	:	High-Resolution Monitor

4.3.2 Software Requirements:

Operating System	:	Windows 7 or higher
Programming	:	python 3.6 and related libraries
Software	:	Anaconda Navigator, Jupyter Notebook or Google colab

4.4 SOFTWARE DESCRIPTION

4.4.1 Python:

Python is one of the most popular programming languages available today. It is an interpreted language that may be used in a variety of settings. PyCharm, PyDev, Jupyter Notebook, Visual Studio Code, and other compilers make it easier to run Python scripts. These compilers are specialized software tools that convert high-level, human-readable code to low-level, machine-readable code. They're made using particular programming languages.

Python does not convert its code into hardware-understandable machine code. Rather, code is compiled into bytecode. Python does not produce machine languages; instead, it goes through a compilation process. The bytecode (.pyc or.pyo) that the CPU generates cannot be directly interpreted by the CPU. Therefore, in order to run the bytecode, an interpreter—more especially, the Python virtual machine—is required.

4.4.2 Pandas:

Due to its resilient data structures, Pandas is a Python toolkit with open-source attributes that provide powerful capabilities for both data manipulation and analysis. The name "Pandas" is derived from "panel data," reflecting its origins in facilitating economic analysis of multidimensional data analyse economically. When developer Wes McKinney realized he needed a versatile, potent tool for data research in 2008, he began working on pandas. Python was widely utilized for preprocessing and data mining before Pandas. It didn't actually improve the data analysis in any meaningful way.

The solution to this has been discovered by pandas. Regardless of the source of the data, pandas can be used to do the five standard phases in data processing and analysis: load, prepare, manipulate, model, and analyze. Numerous academic and professional domains, including finance, economics, statistics, analytics, and other areas, employ Python with Pandas.

One of Pandas' main features is its Data Frame object, which is quick and effective and has both default and customizable indexing.

- Tools for importing data from various file formats into in-memory data objects.
- Aligning data and addressing missing data through an integrated strategy.
- Modifying and rearranging date sets.
- Cutting, indexing, and subsetting huge data sets with labels.
- A data structure's columns can be added or removed.
- Group data in preparation for processing and aggregation.
- Excellent results when it comes to data merging and combining.
- The ability to work with time series.

4.4.3 NumPy:

A flexible tool for general-purpose array processing is NumPy. In addition to a number of tools designed specifically for effective array manipulation, NumPy offers a multidimensional array object with excellent performance. NumPy is a basic Python module for scientific computing that has a number of interesting features. Among these are its outstanding performance and multidimensional array support, which establish it as a fundamental component of array-based operations in numerical and scientific computing.

- Robust N-dimensional array;
- Intricate broadcasting procedures
- Code integration tools for C/C++ and Fortran

NumPy boasts robust attributes for handling N-dimensional arrays, intricate broadcasting functions, and seamless integration of C/C++ and Fortran code functionalities. It offers practical tools for diverse applications, including linear algebra, random number generation, and the Fourier transform. With its strong capabilities in managing N-dimensional arrays and advanced broadcasting operations, NumPy facilitates the combination of C/C++ and Fortran code while providing useful functions for tasks like random number generation and linear algebra. Moreover, beyond its primary role in scientific applications, NumPy serves as a versatile multi-dimensional container for common data, enabling easy integration with a wide array of databases due to its ability to create various data types swiftly and effortlessly.

4.4.4 Sckit- Learn:

i. Easy-to-use and effective data mining and analysis tools.
ii. Reusable in a variety of settings and available to everybody.
iii. based on matplotlib iv, scipy, and numpy.
iv. Open source and useable for business - BSD authorization.

4.4.5 Google Colab

Google offers a cloud-based platform called Google Colab, sometimes known as Google Colaboratory, which enables people to build and run Python programmes together. It provides a Jupyter notebook environment that is hosted and has the following essential features:

i. **Free GPU Access:** Users may speed up their machine learning computations by using Colab's free GPU and Tensor Processing Unit access.

ii. **Cloud-Based:** No local installs are required; users may operate and share Jupyter notebooks straight from the cloud. It makes use of Google's computing infrastructure.

iii. **Google Drive integration:** Colab and Google Drive are smoothly connected, enabling users to store and share their work. Google Drive is where notebooks are kept and may be shared with colleagues with ease.

iv. **Support for Multiple Libraries:** TensorFlow, PyTorch, OpenCV, and other well-known Python libraries are just a few of the ones that Colab supports, which make it appropriate for a variety of data science and machine learning applications.

v. **Real-Time Collaboration:** This feature allows many users to work together in real-time on the same notebook, creating a collaborative environment for coding.

vi. **Access to Pre-trained Models:** Colab is a helpful tool for users working on deep learning and machine learning projects since it provides pre-trained machine learning models, datasets, and other resources.

vii. **Interactive Data Visualisation:** Matplotlib, Seaborn, and Plotly are just a few of the libraries that Colab supports for interactive data visualisation, which improves data exploration and display in notebooks.

viii. **Simple Sharing:** By just sharing the URL, notebooks made in Colab may be shared with others. The code may be seen and executed by recipients without a Colab account being needed.

ix. **Code Snippet Library:** To help users quickly access frequently used code patterns and examples, Colab comes with a library of code snippets for a variety of activities.

Due to its free GPU access, collaborative capabilities, and simplicity of use, Google Colab is very well-liked in the data science and machine learning fields. As such, it's a useful tool for a range of tasks.

4.5 ALGORITHMS

4.5.1 Logistic Regression

A popular statistical technique for forecasting binary outcomes (y = 0 or 1) is logistic regression. It is primarily used for forecasting categorical results, whether binomial or multinomial, and is abbreviated as LogR. Based on input variables x, this method calculates the likelihood that an event will occur (y=1); LogR outcomes range from 0 to 1.

The conventional logistic function, or sigmoid curve—an S-shaped curve that is provided by the equation- is used by LogR to represent the data points.

Assumptions for Logistic Regression:

- A binary dependent variable is necessary for the use of logistic regression.
- In a binary regression, the intended result should be represented by the dependent variable's factor level 1.
- The variables that matter should be the only ones mentioned.
- There should be no relationship between the independent variables.
- Requiring high sample sizes is necessary for logistic regression.
- Despite being often applied to binary variables (two classes), logistic (logit) regression may also be applied to categorical dependent variables that have more than two classes.
- This type of regression is known as multinomial logistic regression.

$$f\left(x\right)=\frac{1}{1+e^{-\left(x\right)}}$$

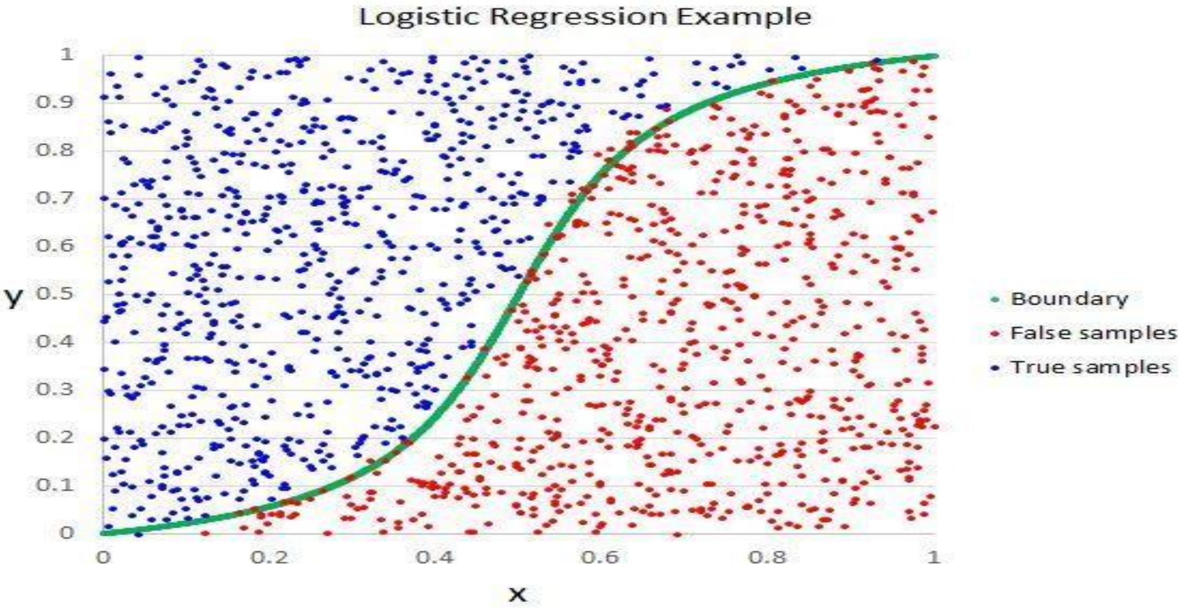


Figure 1: Logistic Regression

4.5.2 SUPPORT VECTOR MACHINE

In order to address regression and classification issues, supervised machine learning techniques such as Support Vector Machines (SVM) are employed. It functions by determining which hyperplane in a high-dimensional space is best for classifying data points into various groups. A "support vector" is the collection of data points required to identify the hyperplane. SVM is resilient and efficient in managing complicated decision boundaries since its goal is to maximize the margin between classes. By using various kernel functions, it may be applied to classification jobs that are both linear and non-linear. SVM is widely used in many different areas, including bioinformatics, text categorization, and picture classification.

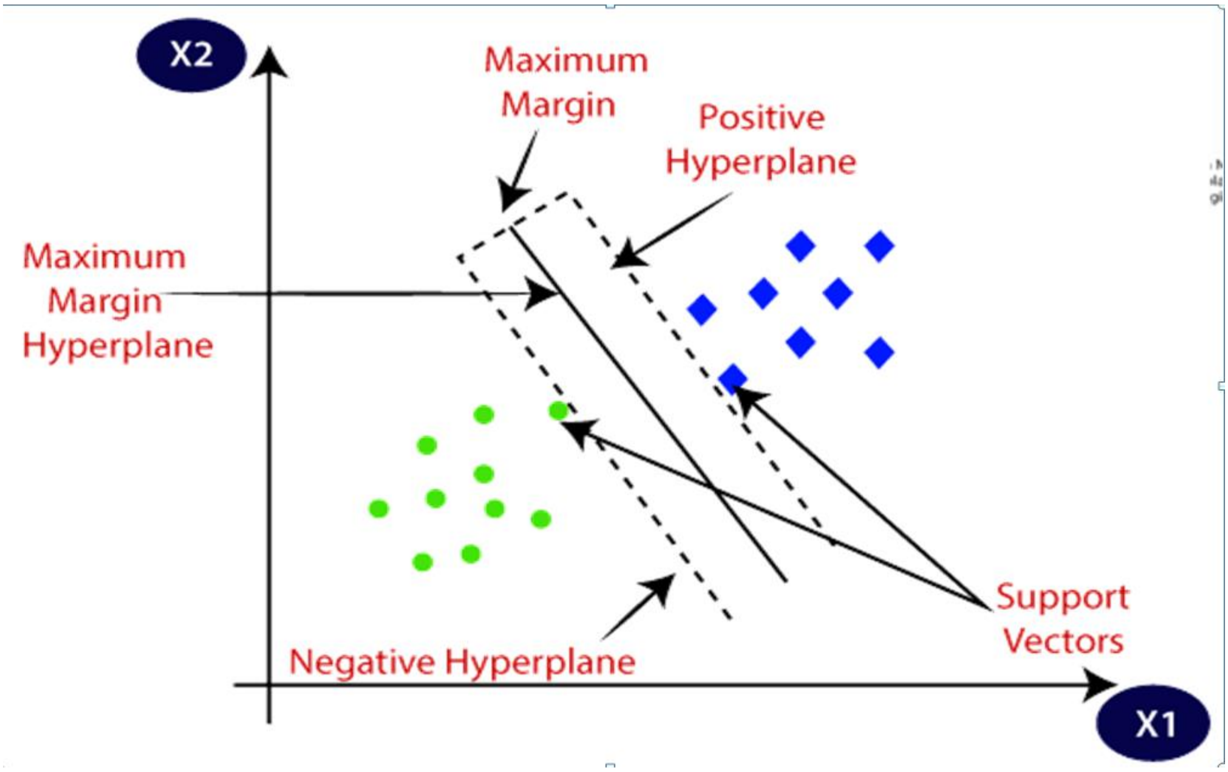


Figure 02: Support Vector Machine

4.5.3 ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are made up of units, also referred to as artificial neurons. These units are arranged in a series of layers to form the Artificial Neural Network of a system. Several dozen to millions of units can make up a layer, depending on how many intricate neural networks are required to search the required hidden patterns in the dataset. The typical architecture of artificial neural networks includes input, output, and hidden layers. The neural network receives external data for analysis or training at the input layer.

The data then passes via one or more hidden layers, which convert the input into knowledge useful to the output layer. The Artificial Neural Networks' reaction to the supplied input data is finally displayed in the output layer.

In most neural networks, units are connected between layers. The degree to which one unit influences the others is shown by the weights given to each of these interactions. As the neural network progresses from one unit to the next, it learns more and more about the data, leading to the output layer's output.

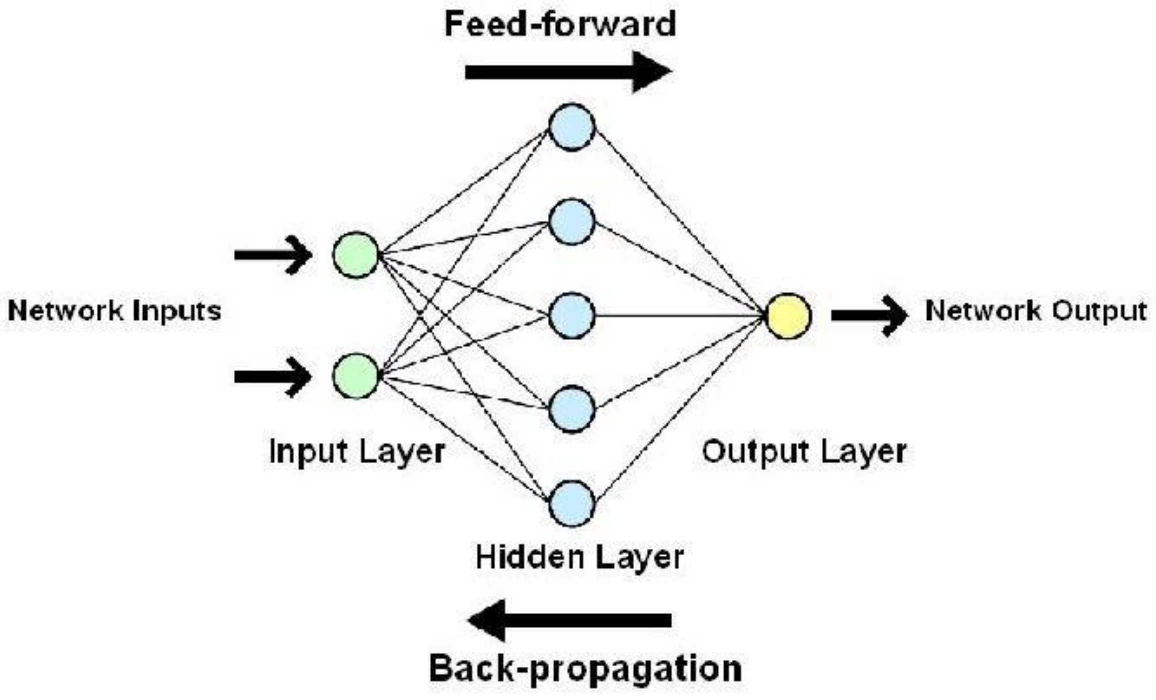


Figure 03: Artificial neural network

4.6 SYSTEM ARCHITECTURE:

The suggested work's flow diagram procedure is depicted in the provided figure. The UCI website served as the first source for the Kaggle Heart Disease Database. After that, the data was pre-processed to remove unnecessary elements. The information was then pre-processed and select important features.

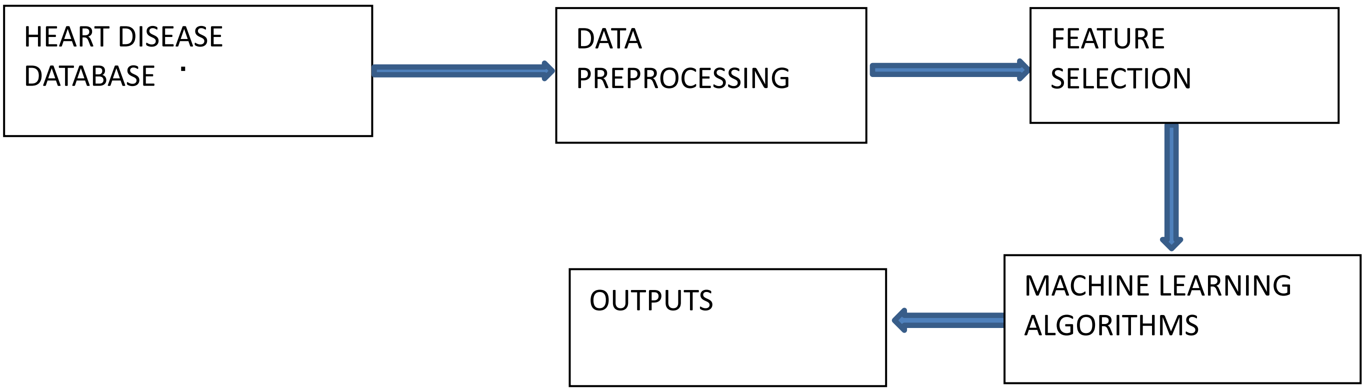


Figure 04: System Architecture

4.7 MODULES:

This project's whole effort is broken down into four components.

- Data pre-processing,
- Features,
- Classification, and
- Prediction are among them.

a. Data Pre-processing:

The file contains all the pre-processing routines needed to handle all incoming texts and documents. After reading the train, test, and validation data files, we performed tokenizing, stemming, and other preprocessing tasks. A few types of exploratory data analysis are performed, including quality checks for null or missing values and response variable distribution.

The process of transforming unprocessed data into an interpretable format is known as data preparation. Since raw data is unusable for data mining, it is also a crucial stage.

It is important to verify the quality of the data prior to utilising machine learning or data mining techniques.

The primary purpose of preprocessing data is to ensure its quality. The following can be used to verify the quality:~

- Accuracy: Assessing the correctness of entered data.
- Completeness: Verifying whether information has been recorded or is missing.
- Consistency: Ensuring uniformity by checking if the same information is consistently stored across all relevant locations.
- Timeliness: Requiring timely and accurate updates to the data.
- Believability: Emphasizing the importance of reliable and trustworthy information.
- Interpretability: Evaluating the ease with which facts can be comprehended and understood.

b. Feature:

We used sci-kit learn Python modules to do feature extraction and selection in this file. We perform the encoding of the categorical values which separates the categorical value from numerical value. Then data feature scaling is done so that the larger value converts into smaller value and is used for prediction with smaller value.

c. Classification:

To analyse the collected features, a number of classifiers from the sklearn package were used, including Random Forest, Naive Bayes, Linear SVM, Logistic Regression, and Stochastic Gradient Descent classifiers. After fitting the models, the confusion matrices were looked at and the F1 scores were compared. Each set of characteristics was applied to each classifier separately. After the assessment, two of the best-performing models from the pool of fitted classifiers were chosen as potential candidates for classifying heart diseases.

By using Grid Search CV techniques to these candidate models, we have optimized the classifiers' parameters and selected the best-performing values. The model that was ultimately chosen was utilized to diagnose heart illness with a probability of true. We have also utilised the top target attributes from our term-frequency data to ascertain which terms are most important in each class. Furthermore, we have used Precision-Recall and learning curves to analyse how training and test sets fare as we increase the amount of data in our classifiers

d. Prediction:

The algorithm that we ultimately chose and found to be the best performer was saved to disc. The heart disease prediction.py file will use this model to categorise cardiac conditions after you shut this repository. The model will also be transferred to the user's computer. The user inputs a news story, and the model is utilised to produce the final categorization output, which is displayed to the user along with the likelihood that it is true.

4.8 DATA FLOW DIAGRAM:

A conventional graphic depiction of the information flows inside a system is called a data flow diagram (DFD). A tidy and easily readable DFD can graphically depict the appropriate level of system requirements. It is possible to execute manually, automatically, or simultaneously.

It displays the locations of data entry and departure destinations, data storage facilities, and data modification procedures. Data flow diagrams (DFDs) are among the most crucial tools used in system analysis methodologies. The many symbols that make up data flow diagrams stand in for different system components. Four types of symbols are used in most data flow modeling techniques: processes, data storage, data flows, and external entities. There are four types of system components that are represented by these symbols. In DFD, circles stand for processes. The DFD shows data flow as a thin line, with a distinct name for each data storage and a square or rectangle symbolizing external entities.

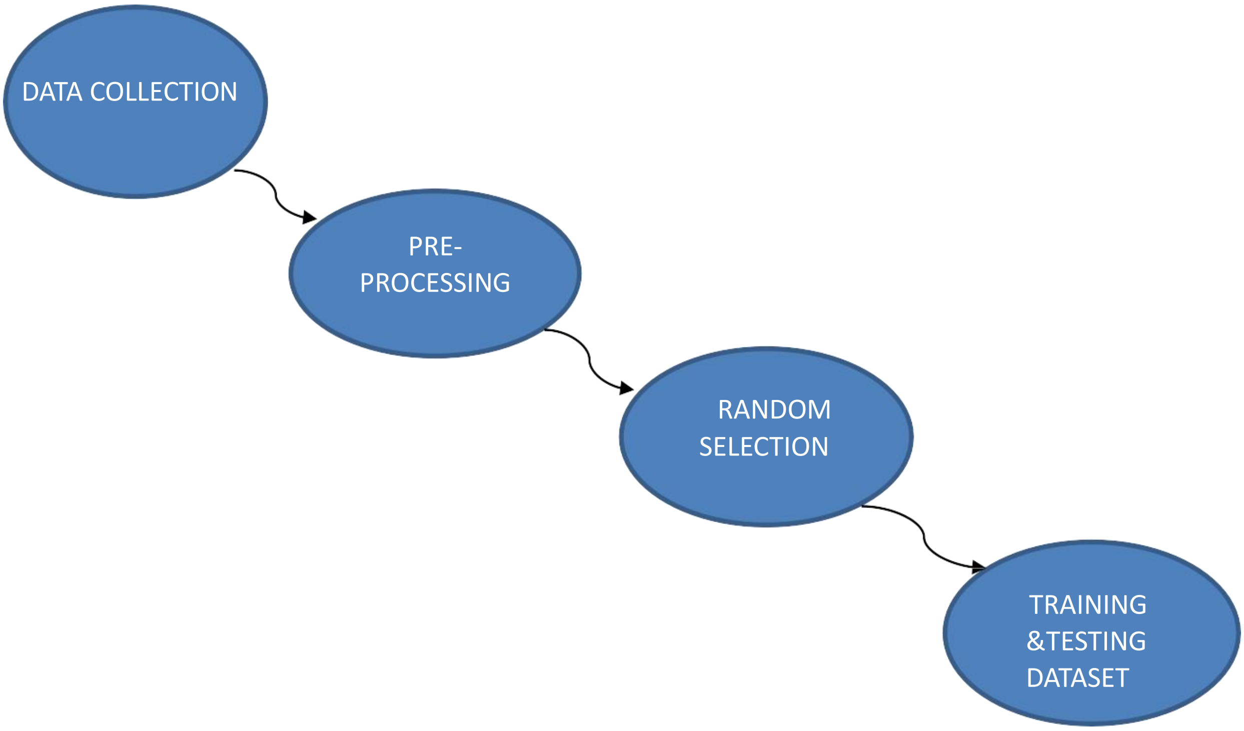


Figure 05: Data Flow Diagram- Level 0

CHAPTER 5
RESULT AND DISCUSSION

After performing the machine learning techniques and approaches for testing and training we found the accuracy of ANN (Artificial Neural Network) is the much effective and efficient approach as compared to the logistic regression and support vector machine. We performed the simulation by the process of data processing and training the model using the help of the python and predicted the accuracy of each model. From the Three models we found the ANN as the best among with the accuracy rate of 89%. The accuracy rate of each model is shown in the table below.

SN. No.	Algorithms	Accuracy
1.	Logistic Regression	77%
2.	Support Vector Machine	80%
3.	Artificial Neural Network	89%

Tab 02: Accuracy comparison

As compared to the reference papers we have taken the accuracy rate of Support Vector Machine (SVM) was higher [1]. But they have concluded the KNN as the best model according to their accuracy [1]. As compared to the other reference paper we have selected, the accuracy of Artificial Neural Network (ANN) is higher according to our model [4]. All the research papers we have selected as reference have discussed the best model according to their model [1][2][3][4]. If we discuss in that context, the Artificial Neural Network is the best according to our model.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Heart Diseases is the major problem according to the present time and has been the leading cause of the death. The prediction about heart diseases is important in the field of science and technology as it is also a major concern of human beings so that the accuracy prediction is one of the way to identify and analyse the performance of algorithms so that it is generalized for detecting heart failure. The development and implementation of a heart disease prediction model is a critical first step toward proactive healthcare and preventative measures. Through the integration of machine learning techniques and the analysis of several datasets, we have developed a predictive model that can identify an individual's risk of heart disease. The logistic regression, support vector machine and artificial neural network are the applied machine learning technique for comparison of accuracy. The proposed ANN method achieved the 89% accuracy and the other methods accuracy rate are listed in the TABLE. 02.

For future work more machine learning method and module will be used and the accuracy will be analysed. The module will be trained using the text data as well as the image data so called hybrid module. The module will be compared by the parameters of accuracy, sensitivity and specificity. The different health parameters related to heart diseases will be studied.

REFERENCES:

[1] Singh, Archana, and Rakesh Kumar. "Improving the Accuracy and Comparing ML Techniques in Heart Disease Prediction." Computer Science and Engineering, Madan Mohan Malaviya University of Technology, Gorakhapur. In 2020 International Conference on Electrical and Electronics Engineering (ICE3), 14-15 February 2020. IEEE, 2020. DOI: 10.1109/ICE348803.2020.9122958.

[2] Qadri, Azam Mehmood; Raza, Ali; Munir, Kashif; Almutairi, Mubarak S. " Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning" IEEE Access, vol. 11, pp. 56214-56224, 30 May 2023. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3281484.

[3] Mohan, Senthilkumar; Thirumalai, Chandrasegar; Srivastava, Gautam. " Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques." IEEE Access, vol. 7, pp. 81542-81554, 19 June 2019. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2923707.

[4] Li, Jian Ping; Haq, Amin Ul; Din, Salah Ud; Khan, Jalaluddin; Khan, Asif; Saboor, Abdus. "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare." IEEE Access, vol. 8, pp. 107562-107582, 09 June 2020. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3001149.

[5] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, “Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools” International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.

[6] Santhana Krishnan J and Geetha S, “Prediction of Heart Disease using Machine Learning Algorithms” ICIICT, 2019

[7] Patel, J., Upadhyay, P. and Patel, “Heart Disease Prediction Using Machine learning and Data Mining Technique” Journals of Computer Science & Electronics , 2016