

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANA SANGAMA, BELAGAVI-590014



A Project Report on

“Heart Disease Prediction”

Submitted in partial fulfillment of the requirement for the award of the degree of

**Bachelor of Engineering in
Information Science and Engineering**

Submitted By

MOULYA G	1DT19IS079
NAYANA SAGAR	1DT19IS086
PRAKRUTHI HR	1DT19IS095
PRIYANSHU SINGH	1DT19IS101

Under the guidance of

Dr.Thirukrishna JT,
Associate Professor,
Dept. of Information Science and Engineering,
DSATM, Bangalore.



DAYANANDA SAGAR ACADEMY of TECHNOLOGY & MANAGEMENT

Udayapura, Kanakapura Road, Opp: Art of Living, Bangalore – 560082

Affiliated to VTU, Belagaavi and Approved by AICTE, New Delhi
2022-23

3 years Accredited by NBA, New Delhi, Validity: 01/07/2022 to 30-06-2025

DAYANANDA SAGAR ACADEMY of TECHNOLOGY & MANAGEMENT

Udayapura, Kanakapura Road, Opp: Art of Living, Bangalore – 560082

Affiliated to VTU, Belagavi and Approved by AICTE, New Delhi

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

3 years Accredited by NBA, New Delhi, Validity: 01/07/2022 to 30-06-2025

CERTIFICATE



This is to certify that the project report entitled “**Heart Disease Prediction** ” is a bonafide work carried out by **Moulya G(1DT19IS079)**, **Nayana Sagar(1DT19IS086)**, **Prakruthi HR(1DT19IS095)** and **Priyanshu Singh(1DT19IS101)** in the partial fulfillment of the requirement for the award of degree in **Bachelor of Engineering in Information Science and Engineering** in college name for Visvesvaraya Technological University,Belagavi,forthe year 2022-2023.It is certified that all corrections/suggestions indicated for the internal assessment have been incorporated in the report. This report has been approved as it satisfies the academic requirements in respect of project work prescribed for Bachelor of Engineering Degree.

Dr.Thirukrishna JT,
Associate Professor,
Dept of ISE,DSATM

Dr.Nandani Prasad KS,
Dean Foreign Affairs & HOD-ISE,
DSATM

Dr.M Ravishankar,
Principal,
DSATM

DAYANANDA SAGAR ACADEMY OF TECHNOLOGY and MANAGEMENT

Udayapura, Kanakapura Road, Opp: Art of Living, Bangalore – 560082

Affiliated to VTU, Belagavi and Approved by AICTE, New Delhi

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

3 years Accredited by NBA, New Delhi, Validity: 01/07/2022 to 30-06-2025



VISION OF THE INSTITUTE

To strive at creating the institution a centre of highest calibre of learning, so as to create an overall intellectual atmosphere with each deriving strength from the other to be the best of engineers, scientists with management & design skills.

MISSION OF THE INSTITUTE

- To serve its region, state, the nation and globally by preparing students to make meaningful contributions in an increasing complex global society challenge.
- To encourage, reflection on and evaluation of emerging needs and priorities with state of art infrastructure at institution.
- To support research and services establishing enhancements in technical, economic, human and cultural development.
- To establish inter disciplinary centre of excellence, supporting/ promoting student's implementation.
- To increase the number of Doctorate holders to promote research culture on campus.
- To establish IIPC, IPR, EDC, innovation cells with functional MOU's supporting student's quality growth.

DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT

Udayapura, Kanakapura Road, Opp: Art of Living, Bangalore – 560082

Affiliated to VTU, Belagavi and Approved by AICTE, New Delhi

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

3 years Accredited by NBA, New Delhi, Validity: 01/07/2022 to 30-06-2025



VISION OF THE DEPARTMENT

Impart magnificent learning atmosphere establishing innovative practices among the students aiming to strengthen their software application knowledge and technical skills.

MISSION OF THE DEPARTMENT

M1: To deliver quality technical training on software application domain.

M2: To nurture team work in order to transform individual as responsible leader and entrepreneur for future trends.

M3: To inculcate research practices in teaching thus ensuring research blend among students.

M4: To ensure more doctorates in the department, aiming at professional strength.

M5: To inculcate the core information science engineering practices with hardware blend by providing advanced laboratories.

M6: To establish innovative labs, start-ups and patent culture.

ABSTRACT

One of the leading causes of mortality is cardiovascular disease. Since the health system produces a significant amount of data, detecting cardiovascular problems is becoming even more crucial. Internet - of - things healthcare systems provide a tough challenge. Machine learning is essential for making correct disease predictions. There has been extensive work in this domain, yet they have not effectively grasped the real potential of machine learning strategies in predicting risk in patients because they have not utilized large quantities of information. In this study, we suggest a unique method for enhancing the accuracy of coronary heart disease diagnosis by identifying significant features using machine learning strategies. Various feature groupings and many well-known classifications techniques have been employed to develop the prediction system. The main goals of the planned study are to improve feature selection and minimize the number of traits while producing improved outcomes. In this work, a better search optimization algorithm with a conceptual methodology is applied to recognize defining factors of cardiovascular diseases.

The prenatal recognition of CVDs can help high-risk patients choose whether to change the way they live, , which can lessen their severity. Using consistent techniques of machine learning, research has sought to identify the most significant risk variables for heart disease as well as effectively estimate the total risk. In order to produce an accurate predictive algorithm for heart disease, the latest research has looked at bringing together these methods using techniques like mixed machine learning (ml) algorithms. These findings suggest a model to evaluate the accuracy of applying individual findings of logistic regression, decision tree algorithms, k-near neighbor, and SVM on the Cleveland Heart Disease Database.

The proposed technique can also be immediately put into practice in the medical world to detect heart disease.

ACKNOWLEDGEMENT

We would like to take this opportunity to express our sincere thanks and gratitude to all those who have been kind enough to guide when needed which has led to the successful completion of the project.

We would like to express our special thanks and gratitude to **Management** of Dayananda Sagar Academy of Technology and Management for providing all the required facility.

We would like to convey our immense gratitude to **Dr. Ravishankar M**, Principal, Dayananda Sagar Academy of Technology and Management and **Dr.Nandini Prasad K S**, Dean-Foreign Affairs & HOD, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore, for their continuous support and encouragement which enabled us to come up with this project and also thank them for providing the right ambience for carrying out the same.

We would like to express our profound gratitude to our Project Guide- **Dr.Thirukrishna JT**, Associate Professor, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore for their continuous support and encouragement.

We extend our sincere gratitude to the project coordinators, **Dr.Manjula G**, Associate Professor and **Mrs.Kavyashree G M**, Assistant Professor, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore, for their guidance and suggestions for successful completion of our project.

We would like to thank our parents and friends who helped us in finalizing this project within the stipulated time frame.

MOULYA G 1DT19IS079

NAYANA SAGAR 1DT19IS086

PRAKRUTHI HR 1DT19IS095

PRIYANSHU SINGH 1DT19IS101

CONTENTS

Chapter	Description	Page No.
1	Introduction	1-3
	1.1 Overview	1
	1.2 Existing system and drawbacks	2
	1.3 Problem Statement	2
	1.4 Proposed System	3
	1.5 Advantages	3
2	Literature Survey	4-12
3	System Requirements Specification	13
	3.1 Hardware requirements	13
	3.2 Software requirements	13
4	System Design	14-18
	4.1 High Level Design	14
	4.1.1 Module Classification	15
	4.1.2 System Architecture	15-16
	4.1.3 Dataflow Diagram	17
	4.2 Detailed Design	18
5	Coding	19-22
6	System Testing	23-25
	6.1 Introduction	23
	6.2 Unit Testing	23
	6.3 Integration Testing	23
	6.4 System Testing	23
	6.5 Output Testing	25
7	Result and Discussion	26-33
8	Conclusion	34
9	Published Journals and certificates	35-54
	References	55

LIST OF TABLES

Serial number	Table Number	Title	Page No.
1	2.1	Literature survey summary	4
2	3.1.1	Hardware requirements	13

LIST OF FIGURES

Serial number	Figure Number	Title	Page No.
1	4.1.2.1	System architecture	17
2	4.1.3.1	Data flow diagram	18
3	5.1	Modules imported	21
4	5.2	Data frames	21
5	5.3	Data processing codes	22
6	5.4	Checking with median	23
7	5.5	Slope	23
8-13	7.1-7.7	Outputs	28-34

CHAPTER 1

INTRODUCTION

1.1 Overview

Heart attacks are to blame for 80% of all fatalities in the country, according to the World Health Organisation (WHO). Each person has a unique heart rate and blood pressure, which vary from 120/80 to 140/90 for blood pressure and 60 to 100 beats per minute for pulse rates.

Diagnosing heart disease is difficult due to the many influencing risk factors such as diabetes, high blood pressure, high cholesterol, irregular heart rate and many other variables. The severity of heart disease in humans has been determined using various data mining and neural network methods. Various methods are used to classify disease severity, including K-nearest neighbor (KNN), decision trees (DT), genetic algorithm (GA), and Naive Bayes (NB). Since heart disease is complex in nature, it requires careful treatment. Failure to do so can damage the heart or lead to premature death. Data mining and a medical research perspective are used to identify different metabolic syndromes. Both cardiovascular disease prediction and data analysis can greatly benefit from data mining and classification.

1.2 Existing System And Drawbacks

Existing System:

Anticipating heart illness is one of the foremost troublesome challenges in pharmaceutical nowadays. Nowadays, around one individual kicks the bucket of heart illness each miniature. Preparing tremendous sums of information in healthcare requires information science. Since predicting heart illness could be a complex errand, the method must be computerized to decrease the dangers and caution the understanding in time. This consider employments the UCI Chord Information Machine Learning Database. The proposed think about employments different information mining procedures counting Gullible Bayes, Choice Trees, Calculated Relapse and Irregular Timberland to anticipate the likelihood of heart infection and classify risk levels of patients. To compare the execution of distinctive machine learning calculations, typically exhausted this paper. Exploratory comes about appear that compared to other utilized ML calculations, the Irregular Timberland strategy has the leading exactness (90.16%).

Drawbacks:

The results can't be correct everytime.

1.3 Problem Statement

The primary goal is to create a heart prediction system by locating and extracting disease-related latent information from a historical heart data collection.

To aid in the prediction of heart disorders, the heart disease prediction system tries to utilise methods on a collection of medical data.

It will lessen the need for medical intervention and make it more affordable.

By making early forecasts, it will also assist in reducing abrupt fatalities.

1.4 Proposed System

An "Online Election Voting" is a website application through which a administrator from the Election Commission monitors and has the control over the database, has the authority to create election environment. The user (who is voting) signs into the website using his/her

Aadhar number. An OTP is sent to the mobile number that is linked to the Aadhar number. This OTP is entered as the password by the user and logs into the website. On logging, the user is directed to the page where the details about the candidates are displayed. The user has the choice to vote to the candidate by activating the radio button that is provided along with the candidate details. On clicking the radio button and submit button, the user votes to the candidate and the database is updated. Further when the users logs in again, they can only view the history of their voting but cannot vote again.

1.5 Advantages

With boosting algorithms, the prediction with ML models for identifying heart attack symptoms is very effective. The prediction was made to assess the forecast's area under the curve, recall, accuracy, and precision. To provide the best forecasts possible, ML models are being trained.

CHAPTER 2

LITERATURE SURVEY

AUTHOR	PURPOSE	TECHNIQUE USED	ACCU RACY
1.Gudadhe et al.	Created a diagnosis system for HD diagnosis utilizing multi-layer Recurrent neural network and support vector machine (SVM)	<ul style="list-style-type: none"> •SVM algorithm •Neural networks 	80.41%
2. Palaniappan et al	A system for professional medical diagnosis was proposed for heart disease identification using ANN,NB and DT	<ul style="list-style-type: none"> •Artificial Neural Networks •Decision Trees (DT) •Navies Bays (NB). 	<ul style="list-style-type: none"> •88.12% •86.12% •80.4%
3. S. U. D. J. K. A. KHAN and A. SABOOR	The authors suggested a NN-based prediction of coronary heart disease (CHD) analysis (NN-FCA) based on feature correlation. The KNHANES-VI dataset created by the Korean Centre for Disease Control and Prevention was used in this study.	<ul style="list-style-type: none"> •Fast Conditional Mutual Information method for selecting features (FCMIM) •FCMIM-SVM 	92.37%
4. Waqar et al.	Suggested using deep learning based on SMOTE. Without feature selection, the author balanced the dataset using the SMOTE technique. A deep neural network was trained and tested to predict the absence and presence of a cardiac arrest using the balanced dataset,	Deep neural network	96%
5. Fitriyani et al	developed a methodology for HD prediction that combines hybrid synthetic minority over-sampling technique-edited nearest neighbor (SMOTE-ENN) and density-based spatial clustering of applications with noise (DBSCAN).	DBSCAN SMOTE-ENN XG BOOST CLASSIFIER	95.9%
6. MOHAN et al.	Developed hybrid machine learning strategy for HD detection. He also put forth a novel methodology for choosing important characteristics from the information for machine learning classifiers to use in training and testing.	HYBRID ALGORITHM	88.07%

Table 2.1 showing literature survey summary

Published paper

An extensive analysis of data mining and machine learning strategies for heart disease prediction

Prakruthi HR [1] , Nayana Sagar [2] ,Moulya G [3] , Priyanshu Singh [4] Dr. ThiruKrishna JT [5]

prakruthihr04@gmail.com[1] , nayanas.1dt19is086@gmail.com[2] , moulyag.1dt19is079@gmail.com[3] , singhpriyanshu073@gmail.com[4] , drthirukrishna@dsatm.edu.in [5]

¹²³⁴UG Scholars, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Karnataka, India

⁵Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Karnataka, India

ABSTRACT

One of the leading causes of mortality is cardiovascular disease. Since the health system produces a significant amount of data, detecting cardiovascular problems is becoming even more crucial. Internet - of - things healthcare systems provide a tough challenge. Machine learning is essential for making correct disease predictions. There has been extensive work in this domain, yet they have not effectively grasped the real potential of machine learning strategies in predicting risk in patients because they have not utilized large quantities of information. In this study, we suggest a unique method for enhancing the accuracy of coronary heart disease diagnosis by identifying significant features using machine learning strategies. Various feature groupings and many well-known classifications techniques have been employed to develop the prediction system. The main goals of the planned study are to improve feature selection and minimize the number of traits while producing improved outcomes. In this work, a better search optimization algorithm with a conceptual methodology is applied to recognize defining factors of cardiovascular diseases. The proposed technique can also be immediately put into practice in the medical world to detect heart disease.

Keyword : - Random forest, SVM, Logistic Regression etc.

1. INTRODUCTION

According to the World Health Organization (WHO), heart attacks account for 80 percent of the overall fatalities nationwide .Each person has a varied blood pressure and heart rate, which range between 60 to 100 beats per minute for pulse rates and 120/80 to 140/90 for blood pressure. Thrombosis, cardiomyopathy, congestive cardiac failure, arrhythmia, pulmonary disease, sudden cardiac death, valve disease, and congenital heart defects are the numerous forms of cvd. It is usually diagnosed by a doctor after reviewing the individual's medical history, the results of their clinical examination, and any alarming problems. However, it is not possible to identify an individual with HD using the outcomes of this clinical diagnosis. Non-laboratory statistics indicate that a variety of risk factors, such as age, gender, smoking, hypertension, high systolic blood pressure, and high blood pressure treatment, raise the risk and body-mass index, might provide useful information for assessing CVD risk [3]. The majority of these parameters might be viewed as onset indicators and cautions to the person, which can add to the risk score obtained by standard biochemical measures (such as cholesterol values). The adoption of self-assessment questionnaires as a complement to most clinical procedures is due to this. When new modes and medical experts are inaccessible, the diagnosis and treatment of heart disease is really quite challenging. As a consequence, timely identification of heart problems can minimize the number of mortality and enable healthcare professionals to prescribe the most appropriate treatment option. However, a number of unidentified elements even cause expertise in heart disease to wrongly identify the condition. However, it is critical to search for accurate procedures to accurately account for all the uncertain risk factors and detect cardiovascular disease. Scientists have explored a diversity of algorithms using machine learning to determine the best combinations of cardiovascular diseases parameters to aid health care professionals in improving computer - aided diagnosis methods and the quality of treatments.

As it takes expertise and in-depth understanding to anticipate cardiac disease, it is a challenging task. The complexity of the problem is categorized using a variety of techniques, including the K-Nearest Neighbor Algorithm (KNN), Naive Bayes (NB), Decision Trees (DT), and Genetic Algorithm (GA). Recently, the importance of optimization to our everyday life has increased. Population- and evolutionary-based optimization approaches are well-liked and frequently employed in various engineering fields. This growth - based finds the best options out of the numerous available options and provides a setting that is good for problem-solving. A mathematical representation of the system is necessary for the majority of optimization strategies. Making a statistical method for complicated processes might be difficult. The high cost forbids employing the solution time even if the model is established. Due to physical events, it is difficult to create an optimization technique to obtain enough global and local search operators.

Due to the CVDs intricate nature, it must be handled with caution. Failure to follow instructions could increase the risk of death or injury to the heart. Many various types of physiologic illnesses are being revealed according to medical research and machine learning (ML). ML with categorization plays an important role in the detection of HD and data processing. An effective machine learning approach is one that performs well on both seen and hidden samples. This happens because a machine learning approach could just understand the data for training otherwise. Several classifiers were placed through data processing, and it was found that they properly classified 50 percent of the total of the instances on average [16]. Additionally, when a model has been trained and evaluated on a dataset, relevant cross validation techniques and performance evaluation metrics are important.

For testing and training, the machine learning prediction models require adequate data. If balanced datasets are used for model training and testing, machine learning model performance can be improved. Furthermore, by incorporating appropriate and significant elements from the data, the model's prediction skills may be improved. In order to increase performance of the model, data balance and extraction of features are therefore critical. Here, we undertake tests to determine the characteristics of a hybrid machine learning algorithm. The outcomes of the experiment indicate that, in comparison to other methods, hybrid methods have a greater capacity to predict heart disease.

The effectiveness of every classifier in the challenge of classifying cardiovascular disease is examined in this study using four large-scale datasets, including the Cleveland heart disease dataset, Cardiovascular dataset, Framingham cardiovascular disease dataset and Cardio train1 dataset. According to experimental results, this gentle group always does well when compared to certain other classifiers, as indicated by a higher measure, particularly with large datasets like the Cardiovascular and the Cardio-Train1 datasets. The other sections of this paper are organized in the following way: Section II presents recent work in the field; Section III describes the methodology technique; Section IV presents the various Algorithms; and Section V reviews and summarizes the paper.

2. RELATED WORK

With the introduction of machine learning technologies, numerous research has been dedicated to identifying heart disease issues. A substantial amount of information has been generated by wearable devices and mobile healthcare systems, which has detailed exploration to gather the health information they need to predict cardiovascular disease. In recent years, numerous studies have been carried out to categorize heart disease with great accuracy using numerous classification algorithms, mostly on the publicly accessible Cleveland dataset. The global evolutionary method and the features selection procedure both were applied to the Cleveland dataset. With the ten most important features chosen by SVM-RFE (Recursive Feature Elimination) and gain ratio methods, Naive Bayes obtains an efficiency of 84.1584%. On the Cleveland sample, the Naive Bayes classification procedure is carried out.

The accuracy of the algorithm for decision trees is the lowest when applying the 10-cross validation technique, coming in at 77.55% when all 13 of the dataset's attributes are utilized. KNN comes in second with an accuracy of 83.16percent of total when $k = 9$. Nevertheless, the accuracy of the decision tree and SVM with boosting is greater, at 82.17% and 84.81%, respectively. The decision tree technique fared poorly with an accuracy of 42.89% compared to the SVM classifier's accuracy of 85.7655%. SVM achieves an f-measure value of 93.5617%.

In different research, Gudadhe et al. [22] created a diagnosis system for HD diagnosis utilizing multi-layer Recurrent neural network and support vector machine (SVM) techniques and achieved accuracy of 80.41%. By combining a neural network with fuzzy logic, Humar et al. [] developed the HD recognition system. A technique for diagnosing heart disease based on ML was created by Akil et al. The ANN-DBP algorithm and FS algorithm both performed well. A system for professional medical diagnosis for HD identification was proposed by Palaniappan et al. Artificial Neural Networks (ANN), Decision Trees (DT), and Navies Bays (NB) were used as predictive machine learning models during the development of the system. NB attained 86.12% efficiency, ANN 88.12% accuracy, and DT classifier 80.4% accuracy.

In Another research by MOHAN et al. [27] developed a hybrid machine learning strategy for HD detection. He also put forth a novel methodology for choosing important characteristics from the information for machine learning classifiers to use in training and testing. They have an 88.07% classification accuracy rate.

A balancing strategy was established in a small number of research to support decision systems that tackled the mentioned issue. To identify and eliminate outliers and equalize distribution of the data, Fitriyani et al. devised an HD prediction method that uses density-based spatial clustering of applications with noise (DBSCAN) and hybrid synthetic minority over-sampling technique-edited nearest neighbor (SMOTE-ENN). The XGBoost classifier also predicts the patient's status using which an accuracy of 95.9% was achieved using the proposed model [16]. To predict cardiac attacks, Waqar et al. suggested using deep learning based on SMOTE. Without feature selection, the author balanced the dataset using the SMOTE technique. A deep neural network was trained and tested to predict the absence and presence of a cardiac arrest using the balanced dataset, and it obtained 96% efficiency.

So in order to cluster relevant healthcare data in the cloud, propose a cloud-based K-means Clustering employed as a MapReduce task. Using an adaptive boosting approach, the authors proposed an ensemble learning classification algorithm. 4 distinct heart disease datasets from the Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology (HIC), Long Beach Medical Center (LBMC), and Switzerland University Hospital were used to test this model (SUH). The same factors are taken into consideration as heart disease causes in all four datasets. The generated model outperformed the accuracy of earlier study by achieving accuracies of 80.14% for CCF, 89.12% for HIC, 77.78% for LBMC, and 96.72% for SUH. For a better understanding of the significance of our suggested methodology, Table 1 summarizes the drawbacks and advantages of the HD detection methodologies that have been presented in the abovementioned literature. To detect HD in its earliest phases, all of these systems in use today employed a variety of techniques. All of these methods, however, have poor predictive performance and take a long time to compute. Table 1 shows that more improvements are needed to the HD detection method's prediction accuracy in order to detect HD effectively and accurately at an early stage, which would lead to better treatment and recovery. Therefore, the main problems with these earlier methods are their poor accuracy and prolonged computation times, which may be caused by the introduction of unnecessary features in the dataset. To address these issues, new HD detection methods are required.

Using feature correlation, the authors of [18] proposed a NN-based prediction of coronary heart disease (CHD) analysis (NN-FCA) (NN-FCA). This research provided use of the KNHANES-VI dataset produced by the Korean Center for Disease Control and Prevention. For CHD prediction, the NNFCFA method incorporates feature correlation analysis and produces a superior ROC Curve (0.7490.010) than the Framingham Risk Score (FRC) (0.3930.010). The features extraction issue is resolved in [19] using a Fast Conditional Mutual Information method for selecting features (FCMIM). The Cleveland heart disease dataset is used to evaluate the FCMIM-based technologies. FCMIM-SVM is more successful than other techniques, such as NB-based HD diagnostic techniques (86.12%), three-phase ANN detection system (88.89%), and the Neural Network Ensemble (89.01%), with a 92.37% accuracy rate.

Technique	Limitation	Advantages	Acc(%)
HD diagnosis using ML classifiers	The Proposed method accuracy is very low.	Computationally less complex.	77
MLP+SVM	Computationally complex.	The performance of the proposed method is high in terms of prediction accuracy.	80.41
ANN-Fuzzy Logic	More execution time required to generate results.	Accuracy is high.	87.4
ANN ensemble based diagnosis system	Computationally complex.	High accuracy.	89.81
HD diagnosis system based on NB, DT and ANN	The NB and DT performance are low.	ANN achieved high performance in terms of accuracy.	88.12
Three phase technique based on ANN	High computation time.	High accuracy.	88.89
ANN-FUZZY-AHP	Computationally complex.	Achieved high accuracy.	91.1
Relief-Rough set based method for HD detection	Computation time is high.	High accuracy due to selection of appropriate feature for training and testing of the model.	92.32
Hybrid ML method	Low accuracy.	Low computation time.	88.87

Fig 1: Various technique with their accuracy score

3. METHODOLOGICAL FRAMEWORK

3.1 Collection of Dataset

The dataset is referred as group of connected data which contains data for each instance. An attribute in the dataset contributes to the factor that determines the outcome. Each attributes contributes a certain level to the final outcome but it is not sure that all the attributes have same level of control over the outcome. The dataset has been obtained from international universities such as University of California Irvine (UCI) 2016–2022, which is recorded from real time observation of the patients suffering from Cardiovascular Disease. The original dataset contains 13 attributes, 270 subjects and an output class. Each and every property present in UCI dataset play an important role in heart disease prediction. Based on the various physical examination and laboratory tests these datasets has been accumulated. Based on the survey conducted by the UCI is used to find risky factors of the disease. If a person is classified under the category to be tested further An individual is classified as 'needs further tests then there are many factors due to which the person has to take further tests some of them can be lack of physical fitness, obesity, diabetes, high blood pressure. A person who is classified as non-healthy are the ones who have already had heart-attack or have prolonged chest pain. According to reports that are made on a daily basis the chances of having heart related problems are diabetes, high blood pressure, and they can be facing some symptoms such as chest pain or chest burn and shortness in breath, The other people who did not experience any of the symptoms are classified under the healthy category.

3.2 preliminary processing of dataset

Pre-processing can be defined as a process that is used to convert raw data into useful format. The major step of preprocessing is formatting the data. Normally, the data we obtain in the raw format contains a lot of missing values, wrong representations, so this data cannot be used for the machine learning models directly and it has to be cleaned up and made suitable for machine learning models. dataset plays a major role in creating machine learning models. The main steps of preprocessing models are data cleaning, data integration, data transformation, data reduction.

3.2.1 Data Cleaning

This technique is used to remove the missing values, Noisy data and the inconsistency in the data points. The result of this is to get an accurate output for machine learning models. The problem of missing value occurs when one or more dataset are used to form larger dataset, the most easier way to resolve the issue is to delete those fields before merging. There is one more technique to fill out the missed values with most probable values and this can be done using logistic regression.

3.2.2 Data Integration

The data will be collected from different sources and it has to be integrated for the proper usage and during this integration, it may lead to several inconsistencies and redundant data. There are three techniques for integrating the data they are data consolidation, data virtualization, data propagation. In data consolidation all the data physically bought together at one place this helps to increase organization and productivity of data integration. data virtualization explains about the viewpoint of the data. In data propagation with the help of some applications we can transfer the data from one location to another location.

3.2.3 Data Reduction

This technique is used to reduce the quantity of data so it helps to reduce the cost associated with it. When we are working with big data this data preprocessing step plays a major role. This technique helps to create faster and more efficient models.

3.2.4 Data Transformation

This technique is used to convert the data from one pattern to another pattern. Some of the strategies used for data transformation are Smoothing, Aggregation, normalization, generalization. In generalization we will convert the data features from low level to high level, Normalization process will convert all variables within some specified unit. Smoothing is used to remove the noise from the data using some algorithms.

The table gives a brief description about chest pain occurring in various age groups along with sex. Here we find the men in the age group of 50-60 are the ones who suffer more from chest pain. Chest pain can again be in four various forms. Chest pain is the most common symptom.

Disease	Female	Male
Stroke	10%	8%
Hypertensive heart disease	2%	1.1%
Rheumatic heart disease	0.15%	0.13%
Cardiovascular and circulatory disease	8%	15%
Endocarditis	0.14%	0.11%
Cardiovascular disease	26.7%	29.2%

Fig 2: Major Categories under Study.

The dataset from the uci is a multivariate dataset that contains 76 attributes which are related to various blood and body conditions out of which a subset containing 10 attributes are been picked like age, sex, blood pressure, blood glucose level, cholesterol level, electrocardiogram, heart rate, angina induced due to exercise, depression caused due to exercising. The four variations in chest pain are marked as a, b, c, d. They are typical angina, atypical angina, non-anginal pain, asymptomatic respectively.

4. MACHINE LEARNING MODEL/ALGORITHM

4.1 PRISMA algo

Without prospectively registering, researchers followed a process that was agreed upon by all authors and adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) declaration. The objective of this algorithm is to summarize and assess the best reliable machine-learning method for ischemic heart disease prediction. PRISMA criteria were followed in conducting this systematic review. Multiple databases, including Science Direct, PubMed, MEDLINE, CINAHL, and IEEE Explore, were used to conduct a thorough search. The inclusion was open to 13 papers that were released between 2017 and 2021. Three topics emerged: the most popular algorithm for ischemic heart disease prediction, the reliability of ischemic heart disease prediction algorithms, and clinical outcomes to raise the standard of treatment. Both supervised and unsupervised machine learning have been used in all approaches.

4.2 L.S.T.M model

The large-scale patient hospital records are not successfully employed to improve the prediction performance, and previous dynamic prediction models seldom handle multi-period data with variable intervals. Some studies use an enhanced long short-term memory (LSTM) model to examine the prediction of cardiovascular disease.

4.3 S.V.M (Support Vector Machine) algo

One of the machine learning algorithms is called the support vector machine. An algorithm for supervised learning is the support vector machine. The provided data is categorized using the support vector machine. A hyper plane is used by the method to distinguish between the various classes. Regression analysis also makes use of support vector machines. Both linear and non-linear data are classified by SVM. The SVM classifier's primary goal is to locate the hyperplane in an n-dimensional space.

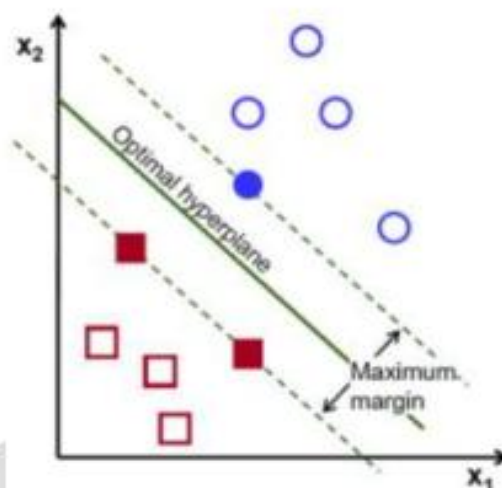


Fig.3: showing S.V. M

4.4 Random Forest

Machine learning algorithms are increasingly being used to forecast different illnesses. This idea is so vital and useful because machine learning algorithms are designed to think like humans. Here, the problem of improving heart disease prediction accuracy is tackled. The Cleveland heart disease dataset's non-linear tendency was taken advantage of when Random Forest was used, yielding an accuracy of 85.81%. We achieve more accuracy when utilizing the Random Forest approach to forecast cardiac illnesses with well-defined features. 303 data instances were used to train Random Forest, and 10-fold cross validation was used to verify correctness. Future lives might be spared by the suggested method for heart disease prediction

4.5 Decision Tree

A typical data mining technique for creating classification and prediction systems based on many explanatory characteristics and creating prediction models for a target instance is the decision tree methodology. This technique divides a population into pieces that resemble branches in a tree that create a root node, internal nodes, and leaf nodes in an inverted tree. A decision tree is a non-parametric technique that can handle large, complex data sets effectively without utilizing several parametric structures. Study data can be split into training and validation data sets if the sample size is big enough. Choose the right tree size to get the best final model by using the training data set to create a decision tree model and the validation data set.

4.6 Logistic Regression

The link between the dependent variable (target), which is categorical data with a nominal or ordinal scale, and the independent variable (predictor), which is categorical data with an interval or ratio scale, is assessed using the predictive model known as logistic regression. To determine the link between the relevant variables, this approach may also be employed in time series modeling. An approach called logistic regression is used to forecast the likelihood of categorical dependent variables.

4.7 ANN (Artificial Neural Network)

ANN algorithm is based on a large number of basic neural units (artificial neurons), which are roughly equivalent to the observed behavior of the axons in a real brain. It is used in computer science and other study areas. Each neuronal unit is interconnected with several others, and these connections can either increase or decrease the level of activity in nearby neural units. The outline function is used to compute for each individual neuronal unit. Each link and the unit itself may have a threshold function or limiting function that requires the signal to exceed before it may reach other neurons. These systems thrive in areas where the solution or feature identification is challenging to describe in a conventional computer programmer because they are self-learning and taught rather than explicitly coded.

4.8 Naive Bayes Algorithm

Data mining is the process of applying a number of approaches to find information or decision-making expertise in a database and extracting it so that it may be used for tasks like decision support, forecasting, estimate, and prediction. The healthcare sector gathers enormous volumes of data, which are regrettably not "mined" to reveal hidden information for wise decision-making. Data mining is the process of identifying relationships between variables in a database. The Decision Support in Heart Disease Prediction System (DSHDPS) established by this study makes use of the Naive Bayes data mining modeling approach. The chance of people developing heart disease may be

predicted using medical profiles including age, sex, blood pressure, and blood sugar. It is implemented as an online survey application. It may be used as a teaching tool to teach nurses and medical students how to diagnose heart disease patients.

S no.	Reference No.	Techniques /Methods used	Accuracy(%)
1	17	ANN	85.53
2	20	Naive Bayes	96.5
3	21	Decision Tree	99.2
4	22	SVM	86.6
5	6	Logistic Regression	83.70
6	9	Random Forest	86.9
7	15	LSTM	92.5
8	16	PRISMA	84.5

5. CONCLUSION

The goal of this study was to determine if patient questionnaires containing historical subjective and examination-based objective health data might be utilised to detect potential risks for heart disease. Such data may support the diagnostic value of physiological-biochemical tests clinically carried out in CVD in addition to screening. SVMs with strict feature selection were taken into account by the evaluation system. The categories of medical condition, cardiovascular health, and fitness have shown good promise in determining the risk of CVD after a number of tests, with the category of fitness demonstrating significant effectiveness.

Researchers have outlined many machine learning techniques for heart disease prediction. They developed a number of machine learning algorithms and then examined their attributes to determine which one was the best. Every algorithm has produced a distinct outcome in a variety of circumstances. Further analysis shows that the prediction model for heart illness only achieves minimal accuracy; hence, more complicated models are required to improve the accuracy of predicting early heart disease. Future methodologies for highly accurate, low-cost, and simple early heart disease prediction will be proposed. The researchers stated many algorithms and the algorithms have problems also. But the best 3 with the accuracies are Decision Tree, Naive Bayes and LSTM.

6. REFERENCES

- [1] WHO, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
- [2] S. Singh, and R. Zeltser, "Cardiac Risk Stratification," in: StatPearls, Treasure Island (FL): StatPearls Publishing, 2020.
- [3] A. Pandya, et al., "A comparative assessment of non-laboratory based versus commonly used laboratory-based cardiovascular disease risk scores in the NHANES III population," PLoS One, vol. 6, no. 5, pp. e20416, May 2011.
- [4] NHANES: <https://www.cdc.gov/nchs/nhanes/index.htm>
- [5] C.Y. Wang, et al., "Cardiorespiratory fitness levels among US adults 20-49 years of age: findings from the 1999-2004 National Health and Nutrition Examination Survey," Am J Epidemiol., vol. 171, no. 4, pp. 426-435, Feb. 2010.
- [6] P.L. Tsou, and C.J. Wu, "Sex-Dimorphic Association of Plasma Fatty Acids with Cardiovascular Fitness in Young and Middle-Aged General Adults: Subsamples from NHANES 2003-2004," Nutrients, vol. 10, no. 10, 1558, Oct. 2018.
- [7] S.S. Yoon, et al., "Trends in the Prevalence of Coronary Heart Disease in the U.S.: National Health and Nutrition Examination Survey, 2001-2012," Am. J. Prev. Med., vol. 51, no. 4, pp. 437-445, Oct. 2016.
- [8] R. Moonesinghe, et al., "Prevalence and Cardiovascular Health Impact of Family History of Premature Heart Disease

in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007-2014," J. Am. Heart Assoc., vol. 8, no. 14, e012364, July 2019.

[9] K. Jindai, et al., "Multimorbidity and Functional Limitations Among Adults 65 or Older, NHANES 2005-2012," Prev. Chronic Dis., vol. 13, 160174, Nov. 2016.

[10] S. Heyden, et al., "Angina Pectoris and the Rose Questionnaire," Arch. Intern. Med., vol. 128, no. 6, pp. 961-964, 1971.

[11] A. Koyanagi, et al., "Correlates of physical activity among community-dwelling adults aged 50 or over in six low- and middle income countries," PLoS ONE, vol. 12, no. 10, e0186992, Oct. 2017. [12] W.-H. Weng, "Machine Learning for Clinical Predictive Analytics," in: Leveraging Data Science for Global Health. L. A. Celi et al. (eds.), 2020, ch. 12.

[13] Support Vector Machines," Machine Learning, vol. 46, pp. 389-422, Jan. 2002. [14] H. Sanz, et al., "SVM-RFE: selection and visualization of the most relevant features through non-linear kernels," BMC Bioinformatics, vol. 19, 432, Nov. 2018.

[15] A. Dinh, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 19, 211, 2019. <https://doi.org/10.1186/s12911-019-0918-5> [16] Tulay Karayilan, Dept of Computer Engineering, Yildirim and Ozkan Kilic, Department of Computer Engineering, "Prediction of heart disease using neural network", IEEE Explorer

[17] J. K. Kim and S. Kang, "Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis," Journal of Healthcare Engineering, vol. 2017, 2017. 23. J. PING LI, A. U. H. [18] S. U. D. J. K. A. KHAN and A. SABOOR, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, 19 June 2020.

[19] M. A. Jabbar, Shirina Samreen, "Heart disease prediction system based on hidden naive bayes classifier", IEEE, 2020

[20] Mai Shouman, Tim Turner, Rob Stocker, School of Engineering and Information Technology, University of New South Wales at Australian Defence Force Academy Canberra ACT 2600, "Using Decision Tree for Diagnosing Heart Disease Patients", 2021

[21] T. Mythili, Dev Mukherji, Nikita Padalia, Abhiram Naidu, International Journal of Computer Applications, "A heart disease prediction model SVM-Decision Trees-Logistic Regression(SDL)", 2013

[22] 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, Jun. 2021.

[23] H. B. Kibria and A. Matin, "The severity prediction of the binary and multi-class cardiovascular disease—A machine learning-based fusion approach," Comput. Biol. Chem., vol. 98, Jun. 2022, Art. no. 107672.

[24] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," IEEE Access, vol. 9, pp. 39707-39716, 2021

CHAPTER 3

SOFTWARE REQUIREMENTS SPECIFICATION

It is a comprehensive document that covers all aspects of your software project, from the project's goals and objectives to the software's specific requirements. It outlines the functional and non-functional requirements of the software and serves as a blueprint for the entire project.

3.1 Hardware Requirements

	Windows requirements	Mac requirements	Linux requirements
Operating system	Windows 7 or later	Mac OS X 10.9.x or later	64-bit Ubuntu 12.04+, Debian 8+, OpenSuSE 12.2++, or Fedora Linux 17
Processor	Intel Pentium 4 or later	Intel	Intel Pentium 3 / Athlon 64 or later
Memory	2 GB minimum, 4 GB recommended		
Screen resolution	1280x1024 or larger		
Application window size	1024x680 or larger		
Internet connection	Required		

Table 3.1.1 showing hardware requirements

3.2 Software Requirements

- Back end: Python
- Front end: HTML, CSS
- Tools used: Jupyter notebook
- Libraries: Bootstrap, JQuery

CHAPTER-4

SYSTEM DESIGN

4.1 High Level Design

The design utilized to construct computer program is portrayed in high-level plan (HLD). An design graph gives a total picture of the framework by highlighting the key components and their intelligent showing up for the item.

An HLD is likely to utilize non-technical or tolerably specialized dialect that supervisors ought to get it. But for engineers, shallow plan still appears the consistent exactness plan of each component.

State-of-the-art plan ought to incorporate all critical changes that got to be made to all stages, frameworks, items, administrations and forms that it depends on. In expansion, all noteworthy commerce, legitimate, natural, security, security and specialized dangers, issues and presumptions ought to be briefly checked on.

4.1.1 Modules Used

Flask module:

Python is utilized to make web applications with Carafe, which is created on Werkzeug and Jinja2. There are benefits to embracing the Carafe system, counting a built-in advancement server and a fast debugger. Lightweight.

Pickle Module:

For the reason of serializing and deserializing a Python question structure, the pickle module underpins parallel conventions. The act of turning a Python protest pecking order into a byte stream is known as "pickling," and the method of turning a byte stream (from a double record or question that looks like it is made of bytes) back into an question chain of command is known as "unpickling." Elective names for pickling (and unpickling) incorporate "serialization," "marshaling," "smoothing," and 1, but for the purposes of this article, "pickling" and "unpickling" will be utilized instep.

Numpy Module:

The Python bundle NumPy is utilized to control clusters. Moreover, it has frameworks, fourier change, and capacities for working within the area of straight polynomial math. Within the year 2005, Travis Oliphant developed NumPy. You'll be able utilize it for complimentary since it is an open source extend. Numerical Python is alluded to as NumPy.

Pandas Module:

Python's open source Pandas library is accessible. It offers high-performance information structures and devices for information examination that are prepared for utilize. The broadly utilized Pandas module for information science and analytics works on best of NumPy.

Matplotlib Module:

Python's Matplotlib bundle gives a total apparatus for building inactive, vivified, and intuitively visualisations. Matplotlib makes troublesome things conceivable and basic things simple. Create plots fit for distributing. Make intelligently charts with zoom, container, and overhaul capabilities.

Seaborn Module:

A Python data visualization package based on Matplotlib is called Seaborn. It provides an advanced plotting tool to create eye-catching and educational statistical images.

4.1.2 System Architecture



Figure(Fig) 4.1.2.1 showing system architecture of Heart Disease(HD) prediction

System architecture, sometimes called system architecture, is a conceptual model that describes the behavior, structure and other aspects of a system. A formal description and representation of a system intended to facilitate the analysis of its structures and behavior is called an architectural description.

System architecture can include extended systems that are designed and work together to implement the entire system.

The architectural description languages (ADLs) collectively refer to efforts to formalise languages that describe system architecture.

4.1.3 Data Flow Diagram

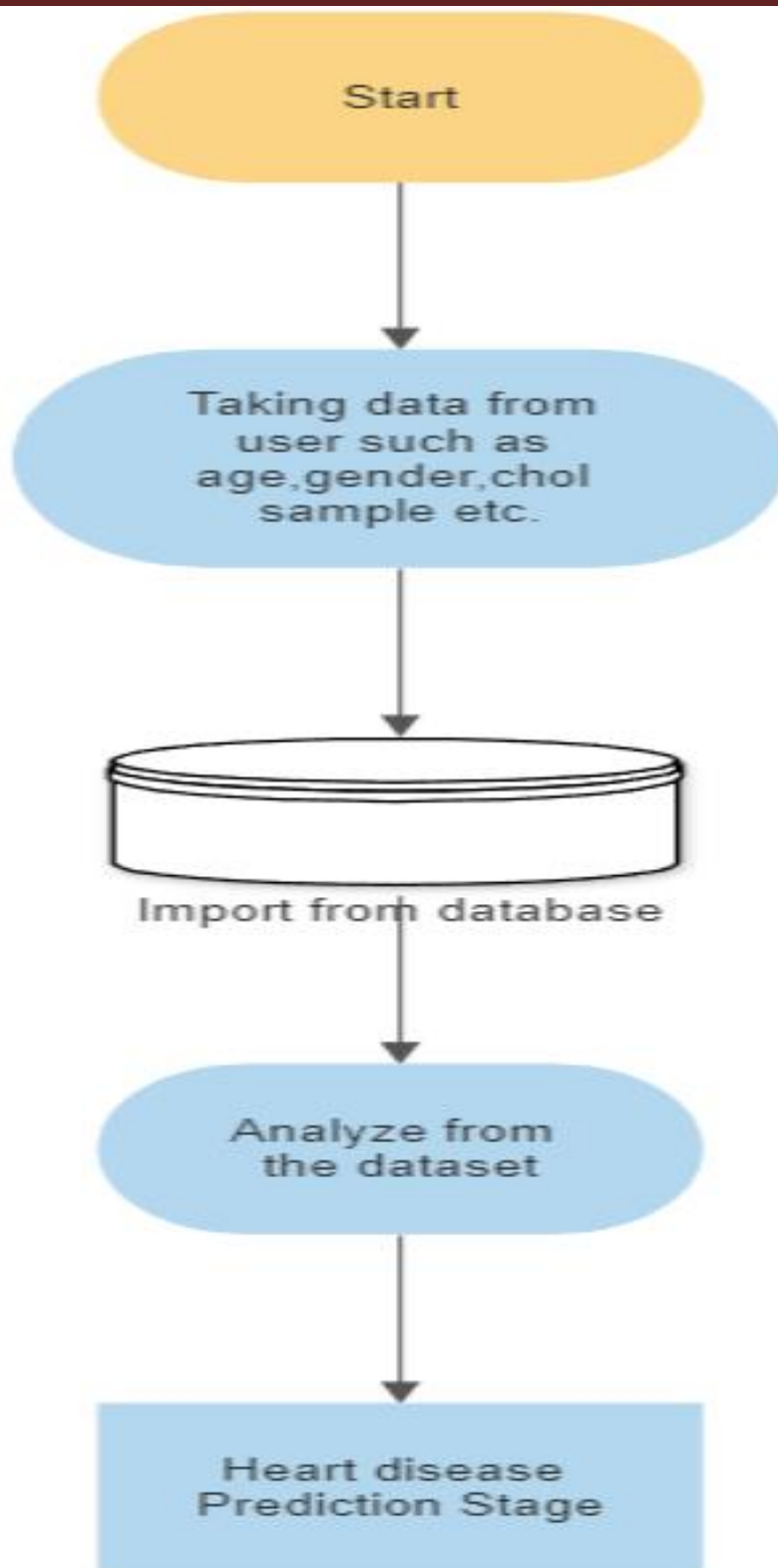


Fig 4.1.3.1 Data flow diagram for HD Prediction

A flowchart is “diagram” that displays the various steps of “process” in “right sequence”. It's a versatile tool that may be utilized for describing a variety of procedures, such as a production process, an administrative or service procedure, or a strategy for a project.

4.2 Detailed Design

- The final design effort before implementation starts is detailed system design.
- The most difficult design issues must be solved by the thorough design; else, the design is incomplete.
- Although compared to source code, the detailed design is still an abstraction, it should be sufficiently detailed to guarantee that the translation to source is a precise mapping rather than a loose interpretation.
- Every perspective in the detailed design should utilise a distinct modelling approach to depict the system design in a range of viewpoints.
- Different views can help to make different aspects of the system more understandable.
- While other views are better at illustrating how data flows within the system, some views are better at elaborating a system's states.
- For systems that are created using an object-oriented methodology, other perspectives do a better job of illustrating how various system items connect to one another through class taxonomies.

CHAPTER 5

CODING

5.1 Coding files

App.py:-

```
from flask import Flask, request, render_template
import pickle
import numpy as np
from sklearn.ensemble import RandomForestClassifier
app = Flask(__name__)
forest = pickle.load(open('heart.pkl', 'rb'))

@app.route('/')
def home():
    return render_template('index.html')

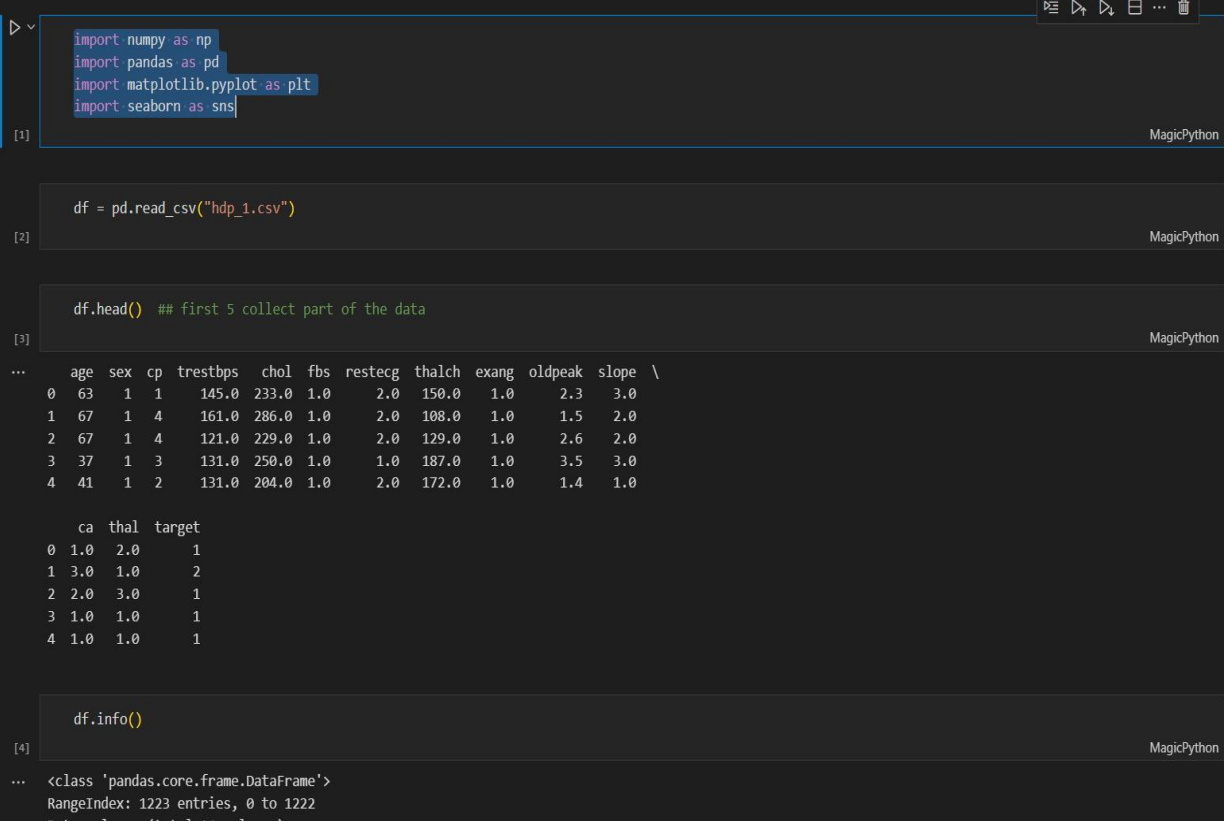
@app.route('/predict' , methods = ['GET', 'POST'])
def predict():

    int_features = [x for x in request.form.values()]
    final_features = [np.array(int_features)]
    prediction = forest.predict(final_features)

    output = prediction[0]

    if output == 0:
        return render_template('index.html', prediction_text= 'No_RISK')
    elif output == 1:
        return render_template('index.html', prediction_text='Stage 1')
    elif output == 2:
        return render_template('index.html', prediction_text='Stage 2')
    elif output == 3:
        return render_template('index.html', prediction_text='Stage 3')
    else:
        return render_template('index.html', prediction_text= 'Final Stage')

if __name__ == "__main__":
    app.run(debug=True)
```

project final.ipynb:-


```

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("hdp_1.csv")

df.head() ## first 5 collect part of the data

age sex cp trestbps chol fbs restecg thalach exang oldpeak slope \
0 63 1 1 145.0 233.0 1.0 2.0 150.0 1.0 2.3 3.0
1 67 1 4 161.0 286.0 1.0 2.0 108.0 1.0 1.5 2.0
2 67 1 4 121.0 229.0 1.0 2.0 129.0 1.0 2.6 2.0
3 37 1 3 131.0 250.0 1.0 1.0 187.0 1.0 3.5 3.0
4 41 1 2 131.0 204.0 1.0 2.0 172.0 1.0 1.4 1.0

ca thal target
0 1.0 2.0 1
1 3.0 1.0 2
2 2.0 3.0 1
3 1.0 1.0 1
4 1.0 1.0 1

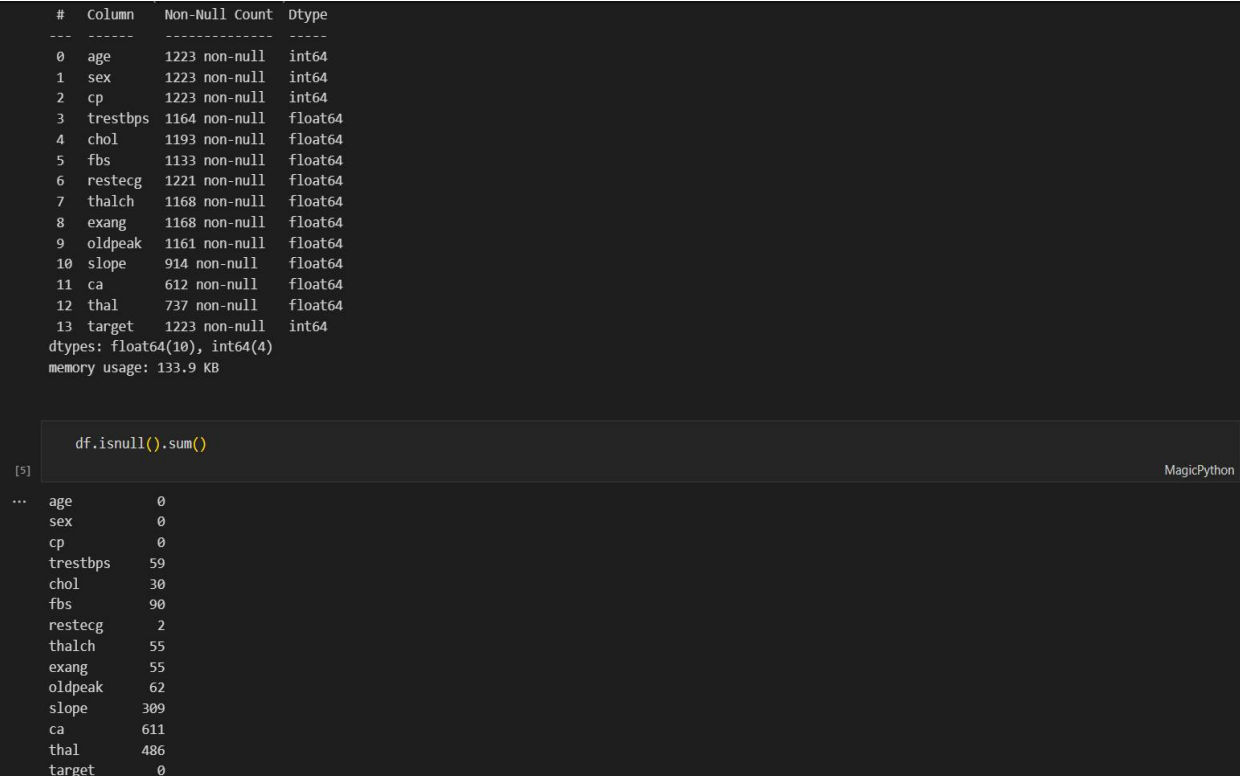
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1223 entries, 0 to 1222
Data columns (total 14 columns):

```

Fig 5.1 showing modules imported

Here we are importing the required for the program to proceed.



```

# Column Non-Null Count Dtype
0 age 1223 non-null int64
1 sex 1223 non-null int64
2 cp 1223 non-null int64
3 trestbps 1164 non-null float64
4 chol 1193 non-null float64
5 fbs 1133 non-null float64
6 restecg 1221 non-null float64
7 thalach 1168 non-null float64
8 exang 1168 non-null float64
9 oldpeak 1161 non-null float64
10 slope 914 non-null float64
11 ca 612 non-null float64
12 thal 737 non-null float64
13 target 1223 non-null int64
dtypes: float64(10), int64(4)
memory usage: 133.9 KB

df.isnull().sum()

age 0
sex 0
cp 0
trestbps 59
chol 30
fbs 90
restecg 2
thalach 55
exang 55
oldpeak 62
slope 309
ca 611
thal 486
target 0

```

Fig 5.2 showing data frame

Here we can see the data frame being used.

```

dtype: int64

Data_preprocessing

df['trestbps'].isnull().mean()

[6]
... 0.0482420278004906

def fun(df):
    mean = df['trestbps'].mean()
    print(df['trestbps'].std())
    df['trestbps_replaced'] = df['trestbps'].fillna(mean)
    print(df['trestbps_replaced'].std())
    ## visualization technique
    df['trestbps'].plot(kind = 'kde' , color = 'r' , legend = 'trestbps')
    df['trestbps_replaced'].plot(kind = 'kde' , color = 'green' , legend = 'trestbps_replaced')
    plt.legend(loc = 'best')
    plt.show()

fun(df)

[7]
... 18.412276951563168
    17.962292044594445

</> <Figure size 640x480 with 1 Axes>

```

Fig 5.3 showing data processing codes

Here we have the code for data processing.

```

cheking with median

def fun(df):
    median = df['trestbps'].median()
    print(df['trestbps'].std())
    df['trestbps_replaced_median'] = df['trestbps'].fillna(median)
    print(df['trestbps_replaced_median'].std())
    ## visualization technique
    df['trestbps'].plot(kind = 'kde' , color = 'r' , legend = 'trestbps')
    df['trestbps_replaced_median'].plot(kind = 'kde' , color = 'green' , legend = 'trestbps_replaced_median')
    plt.legend(loc = 'best')
    plt.show()

fun(df)

[8]
... 18.412276951563168
    17.96675758641796

</> <Figure size 640x480 with 1 Axes>

3rd_technique()

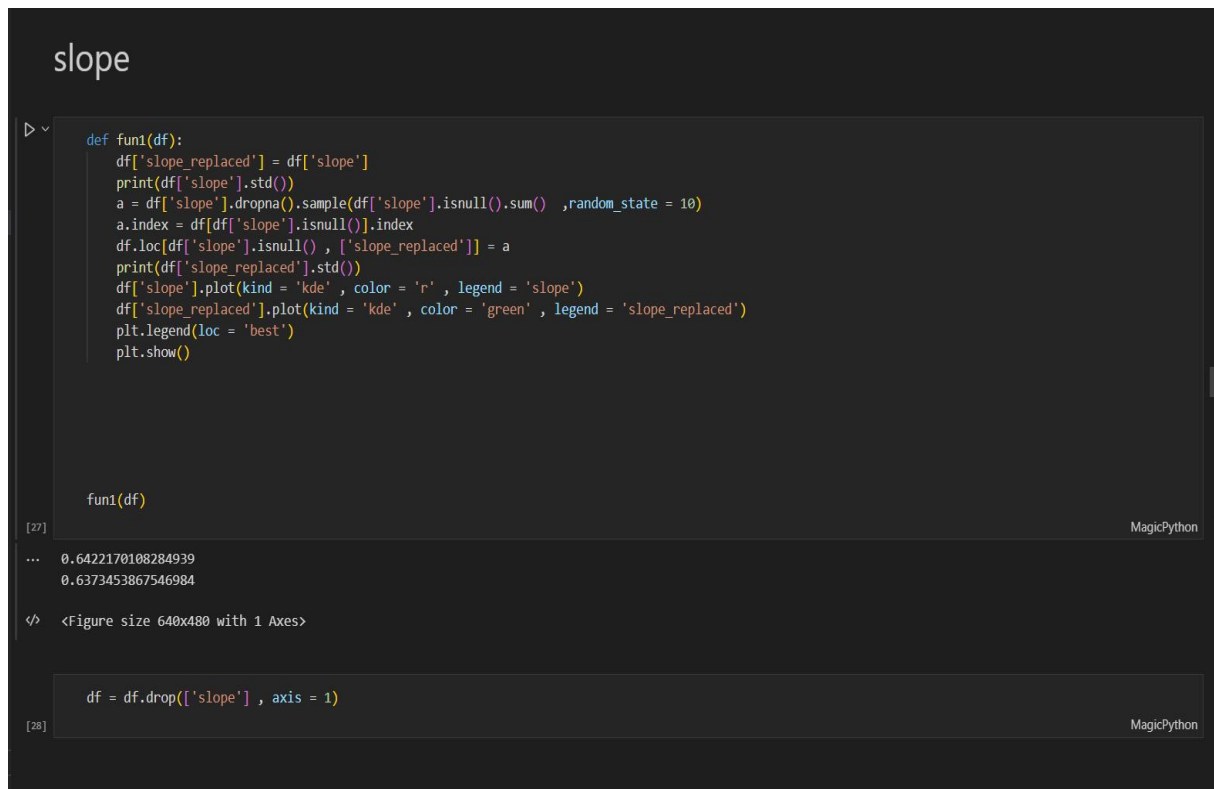
df['trestbps'].dropna().sample(df['trestbps'].isnull().sum() , random_state = 10)

[9]
... Output exceeds the size limit. Open the full output data in a text editor
    228    111.0

```

Fig 5.4 showing checking with median

Here we have the code for the checking of the median of different views



```
slope

def fun1(df):
    df['slope_replaced'] = df['slope']
    print(df['slope'].std())
    a = df['slope'].dropna().sample(df['slope'].isnull().sum(), random_state = 10)
    a.index = df[df['slope'].isnull()].index
    df.loc[df['slope'].isnull(), ['slope_replaced']] = a
    print(df['slope_replaced'].std())
    df['slope'].plot(kind = 'kde', color = 'r', legend = 'slope')
    df['slope_replaced'].plot(kind = 'kde', color = 'green', legend = 'slope_replaced')
    plt.legend(loc = 'best')
    plt.show()

fun1(df)

[27]

... 0.6422170108284939
     0.6373453867546984

<Figure size 640x480 with 1 Axes>

df = df.drop(['slope'], axis = 1)

[28]
```

Fig 5.5 showing slope

The slope is very important to be determined for the process to continue.

CHAPTER-6

TESTING

6.1 Introduction

“Testing” is done to look for ‘mistakes’. Testing is “process” of looking for “defects or weaknesses” in work. It is “software training” process designed to ensure that software system meets its requirements and user expectations and does not fail undesirably. It provides a mechanism for testing the functionality of components, subassemblies, assemblies, and/or the final product.

Tests come in assortment of shapes. Each test sort reacts to certain testing require. Computer program and equipment framework testing is testing done on whole coordinates framework to see in event that it complies with the criteria. Framework testing falls beneath category of "dark box testing," and as such, it shouldn't require understanding the internal workings of the rationale or code. System testing regularly employments the package itself combined with any appropriate equipment framework as well as all of the "coordinates" computer program components that have passed testing integration as input.

6.2 Unit testing

Planning test cases for unit testing guarantees that the center program rationale is working accurately which program inputs result in genuine yields. It is imperative to confirm the inner code stream and all choice branches. It is the testing that's done on the application's isolated program modules after they have been wrapped up but some time recently integration. This sort of meddlesome basic testing requires an understanding of how it was built. Unit tests carry out crucial tests at the component level and look at a specific arrangement of a framework, application, or trade handle. Unit tests make confirmation that each unmistakable course of a trade handle follows accurately to the expressed details and has inputs and yields that are well-defined. Ordinarily, unit testing is done as portion of a combined code.

6.3 Integration Testing

Program components that have been consolidated are tried in integration tests to see on the off chance that they truly work as a single program. Testing is event-driven and centers more on the basic comes about of screens or areas. Integration tests demonstrate that indeed in spite of the fact that the person components were palatable, the combination of the components is exact and steady, as prove by fruitful unit testing. Integration testing is particularly planned to highlight issues that result from combining distinctive components.

Software integration testing involves incrementally integrating two or more software components that have been integrated on a single platform in order to induce failures brought on by interface flaws. The purpose of an integration test is to ensure that software applications or, as an intermediate step, software system components, are functioning as intended.

6.4 System Test

Framework testing makes guaranteeing that coordinates software as entire complies with determinations. In arrange to supply known and unsurprising results, it assesses setup. The configuration-oriented framework integration test is outline of framework testing. Framework testing places accentuation on pre-driven handle associations and integration focuses and is based on forms, portrayals, and streams.

6.4.1 White box testing

A computer program testing method called white box testing (too alluded to as clear box testing, glass box testing, straightforward box testing, and basic testing) analyzes application's inside instruments as restricted to its usefulness (i.e., dark box testing). In white box testing, test cases are made utilizing programming information and inner perspective of the framework. The analyzer picks inputs to undertake out code routes and choose the proper yields. This is often comparable to two testing hubs (i.e. CT) in circuit.

The unit, integration, and framework stages of the program testing prepare may all utilize white-box testing. White-box testing is presently progressively broadly used for integration and framework testing, in spite of the truth that past analyzers tended to think of it as being performed at the unit level. It may test the associations between subsystems amid a system-level test, as well as the associations between units amid integration. In spite of the fact that this approach to test plan can discover parcel of mistakes or issues, it runs the chance of lost

prerequisites or parts of the detail that haven't been actualized.

6.4.2 Black box testing

Black-box Computer program testing that surveys application's working without looking at its inner components is known as "black-box" testing. For all intents and purposes each level of computer program testing, counting unit, integration, framework, and acknowledgment testing, can be conducted utilizing this test strategy. Testing that's based on determinations is another title for it.

CHAPTER 7

RESULTS AND DISCUSSIONS

The goal of this study was to determine if patient questionnaires containing historical subjective and examination-based objective health data might be utilised to detect potential risks for heart disease. Such data may support the diagnostic value of physiological-biochemical tests clinically carried out in CVD in addition to screening. SVMs with strict feature selection were taken into account by the evaluation system. The categories of medical condition, cardiovascular health, and fitness have shown good promise in determining the risk of CVD after a number of tests, with the category of fitness demonstrating significant effectiveness.

Researchers have outlined many machine learning techniques for heart disease prediction. They developed a number of machine learning algorithms and then examined their attributes to determine which one was the best. Every algorithm has produced a distinct outcome in a variety of circumstances. Further analysis shows that the prediction model for heart illness only achieves minimal accuracy; hence, more complicated models are required to improve the accuracy of predicting early heart disease. Future methodologies for highly accurate, low-cost, and simple early heart disease prediction will be proposed. The researchers stated many algorithms and the algorithms have problems also. But the best 3 with the accuracies are Decision Tree, Naive Bayes and LSTM.

Heart Disease Prediction

Enter your age	<input type="text" value="Your current age in years"/>
Enter your Gender	<input type="text" value="Male"/>
Chest pain type?	<input type="text" value="No chest pain"/>
Serum Cholestrol in mg/dl	<input type="text" value="Cholestrol"/>
Rest ECG results	<input type="text" value="Normal"/>
Maximum heart rate achieved during ecg	<input type="text" value="Thalch"/>
Chest pain during exercise?	<input type="text" value="Yes"/>
Oldpeak	<input type="text" value="Oldpeak"/>
Slope[0-3]	<input type="text" value="slope"/>
Ca	<input type="text" value="Ca"/>

Predict

Fig 7.1 showing welcome page of the website

The welcome page of our paper contains the UI which takes the input from user such as age,gender,serum cholesterol,heart rate etc.

Heart Disease Prediction

Enter your age	<input type="text" value="50"/>
Enter your Gender	<input type="text" value="Male"/>
Chest pain type?	<input type="text" value="No chest pain"/>
Serum Cholestrol in mg/dl	<input type="text" value="100"/>
Rest ECG results	<input type="text" value="Normal"/>
Maximum heart rate achieved during ecg	<input type="text" value="100"/>
Chest pain during exercise?	<input type="text" value="Yes"/>
Oldpeak	<input type="text" value="100"/>
Slope[0-3]	<input type="text" value="100"/>
Ca	<input type="text" value="100"/>

Predict

No_RISK

Fig 7.2 showing No Risk Case

No Risk is the case where the health is considered to be fine.

Heart Disease Prediction

Enter your age	<input type="text" value="67"/>
Enter your Gender	<input type="text" value="Male"/>
Chest pain type?	<input type="text" value="Typical angina(Chest pain due to emotional or physical stress)"/>
Serum Cholestrol in mg/dl	<input type="text" value="161"/>
Rest ECG results	<input type="text" value="Normal"/>
Maximum heart rate achieved during ecg	<input type="text" value="286"/>
Chest pain during exercise?	<input type="text" value="Yes"/>
Oldpeak	<input type="text" value="200"/>
Slope[0-3]	<input type="text" value="1.5"/>
Ca	<input type="text" value="1"/>

Stage 1

Fig 7.3: Stage 1 Case

Stage 1 is the stage where the food and diet needs to be proper so that the health won't get affected.

Heart Disease Prediction

Enter your age	<input type="text" value="62"/>
Enter your Gender	<input type="text" value="Male"/>
Chest pain type?	<input type="text" value="Typical angina(Chest pain due to emotional or physical stress)"/>
Serum Cholestrol in mg/dl	<input type="text" value="161"/>
Rest ECG results	<input type="text" value="Normal"/>
Maximum heart rate achieved during ecg	<input type="text" value="286"/>
Chest pain during exercise?	<input type="text" value="Yes"/>
Oldpeak	<input type="text" value="200"/>
Slope[0-3]	<input type="text" value="1.5"/>
Ca	<input type="text" value="1"/>

Stage 2

Fig 7.4 showing Stage 2 Case

Stage 2 is the stage the daily routine need to be checked to get the healthy life.

Heart Disease Prediction

Enter your age	<input type="text" value="70"/>
Enter your Gender	<input type="text" value="Male"/>
Chest pain type?	<input type="text" value="Typical angina(Chest pain due to emotional or physical stress)"/>
Serum Cholestrol in mg/dl	<input type="text" value="161"/>
Rest ECG results	<input type="text" value="Normal"/>
Maximum heart rate achieved during ecg	<input type="text" value="286"/>
Chest pain during exercise?	<input type="text" value="Yes"/>
Oldpeak	<input type="text" value="200"/>
Slope[0-3]	<input type="text" value="1.5"/>
Ca	<input type="text" value="1"/>

Stage 3

Fig 7.5 showing Stage 3 Case

Stage 3 is the case the client needs to go to the hospital for initial check.

Heart Disease Prediction

Enter your age	<input type="text" value="75"/>
Enter your Gender	<input type="text" value="Male"/>
Chest pain type?	<input type="text" value="Typical angina(Chest pain due to emotional or physical stress)"/>
Serum Cholestrol in mg/dl	<input type="text" value="161"/>
Rest ECG results	<input type="text" value="Normal"/>
Maximum heart rate achieved during ecg	<input type="text" value="286"/>
Chest pain during exercise?	<input type="text" value="Yes"/>
Oldpeak	<input type="text" value="200"/>
Slope[0-3]	<input type="text" value="1.5"/>
Ca	<input type="text" value="1"/>

Stage 4

Fig 7.6 showing Stage 4 case

Stage 4 is the stage where its time to get to doctor as early as possible

Heart Disease Prediction

Enter your age	<input type="text" value="81"/>
Enter your Gender	<input type="text" value="Male"/>
Chest pain type?	<input type="text" value="Typical angina(Chest pain due to emotional or physical stress)"/>
Serum Cholestrol in mg/dl	<input type="text" value="161"/>
Rest ECG results	<input type="text" value="Normal"/>
Maximum heart rate achieved during ecg	<input type="text" value="286"/>
Chest pain during exercise?	<input type="text" value="Yes"/>
Oldpeak	<input type="text" value="200"/>
Slope[0-3]	<input type="text" value="1.5"/>
Ca	<input type="text" value="1"/>

Stage 5

Fig 7.7 showing Stage 5 Case

Stage 5 is the critical case they need to go to hospital immediately and might need to be admitted also.

CHAPTER 8

CONCLUSION AND FUTURE WORK

The analysis of the literature demonstrates the need for combinational and more sophisticated algorithms to increase the precision of predicting the early onset of cardiovascular disorders.

The purpose of the study was to discover if it would be possible to identify potential risks for heart disease using patient questionnaires that contained history subjective and examination-based objective health data. In order to make a precise prediction of cardiac illness, this research offers a framework that combines decision trees, random forests, logistic regression, and support vector machines. This paper offers recommendations for training and testing the system, resulting in the best effective model among the various rule-based combinations, using the Heart Disease database. This research also suggests comparing the various results, including sensitivity, specificity, and accuracy. The system will need to be developed using the above approaches, and this will require training and testing the system. It also includes development of a tool to estimate a potential patient's illness risk. Future study on this topic may combine various methods for machine learning in order to improve prediction tools in order to improve the accuracy of coronary artery disease prediction and get a deeper knowledge of the crucial factors, new feature-selection algorithms may also be developed.

CHAPTER 9

PUBLICATION JOURNALS AND CERTIFICATES

9.1 Implementation Research paper

Vol-9 Issue-3 2023

IJARIE-ISSN(O)-2395-4396

A detailed examination of machine learning techniques for predicting heart illness

Dr. ThiruKrishna JT^[1],Prakruthi HR^[2], Nayana Sagar^[3],Moulya G^[4], Priyanshu Singh^[5]

drthirukrishna@dsatm.edu.in^[1],prakruthihr04@gmail.com^[2],nayanas.1dt19is086@gmail.com^[3],
moulyag.1dt19is079@gmail.com^[4],singhpriyanshu073@gmail.com^[5]

¹ Associate Professor , Dayananda Sagar Academy of Tech & MGMT, Bengaluru, India

^{2,3,4,5} Students, Dayananda Sagar Academy of Technology & Management, Bengaluru, India

ABSTRACT

The prenatal recognition of CVDs can help high-risk patients choose whether to change the way they live , which can lessen their severity. Using consistent techniques of machine learning, research has sought to identify the most significant risk variables for heart disease as well as effectively estimate the total risk. In order to produce an accurate predictive algorithm for heart disease, the latest research has looked at bringing together these methods using techniques like machine learning (ml) algorithms. These findings recommend a framework for assessing the precision of implementing particular outcomes from the use of decision trees, k-near neighbor, and logistic regression and SVM on the Cleveland Heart Disease Database.

Keywords:- Cardiovascular disorders, decision trees, logistic regression (LR), machinelearning(ML) with support vector machines (SVM),KNN

1. INTRODUCTION

According to statistics from the WHO, coronary heart disease is the leading cause of fatalities globally, resulting in 17.9 million fatalities [1]. The biggest lifestyle risk factors for cardiovascular disease and stroke include poor eating habits, inactivity, cigarette smoking, and excessive drinking [1]. A cardiac event occurs when the heart's capacity to pump blood is compromised by arterial plaque formation. A stroke can happen when there is a thrombus in an artery that blocks the flow of blood to the brain [2]. As a result of the symptoms' resemblance to those of other conditions and potential confusion with aging symptoms, diagnosing patients can be challenging for health providers. Heart disease is difficult to pin down due to a number of risk factors that are connected to it, such as high cholesterol levels, diabetes, high blood pressure, irregular heartbeat, and numerous other factors. Relevant coronary artery disease prediction and early reconnaissance have become crucial to boosting patient rates of survival.

ML has now established itself as a key instrument in the healthcare sector for aiding in patient diagnosis. The bulk of the time, the existing methods for anticipating and diagnosing cardiac disease rely on practitioners' assessments of the medical history of a patient, symptoms, and results from health screenings. Clinical evaluations and other patient information are openly accessible and expanding daily in databases used by the healthcare sector today. The severity of the disease is assessed using a variety of methods, including the K-Nearest Neighbour Technique (KNN), Decision Trees (DT), random forest algorithm (RF), and naive Bayes (NB) algorithms []. Because heart disease has a complex nature, it needs careful supervision. Failure to do so might harm the heart or result in a premature death.

Numerous techniques have been attempted to acquire knowledge using well-known ML techniques for heart disease prediction. In order to establish a prediction model, numerous analyses have been done throughout this investigation using an assortment of methods as well as by linking multiple strategies.

2. LITERATURE SURVEY

AUTHOR	PURPOSE	TECHNIQUE USED	ACCURACY
1.Gudadhe et al.	Created a diagnosis system for HD diagnosis utilizing recurrent neural networks with multiple levels and support vector machines (SVM).	•SVM algorithm •Neural networks	80.41%
2. Palaniappan et al	A system for professional medical diagnosis was proposed for heart disease identification using ANN,NB and DT	•Artificial Neural Networks •Decision Trees (DT) • Navies Bays (NB).	•88.12% •86.12% •80.4%
3. S. U. J. K. A. KHAN and A. SABOOR	The authors suggested a feature correlation-based NN-based forecasting of coronary heart disease (CHD) research. The present research utilised of the KNHANES-VI sample produced by the Korean Institute for Disease Prevention and Control.	•Fast Conditional Mutual Information method for selecting features (FCMIM) •FCMIM-SVM	92.37%
4. Waqar et al.	Suggested using deep learning based on SMOTE. Without feature selection, the author balanced the dataset using the SMOTE technique. A deep neural network was trained and tested to predict the absence and presence of a cardiac arrest using the balanced dataset,	Deepneural network	96%
5. Fitriyani et al	developed a methodology for HD prediction that combines hybrid synthetic minority over-sampling technique-edited nearest neighbor (SMOTE-ENN) and density-based spatial clustering of applications with noise (DBSCAN).	DBSCAN SMOTE-ENN XG CLASSIFIER BOOST	95.9%

6. MOHAN et al.	Developed hybrid machine learning strategy for HD detection. He also put forth a novel methodology for choosing important characteristics from the information for machine learning classifiers to use in training and testing.	HYBRID ALGORITHM	88.07%
-----------------	---	------------------	--------

3. PROPOSED FRAMEWORK

3.1 Algo Description & Equations

3.1.1 Support Vect. Machine

A type of model known as SV machine is employed in classification and regression analysis to examine data and identify trends. When your data contains precisely two class, S.V.M is employed. By locating the ideal hyperplane that differentiates all of the information's data points in a particular category from those in another, . The mathematical model is accurate to a greater extent the more distance separating the two groups. The inner region of a margin cannot contain any points. The data points on the margin's edge are the support vectors. SVM is a computational modeling strategy that represents tough, practical issues. It is based on mathematical functions. Support The training data is translated into kernel space using vector machines. There are several other kernel spaces that may be employed, including the linear (dot product) kernel, quadratic kernel, polynomial kernel, radial basis function kernel, multilayer perceptron kernel, etc. Moreover, there are other ways to put SVM into practice, including least squares, sequential minimal optimisation, and quadratic programming. The difficult part of SVM is choosing a kernel and a technique such that your model isn't overly optimistic or pessimistic..

It is debatable if the selected kernel is RBF or linear because the CHDD comprises a substantial number of instances and characteristics. Despite the nonlinear relationship between characteristics and class labels, RBF kernel performance may not be enhanced by the sheer amount of features. It is advised to test both kernels before choosing the one that is more effective.

Assume that the information for the trained samples is $\text{Data} = y_i, x_i, i=1, 2, 3, \dots, n$, where $x_i \in \mathbb{R}^n$ represents the i th vector and $y_i \in \mathbb{R}$ defines the target element. The linear Support Vector Machine (SVM) is used to identify the ideal hyperplane with the shape $f(x) = w^T x + b$, where w is a multidimensional parameter vector and b is an interval. By fixing the subsequent optimizations issue, this is achieved:

$$\begin{aligned} \min_{w, b, \xi_i} \quad & \frac{1}{2} w^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t. } & y_i (w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall i \in \{1, 2, \dots, m\} \end{aligned}$$

3.1.2 Decision Trees(DT)

A decision tree(DT) is an approach for categorizing or forecasting what will happen to data using a regression analysis or classifier. When there is continuous data, regression is utilised, while classification is used when the characteristics are clustered. One of the key techniques for data mining is the decision tree. A root node, branches, and leaf nodes make up a decision tree. Follow the path from the root node to a leaf node to assess the data.

A purity index, which will divide the nodes as stated in the training section, must be used to generate decision trees. Each of the 297 tuples is assessed for heart disease using the CHDD decision tree, which results in a positive or negative judgment for each. The reliability, precision, and degree of sensitivity of the predicted outcome are evaluated by comparing them to the initial selection parameter in the CHDD in order to check for erroneous positives or misleading negatives. The splitting parameter that has been utilised further demonstrates the value of each attribute. The trees are built utilising inputs with high entropy for the training instance of D. The top-down recursion division and conquering (DAC) method is used to swiftly and easily produce these trees. To get rid of the unnecessary samples, D is pruned.

$$Entropy = - \sum_{j=1}^m p_{ij} \log_2 p_{ij}$$

3.1.3 Logistic Regression

It is typical to refer to this kind of statistical framework as a log it model, and it is frequently utilised in reclassification and analytical forecasting. Based on a variety of factors that are independent, logistic regression estimates the possibility that a situation, such as voting or not voting, is going to occur. Men are more likely than women to get heart disease, according to the results of the logistic regression analysis. The risk factors for CHD are age, daily cigarette consumption, and systolic blood pressure. Yet, neither the total cholesterol tier1 nor the blood glucose level have changed much.

3.1.4 Naive Bayes

For task classification like categorizing texts, the Naive Bayes classification model is a supervised artificial intelligence methodology. Furthermore, it belongs to the family of generative learning methods, which duplicates the input distribution within an identified group or categories. The NB algorithm could recognise the features associated with heart disease. It displays the potential for each of the 15 input attributes for the predetermined condition.

3.1.5 Random Forest

Leo Breiman and Adele Cutler created the widely used method for machine learning known as random forest modelling, which integrates the outcomes of numerous decision trees to get one final decision. Its versatility and effectiveness, which can handle issues with regression and classification, are what fuel its widespread usage. The random forest (RF) methodology is used in ROC curve. For both the true positive rate and the rate of false positives at various sensitivity settings, the area under the ROC curve is shown. The simulation properly determined whether an individual had coronary artery bypass graft or not, according to the ROC curve's AUC measurement of 93.3%.

3.1.6 K-nearest neighbors

The k-nearest neighbours technique, commonly referred to as KNN or k-NN, is a classifier developed using supervised learning that anticipates or groups how a single data point will be categorized. KNN is a simple classifier in which samples are categorized according to the class of their closest neighbor. High volume is a characteristic of medical databases. Classification may result in less accurate results if the data collection contains redundant and unnecessary properties.

3.2 Architecture

The general layout of the software architecture is shown in Fig 3.2.1. The proposed system's core modules are made

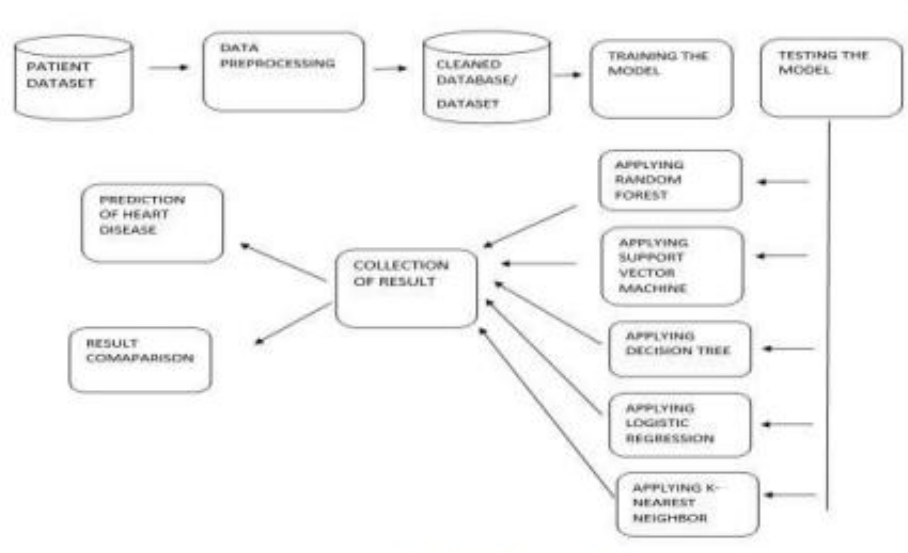


fig 3.2.1 Architecture of HD prediction system

from :

Knowledge of Domain Dataset

The system is given a dataset as input, which is covered in depth. It also has the option of accepting human entries

Processing of Data

Data preparation is the process of transforming data so that it may be used for future analysis.

Module

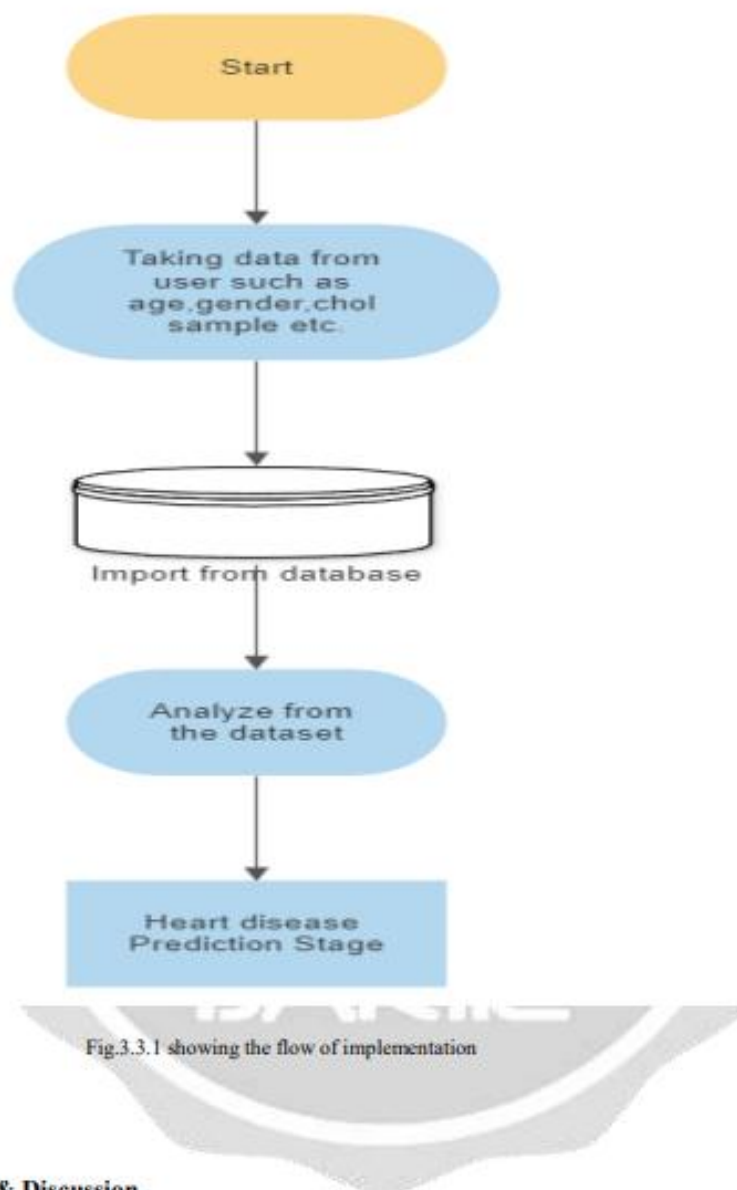
It discusses the algorithmic methodology used on the system to provide very precise findings. In machine learning techniques, we employ SVM, KNN, Random forest, Decision tree as algorithmic approaches.

Assessment and Implementation

Information about the outcome is included in the concluding analysis modules. Our method compares and draws conclusions based on quantifiable results like after getting a confusion matrix, the levels of sensitivity, specificity, accuracy, true positive rate, and the false-positive rate.

3.3 Flowchart

A flowchart is a diagrammatic representation of the steps of a process in a sequence. It's a tool that may be utilized for describing a variety of procedures, such as a production process, service procedure, or a strategy for a project.



4. Results & Discussion

4.1 Simulation Para

The random variable inputs in simulation are typically not precisely understood, although the model is frequently. Inputs are precisely known in machine learning, but the model is unknown before training. The variations in production are relatively slight. Both provide an output, but there are several sources of uncertainty. The final results are detected by the medical experts .

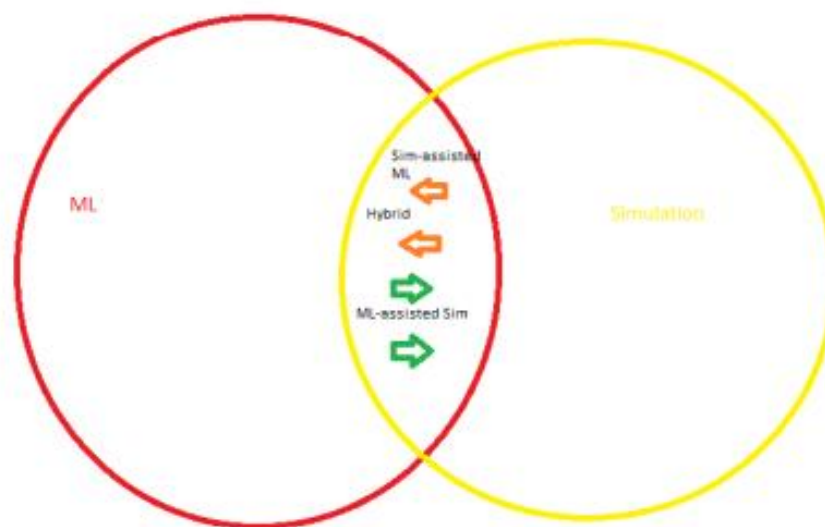


Fig 4.1.1 showing ML and simulation relation

4.2 Output Screenshot

A screenshot of a web application titled 'Heart Disease Prediction'. The interface features a dark green header with the title in white. Below the header, there is a light green background with ten input fields arranged vertically. The input fields are labeled: 'Age of a patient', 'Gender', 'CP', 'cholest', 'resting_blood_pressure', 'Thalach', 'Exersing', 'Oldpeak', 'slope', and 'Ca'. A 'Predict' button is located at the bottom right of the input fields.

Fig4.2.1: Some standard parameters are present and by entering the values against each parameter, The algorithm predicts the type of heart disease.

A screenshot of a web-based heart disease prediction interface. It features a light green background with a central column of ten input fields. The values entered in these fields are: 51, 1, 0, 140, 290, 0, 1, 4.2, 1, and 3. Below the input fields is a black button with the word 'Predict' in white. Underneath the button, the text 'No_RISK' is displayed in a bold, black font.

51
1
0
140
290
0
1
4.2
1
3

Predict

No_RISK

Fig 4.2.2: The figure contains values against each parameter and the algorithm has detected no risk in the patient's heart condition.

A screenshot of a web-based heart disease prediction interface, similar to the one above. It features a light green background with a central column of ten input fields. The values entered in these fields are: 59, 1, 4, 161, 176, 3, 1, 2, 2, and 2. The fourth input field, containing the value '161', is highlighted with a light blue background. Below the input fields is a black button with the word 'Predict' in white. Underneath the button, the text 'Stage 2' is displayed in a bold, black font.

59
1
4
161
176
3
1
2
2
2

Predict

Stage 2

Fig 4.2.3: The figure contains values which have been detected as stage 2 heart condition.

Fig 4.2.4: The figure shows the entries which have the highest risk of having heart disease.

5. APPLICATION

One of the most deadly and curable chronic diseases, heart disease is a leading cause of mortality in both economically developed and underdeveloped nations. If the patient is identified early on and receives the appropriate care, the harm can be significantly mitigated. Hence, early identification can help people make lifestyle adjustments and, if necessary, provide optimal medical care. Cardiovascular disease prediction supports practitioners in making more accurate health decisions for their patients. By processing enormous volumes of complicated health data and exposing clinically meaningful information regarding CVDs, machine learning techniques not only help doctors make more efficient and precise clinical choices but also considerably advance clinical understanding. It is a realistic alternative for limiting and comprehending heart clinical symptoms is through detection using machine learning (ML).

6. CONCLUSION AND FUTURE WORK

By using the review of literature we can draw a conclusion that combinational and more advanced algorithms are to be used to increase the efficiency and accuracy of the predicting system to detect the heart diseases in the earlier stages.

The purpose of the study was to discover if it would be possible to identify potential risks for heart disease using patient questionnaires that contained history subjective and examination-based objective health data. In order to make a precise prediction of cardiac illness, this research offers a framework that combines decision trees, random forests, logistic regression, and support vector machines. This paper offers recommendations for training and testing the system, resulting in the best effective model among the various rule-based combinations, using the Heart Disease database. This research also suggests comparing the various results, including sensitivity, specificity, and accuracy. The system will need to be developed using the above approaches, and this will require training and testing the system. It also includes development of a tool to estimate a potential patient's illness risk. Future study on this topic may combine various methods for machine learning in order to improve prediction tools in order to improve the accuracy of coronary artery disease prediction and get a deeper knowledge of the crucial factors, new feature-selection algorithms may also be developed.

7. REFERENCES

- 1.C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2022,
- 2.V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177-181.
- 3.A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL) Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu School of Computing Sciences and Engineering, VIT University Vellore – 632014, Tamil Nadu, India.
- 4.NHANES: <https://www.cdc.gov/nchs/nhanes/index.htm>
- 5.WHO, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
- 6.S. Singh, and R. Zeltser, "Cardiac Risk Stratification," in: *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2020.
- 7.A. Pandya, et al., "A comparative assessment of non-laboratory- based versus commonly used laboratory-based cardiovascular disease risk scores in the NHANES III population," *PLoS One*, vol. 6, no. 5, pp. e20416, May 2011.
- 8.C.Y. Wang, et al., "Cardiorespiratory fitness levels among US adults 20-49 years of age: findings from the 1999-2004 National Health and Nutrition Examination Survey," *Am J Epidemiol.*, vol. 171, no. 4, pp. 426-435, Feb. 2010.
- 9.P.L. Tsou, and C.J. Wu, "Sex-Dimorphic Association of Plasma Fatty Acids with Cardiovascular Fitness in Young and Middle-Aged General Adults: Subsamples from NHANES 2003-2004," *Nutrients*, vol. 10, no. 10, 1558, Oct. 2018.
- 10.S.S. Yoon, et al., "Trends in the Prevalence of Coronary Heart Disease in the U.S.: National Health and Nutrition Examination Survey, 2001-2012," *Am. J. Prev. Med.*, vol. 51, no. 4, pp. 437-445, Oct. 2016.
- 11.R. Moonesinghe, et al., "Prevalence and Cardiovascular Health Impact of Family History of Premature Heart Disease in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007-2014," *J. Am. Heart Assoc.*, vol. 8, no. 14, e012364, July 2019.
- 12.K. Jindai, et al., "Multimorbidity and Functional Limitations Among Adults 65 or Older, NHANES 2005-2012," *Prev. Chronic Dis.*, vol. 13, 160174, Nov. 2016.
- 13.S.Heyden, et al., "Angina Pectoris and the Rose Questionnaire," *Arch. Intern. Med.*, vol. 128, no. 6, pp. 961-964, 1971.
- 14.A. Koyanagi, et al., "Correlates of physical activity among community-dwelling adults aged 50 or over in six low- and middle- income countries," *PLoS ONE*, vol. 12, no. 10, e0186992, Oct. 2017.
- 15.W.-H. Weng, "Machine Learning for Clinical Predictive Analytics," in: *Leveraging Data Science for Global Health*. L. A. Celi et al. (eds.), 2020, ch. 12

Certificates

INTERNATIONAL JOURNAL OF ADVANCE
RESEARCH AND INNOVATIVE IDEAS IN EDUCATION
★★★

CERTIFICATE

of

PUBLICATION

The Board of International Journal of Advance Research and Innovative Ideas in Education
is hereby Awarding this Certificate to

PRIYANSHU SINGH

In Recognition of the Publication of the Paper Entitled

**A DETAILED EXAMINATION OF MACHINE LEARNING TECHNIQUES FOR PREDICTING HEART
ILLNESS**

Published in E-Journal

Volume-9 Issue-3 2023

Paper Id : 20017
ISSN(O) : 2395-4396



www.ijarie.com

Editor In Chief

(N Patel)



**INTERNATIONAL JOURNAL OF ADVANCE
RESEARCH AND INNOVATIVE IDEAS IN EDUCATION**
★★★
CERTIFICATE

of

PUBLICATION

The Board of International Journal of Advance Research and Innovative Ideas in Education
is hereby Awarding this Certificate to

PRAKRUTHI HR

In Recognition of the Publication of the Paper Entitled

**A DETAILED EXAMINATION OF MACHINE LEARNING TECHNIQUES FOR PREDICTING HEART
ILLNESS**

Published in E-Journal

Volume-9 Issue-3 2023

Paper Id : 20017
ISSN(O) : 2395-4396



www.ijarie.com

Editor In Chief

(N Patel)



**INTERNATIONAL JOURNAL OF ADVANCE
RESEARCH AND INNOVATIVE IDEAS IN EDUCATION**

CERTIFICATE

of

PUBLICATION

*The Board of International Journal of Advance Research and Innovative Ideas in Education
is hereby Awarding this Certificate to*

NAYANA SAGAR

In Recognition of the Publication of the Paper Entitled

**A DETAILED EXAMINATION OF MACHINE LEARNING TECHNIQUES FOR PREDICTING HEART
ILLNESS**

Published in E-Journal

Volume-9 Issue-3 2023

Paper Id : 20017
ISSN(O) : 2395-4396



www.ijariie.com

Editor In Chief

(N Patel)

INTERNATIONAL JOURNAL OF ADVANCE
RESEARCH AND INNOVATIVE IDEAS IN EDUCATION
★★★

CERTIFICATE

of

PUBLICATION

The Board of International Journal of Advance Research and Innovative Ideas in Education
is hereby Awarding this Certificate to

MOULYA G

In Recognition of the Publication of the Paper Entitled

**A DETAILED EXAMINATION OF MACHINE LEARNING TECHNIQUES FOR PREDICTING HEART
ILLNESS**

Published in E-Journal

Volume-9 Issue-3 2023

Paper Id : 20017
ISSN(O) : 2395-4396



www.ijarjie.com

Editor In Chief

(N Patel)



Literature Survey Published Paper Certificates

**INTERNATIONAL JOURNAL OF ADVANCE
RESEARCH AND INNOVATIVE IDEAS IN EDUCATION**

CERTIFICATE

of

PUBLICATION

*The Board of International Journal of Advance Research and Innovative Ideas in Education
is hereby Awarding this Certificate to*

PRIYANSHU SINGH

In Recognition of the Publication of the Paper Entitled

**AN EXTENSIVE ANALYSIS OF DATA MINING AND MACHINE LEARNING STRATEGIES FOR HEART
DISEASE PREDICTION**

Published in E-Journal

Volume-9 Issue-3 2023

Paper Id : 20147
ISSN(O) : 2395-4396



www.ijarjie.com

Editor In Chief

(Nare)

**INTERNATIONAL JOURNAL OF ADVANCE
RESEARCH AND INNOVATIVE IDEAS IN EDUCATION**

CERTIFICATE

of

PUBLICATION

*The Board of International Journal of Advance Research and Innovative Ideas in Education
is hereby Awarding this Certificate to*

PRAKRUTHI HR

In Recognition of the Publication of the Paper Entitled

**AN EXTENSIVE ANALYSIS OF DATA MINING AND MACHINE LEARNING STRATEGIES FOR HEART
DISEASE PREDICTION**

Published in E-Journal

Volume-9 Issue-3 2023

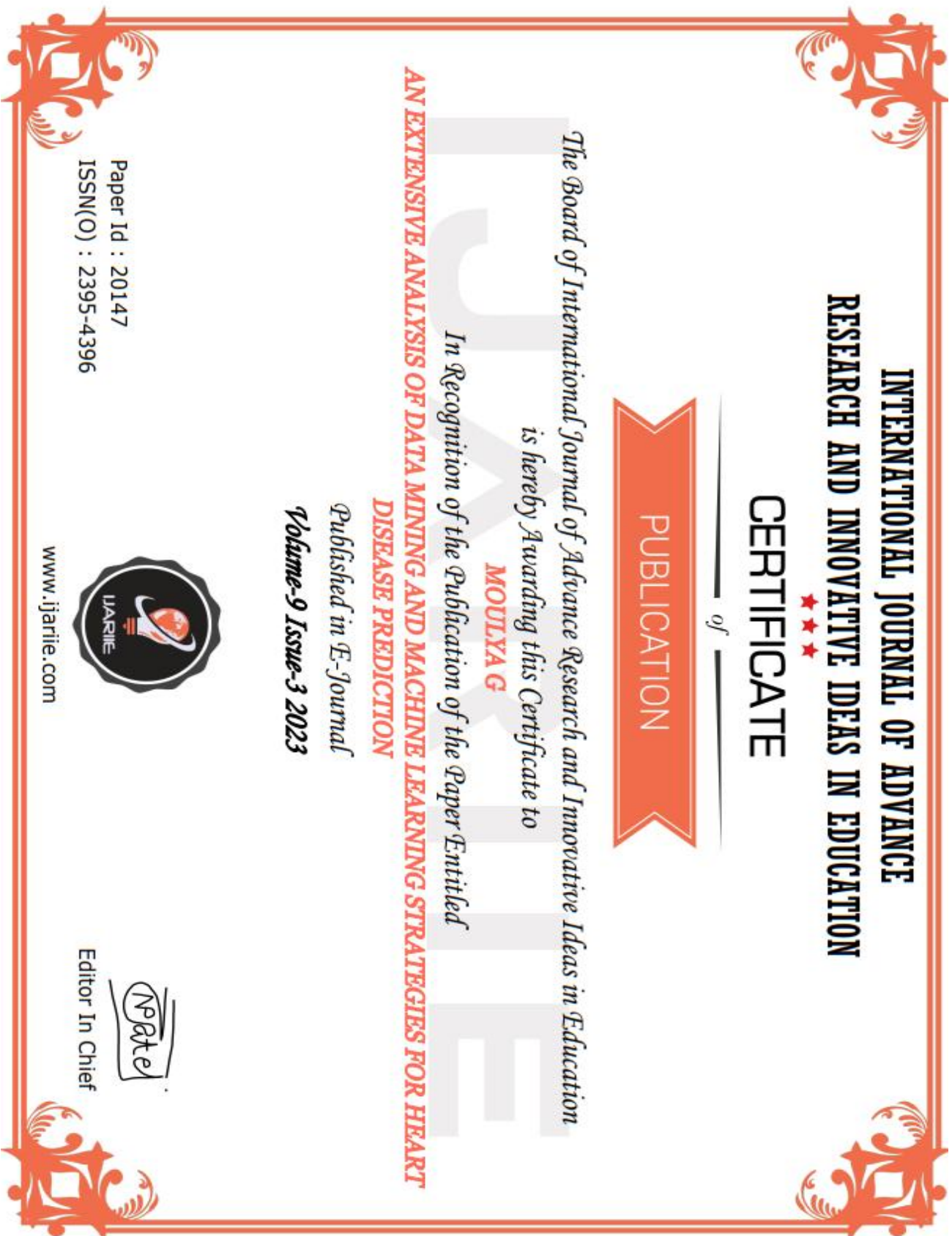
Paper Id : 20147
ISSN(O) : 2395-4396



www.ijariie.com

Editor In Chief

(N Patel)



**INTERNATIONAL JOURNAL OF ADVANCE
RESEARCH AND INNOVATIVE IDEAS IN EDUCATION**

CERTIFICATE

of

PUBLICATION

*The Board of International Journal of Advance Research and Innovative Ideas in Education
is hereby Awarding this Certificate to*

NAYANA SGAR

In Recognition of the Publication of the Paper Entitled

**AN EXTENSIVE ANALYSIS OF DATA MINING AND MACHINE LEARNING STRATEGIES FOR HEART
DISEASE PREDICTION**

*Published in E-Journal
Volume-9 Issue-3 2023*

Paper Id : 20147
ISSN(O) : 2395-4396



www.ijariie.com

Editor In Chief

(Signed)

**INTERNATIONAL JOURNAL OF ADVANCE
RESEARCH AND INNOVATIVE IDEAS IN EDUCATION**

CERTIFICATE

of

PUBLICATION

*The Board of International Journal of Advance Research and Innovative Ideas in Education
is hereby Awarding this Certificate to*

DR. THIRUKRISHNA JT

In Recognition of the Publication of the Paper Entitled

**AN EXTENSIVE ANALYSIS OF DATA MINING AND MACHINE LEARNING STRATEGIES FOR HEART
DISEASE PREDICTION**

Published in E-Journal

Volume-9 Issue-3 2023

Paper Id : 20147
ISSN(O) : 2395-4396



www.ijariie.com

Editor In Chief

(N Patel)

REFERENCES

- [1] Y. I. Mir and S. Mitta, “Thyroid disease prediction using hybrid machine learning techniques: An effective framework,” *International Journal of Science and Technology*, Vol. 9, No. 2, pp. 2868–2874, 2020, doi: 10.1109/ACCESS.2022.3190416
- [2] V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques,"2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177-181.
- [3] A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL) Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu School of Computing Sciences and Engineering, VIT University Vellore – 632014, Tamil Nadu, India.
- [4] NHANES: <https://www.cdc.gov/nchs/nhanes/index.htm>
- [5] WHO: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
- [6] C.Y. Wang, et al., “Cardiorespiratory fitness levels among US adults 20-49 years of age: findings from the 1999-2004 National Health and Nutrition Examination Survey,” *Am J Epidemiol.*, vol. 171, no. 4, pp. 426-435, Feb. 2010.
- [7] S.S. Yoon, et al., “Trends in the Prevalence of Coronary Heart Disease in the U.S.: National Health and Nutrition Examination Survey, 2001-2012,” *Am. J. Prev. Med.*, vol. 51, no. 4, pp. 437-445, Oct. 2016.
- [8] R. Moonesinghe, et al., “Prevalence and Cardiovascular Health Impact of Family History of Premature Heart Disease in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007-2014,” *J. Am. Heart Assoc.*, vol. 8, no. 14, e012364, July 2019.
- [9] K. Jindai, et al., “Multimorbidity and Functional Limitations Among Adults 65 or Older, NHANES 2005–2012,” *Prev. Chronic Dis.*, vol. 13, 160174, Nov. 2016.
- [10] H. Sanz, et al., “SVM-RFE: selection and visualization of the most relevant features through non-linear kernels,” *BMC Bioinformatics*, vol. 19, 432, Nov. 2018.

ORIGINALITY REPORT

24%

SIMILARITY INDEX

12%

INTERNET SOURCES

8%

PUBLICATIONS

18%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Visvesvaraya Technological University, Belagavi Student Paper	5%
2	Submitted to University of Westminster Student Paper	2%
3	github.com Internet Source	1%
4	Sai Ramya Akula. "Semi supervised machine learning approach for DDOS detection", International Journal of Innovative Research in Education, 2021 Publication	1%
5	www.humanitarianresponse.info Internet Source	1%
6	Syed Ammad Ali Shah, Ayat Hama Saleh, Mahsa Ebrahimian, Rasha Kashef. "Early Detection of Heart Disease Using Advances of Machine Learning for Large-Scale Patient Datasets", 2022 IEEE Canadian Conference on	1%

Electrical and Computer Engineering (CCECE), 2022

Publication

7	Submitted to University of Hertfordshire Student Paper	1 %
8	dspace.uiu.ac.bd Internet Source	1 %
9	Submitted to Purdue University Student Paper	1 %
10	research.ijcaonline.org Internet Source	1 %
11	Submitted to Harrisburg University of Science and Technology Student Paper	1 %
12	paperhost.org Internet Source	1 %
13	www.researchgate.net Internet Source	1 %
14	Submitted to AlHussein Technical University Student Paper	<1 %
15	Jingran Li, Fei Tao, Ying Cheng, Liangjin Zhao. "Big Data in product lifecycle management", The International Journal of Advanced Manufacturing Technology, 2015 Publication	<1 %

16	ejmcm.com Internet Source	<1 %
17	Submitted to International School of Management and Technology Student Paper	<1 %
18	Submitted to Kingston University Student Paper	<1 %
19	Submitted to University of Huddersfield Student Paper	<1 %
20	cbio.ensmp.fr Internet Source	<1 %
21	www.ijasret.com Internet Source	<1 %
22	www.ijrti.org Internet Source	<1 %
23	Submitted to Arab Academy for Science, Technology & Maritime Transport CAIRO Student Paper	<1 %
24	Submitted to British University In Dubai Student Paper	<1 %
25	Submitted to University of Leicester Student Paper	<1 %
26	www.freecodecamp.org Internet Source	<1 %

27	fts.vau.ac.lk Internet Source	<1 %
28	Bhavesh Dhande, Kartik Bamble, Sahil Chavan, Tabassum Maktum. "Diabetes & Heart Disease Prediction Using Machine Learning", ITM Web of Conferences, 2022 Publication	<1 %
29	ir.xjtlu.edu.cn Internet Source	<1 %
30	link.springer.com Internet Source	<1 %
31	Submitted to The British College Student Paper	<1 %
32	Submitted to University of North Texas Student Paper	<1 %
33	hcisj.com Internet Source	<1 %
34	www.ostack.cn Internet Source	<1 %
35	www0.cs.ucl.ac.uk Internet Source	<1 %
36	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2020 Publication	<1 %

37

jcdronline.org

Internet Source

<1 %

38

studentsrepo.um.edu.my

Internet Source

<1 %

39

www.slideshare.net

Internet Source

<1 %

40

G. A. Klados, K. Politof, E. S. Bei, K. Moirogiorgou, N. Anousakis-Vlachochristou, G. K. Matsopoulos, M. Zervakis. "Machine Learning Model for Predicting CVD Risk on NHANES Data", 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021

Publication

<1 %

Exclude quotes On

Exclude matches Off

Exclude bibliography On