# HEART DISEASE IDENTIFICATION METHOD
# USING MACHINE LEARNING
# CLASSIFICATION IN E-HEALTHCARE

**A PROJECT REPORT**

***Submitted by***

| | |
|---|---|
| **JAYANT.H** | **113319104032** |
| **SANJAY.R** | **113319104074** |
| **VAJJA HEMASAI** | **113319104097** |
| **VENKATASRINIVAS.P.V** | **113319104103** |

***In partial fulfilment for the award of the degree***

***Of***

**BACHELOR OF ENGINEERING**

***In***

**COMPUTER SCIENCE AND ENGINEERING**

**VELAMMAL INSTITUTE OF TECHNOLOGY**

**CHENNAI 601 204**

**ANNA UNIVERSITY**

**MAY-2023**

# BONAFIDE CERTIFICATE

Certified that this project report **“ HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE ”** is the bonafide work of **“ JAYANT H 113319104032, SANJAY R 113319104074, VAJJA HEMASAI 113319104097, VENKATASRINIVAS P V 113319104103 ”** who carried out the project work under my supervision.

**SIGNATURE**

**Dr.V.P.GLADIS PUSHPARATHI,**
**HEAD OF THE DEPARTMENT**
Computer Science and Engineering,
Velammal Institute of Technology,
Velammal Knowledge Park, Kolkata
Highway,Panchetti,Chennai-601204.

**SIGNATURE**

**Mr. J.A.JEVIN, M.E,**
**ASSISTANT PROFESSOR**
Computer Science and Engineering,
Velammal Institute of Technology,
Velammal Knowledge Park, Kolkata
Highway,Panchetti,Chennai-601204.

# HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE

## VIVA-VOCE EXAMINATION

The viva-voce examination of this project work was done as a part of the Bachelors Degree in Computer Science and Engineering held on ________________________________.

| | |
|---|---|
| **JAYANT.H** | **113319104032** |
| **SANJAY.R** | **113319104074** |
| **VAJJA HEMASAI** | **113319104097** |
| **VENKATASRINIVAS.P.V** | **113319104103** |

**INTERNAL EXAMINER** **EXTERNAL EXAMINER**

# ACKNOWLEDGEMENT

We are personally indebted to many who had helped us during the course of this project work. Our deepest gratitude to the God Almighty.

We are greatly and profoundly thankful to our beloved Chairman **Thiru.M.V.Muthuramalingam** for facilitating us with this opportunity. Our sincere thanks to our respected Director **Thiru.M.V.M.Sasi Kumar** who took keen interest on us. We are also thankful to our Advisors **Prof.K.Razak, Shri.M.Vasu** and our Principal **Dr.N.Balaji** and Vice Principal **Dr.Soundararajan** for their never-ending encouragement which accelerates us towards innovation.

We are extremely thankful to our Head of the Department **Dr.V.P.Gladis Pushparathi** Project Coordinator **Mr.A.Anbumani**, Assistant Professor, Computer Science and Engineering for their valuable teachings and suggestions.

From the bottom of our heart, we would like to thank our guide **Mr.J.A.Jevin, Assistant Professor ,Computer Science and Engineering** who has been the pillar of this project without whom we would not have been able to complete the project successfully.

The Acknowledgement would be incomplete if we could not mention word of thanks to our parents, Teaching and Non-Teaching staffs, administrative staffs and Friends who had motivated and lead their support throughout the project Thank you one and all.

# ABSTRACT

The proposed methodology to handle the mining of distributed medical data sources using Association Rules is a promising approach to address the limitations of centralized EHRs. By decomposing global computations into local computations, it can ensure that sensitive medical data is kept private and secure. The use of agents to represent each distributed data source is an interesting concept that can facilitate the exchange of minimal summaries with other agents or allow local tasks to be performed at each site.

This approach can reduce the need for communication or travel by participating agents across the network, which can reduce the risk of data breaches and other security concerns. Association Rules can be a useful tool in this methodology as it can identify patterns and relationships in large datasets, which can lead to insights and better treatment planning. However, it is important to ensure that the rules generated are accurate and reliable, and that the methodology is designed to handle different distributions of data and different participating nodes. Overall, this proposed methodology can be a valuable contribution to the field of EHRs and can address some of the limitations of centralized EHRs. However, further research and testing may be needed to determine its feasibility and effectiveness in real-world applications.

# ஆய்வுசுருக்கம்

மருத்துவ மாற்றுக்களுக்கு எளிய தடை வழங்கி கொண்டு கொள்ள நோக்கம் உள்ள மின் மருத்துவ பதிவுகள் (EHRs) சேமிக்கப்படுகின்றன. தகுந்து சேகரிக்கப்பட்ட EHRs எல்லாம் சேர்க்கப்பட்டு விநியோகத்திற்கு அனுப்பப்படும், பொருளாதாரத்திற்கு பயன்படுத்தப்படுகின்றன. இதுவரை, மருத்துவக் கண்காணிப்புகள் மற்றும் போதுமான சிகிச்சைகளுக்கு ஏற்றுமதி வழங்கப்பட்டுள்ளது. எனவே, EHRs வகைகள் சென்று முன்னனுப்ப வேண்டிய தடைகளுக்கு மறுபடியும் குறைவாக பயன்படுத்தப்பட வேண்டும். இந்த கட்டுரையில், நடுத்தர மருத்துவ தரவு மூலங்களை சேர்ந்து தேடுதல் மூலம் Association Rules பயன்படுத்தி மருத்துவ தரவுகளின் கணிப்புகளை பரிசோதிக்கும் புதிய முறை இது.

மருத்துவ பரிசோதனைக்கு எலக்ட்ரானிக் மருத்துவ பதிவுகள் (EHRs) பயன்படுகின்றன. மொழிபெயர்க்கப்பட்ட நகர்ப்புற மருத்துவ அமைப்புகளில் இது பயன்படுகின்றது. EHRs கிடைத்தனம், சிகிச்சை மற்றும் பாதிப்புகளின் உரிமைகளை சரிசெய்வதற்கு உதவும். ஆனால், ஏற்கனவே உள்ள EHRs மாதிரிகள் மையங்கள் கொண்டு உள்ளன. இவ்வளவு பரிமாற்றங்கள் இருந்தும், பிரச்சினைகள் இருக்கும். எனவே, மருத்துவ நோய் கணிப்பு மற்றும் முன்னணி வழிகாட்டுதலுக்கு EHRs நன்மைகள் இருக்குமானாலும், மருத்துவர்கள் விரும்பப்படமாட்டார்கள் என்பது சில தடைகள் உள்ளன. நம்பிக்கையாக, மருத்துவர்கள் எங்கும் EHRs ஐ அணுக முடியாது.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **EHR** | ELECTRONIC HEALTH CARE |
| **HRV** | HEART RATE VARIABILITY |
| **CHF** | CONGESTIVE HEART FAILURE |
| **ML** | MACHINE LEARNING |
| **WHO** | WORLD HEALTH ORGANIZATION |
| **IDE** | INTEGRATED DEVELOPMENT ENVIRONMENT |
| **RF** | RANDOM FOREST |
| **PVM** | PYTHON VIRTUAL MACHINE |

# CHAPTER 1

# INTRODUCTION

The growth in medical data collection presents a new opportunity for physicians to improve patient diagnosis. In recent years, practitioners have increased their usage of computer technologies to improve decision-making support. In the health care industry, machine learning is becoming an important solution to aid the diagnosis of patients. Machine learning is an analytical tool used when a task is large and difficult to program, such as transforming medical record into knowledge, pandemic predictions, and genomic data analysis. Recent studies have used machine learning techniques to diagnose different cardiac problems and make a prediction. A major problem of machine learning is the high dimensionality of the dataset. The analysis of many features requires a large amount of memory and leads to an over fitting, so the weighting features decrease redundant data and processing time, thus improving the performance of the algorithm. Finding a small set of features characterizes different diseases of health management, genome expression, medical images, and IoT. Dimensionality reduction uses feature extraction to transform and simplify data, while feature selection reduces the dataset by removing useless features

# CHAPTER 2

# LITERATURE SURVEY

A literature review is a text of a scholarly paper, which includes the current knowledge including substantive findings, as well as theoretical and methodological contribute ions to a particular topic. Literature reviews are secondary sources and do not report new or original experimental work.

**1.Title:** Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques

**Authors:** Pronab Ghosh1, Sami Azam, Mirjam Jonkman, Asif Karim, F. M. Javed Mehedi Hamrat, Eva Ignatious, Shahana Shultana, Abhijith Reddy Beeravolu, and Friso de boer

**Journal:** IEEE Journal on digital health(Volume: 9)

**Year:** 2021

**Abstract:** Cardiovascular diseases (CVD) are among the most common serious illnesses affecting human health. CVDs may be prevented or mitigated by early diagnosis, and this may reduce mortality rates. Identifying risk factors using machine learning models is a promising approach. We would like to propose a model that incorporates different methods to achieve effective prediction of heart disease. For our proposed model to be successful, we have used efficient Data Collection, Data Pre-processing and Data Transformation methods to create accurate information for the training model. We have used a combined dataset (Cleveland, Long Beach VA, witzerland, Hungarian and Stat log). Suitable features are selected by using the Relief, and Least Absolute Shrinkage and Selection Operator (LASSO) techniques. New hybrid classifiers like Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), K-Nearest Neighbors Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM), and Gradient Boosting Boosting Method (GBBM) are developed by integrating the traditional classifiers with bagging and boosting methods, which are used in the training process. We have also instrumented some machine learning algorithms to calculate the Accuracy (ACC), Sensitivity (SEN), Error Rate, Precision (PRE) and F1 Score (F1) of our model, along with the Negative Predictive Value (NPR), False Positive Rate (FPR), and False Negative Rate (FNR). The results are shown separately to provide comparisons.

Based on the result analysis, we can conclude that our proposed model produced the highest accuracy while using RFBM and Relief feature selection methods.

**2. Title:** Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method

**Authors:** Rahma Atallah, Amjed Al-Mousa

**Journal:** IEEE Journal on digital health(Volume: 7)

**Year:** 2021

**Abstract:** This paper presents a majority voting ensemble method that is able to predict the possible presence of heart disease in humans. The prediction is based on simple affordable medical tests conducted in any local clinic. Moreover, the aim of this project is to provide more confidence and accuracy to the Doctor's diagnosis since the model is trained using real-life data of healthy and ill patients. The model classifies the patient based on the majority vote of several machine learning models in order to provide more accurate solutions than having only one model. Finally, this approach produced an accuracy of 90% based on the hard voting ensemble model.

**3. Title:** MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis

**Authors:** Ankur Gupta, Rahul Kumar, Harkirat Singh Arora, And Balasubramanian Raman

**Journal:** IEEE Journal on digital health (Volume: 8)

**Year:** 2020

**Abstract:** Cardiovascular disease tops the list among all major causes of deaths worldwide. Though, prognostication and in-time diagnosis can help in reducing the mortality rate as well as increases the survival rate of patients. Unavailability or scarcity of radiologists and doctors in different countries due to several reasons is a significant factor for hindrance in early diagnosis. Among various efforts of developing the decision support systems, computational intelligence is an emerging trend in the field of medical imaging to detect, prognosticate and diagnose the disease. It helps radiologists and doctors to get relief from being over-burdened and minimizes the induced delays for in-time diagnosis of patients. In this work, a machine intelligence framework for heart disease diagnosis MIFH has been proposed. MIFH utilizes the factor analysis of mixed data (FAMD) to extract as well as derive features from the UCI heart disease Cleveland dataset and train the machine learning predictive models. The framework MIFH is validated using the holdout validation scheme. Experimentation results show that MIFH performed well over several baseline methods of recent times in terms of accuracy and comparable in terms of sensitivity and specificity. MIFH returns best possible solution among all input predictive models considering performance criteria and improves the efficacy of the system, hence can assist doctors and radiologists in a better way to diagnose heart patients.

**4. Title:** Deep Ensemble Detection of Congestive Heart Failure Using Short-Term RR Intervals

**Authors:** Ludi Wang , Wei Zhou, Qing Chang, Jiangen Chen, And Xiaoguang Zhou1.

**Journal:** IEEE on data-enabled for digital health (Volume: 7)

**Year:** 2019

**Abstract:** Heart rate variability (HRV) is an effective predictor of congestive heart failure (CHF). However, important challenges exist regarding the effective temporal feature extraction and efficient classification using high-dimensional HRV representations. To solve these challenges, an ensemble method for CHF detection using short-term HRV data and deep neural networks was proposed. In this paper, five open- source databases, the BIDMC CHF database (BIDMC-CHF), CHF RR interval database (CHF-RR), MIT- BIH normal sinus rhythm (NSR) database, fantasia database (FD), and NSR RR interval database (NSR- RR), were used. Additionally, three RR segment length types (N = 500, 1000, and 2000) were used to evaluate the proposed method. First, we extracted the expert features of RR intervals (RRIs) and then built a long short-term memory-convolutional neural network-based network to extract deep-learning (DL) features automatically. Finally, an ensemble classifier was used for CHF detection using the above features. With blindfold validation (three CHF subjects and three normal subjects), the proposed method achieved 99.85%, 99.41%, and 99.17% accuracy on N = 500, 1000, and 2000 length RRIs, respectively, using the BIDMC-CHF, NSR, and FD databases. With blindfold validation (six CHF subjects and six normal subjects), the proposed method achieved 83.84%, 87.54%, and 85.71% accuracy on N = 500, 1000, and 2000 length RRIs, respectively, using the NSR-RR and CHF-RR n-databases. Based on feature ranking, the significant effectiveness provided by the DL features has been proven. The results have shown that the deep ensemble method can achieve reliable CHF detection using short-term heart rate signals and enable CHF detection through intelligent hardware.

**5. Title:** Machine Learning Techniques For Heart Disease Prediction

**Authors:** A. Lakshmanarao,Y.Swathi, P.Sri Sai Sundareswar

**Journal:** IEEE Journal on digital health(Volume: 8,Issue: 11)

**Year:** 2019

**Abstract:** According to WHO (World Health Organization), Heart diseases are the reason for 12 million deaths every year. In most of the countries, half of the deaths are due to cardiovascular diseases. The early diagnosis of cardiovascular sicknesses can help in settling on choices on the way of life changes in high hazard patients and thusly diminish the difficulties. In this paper, machine learning techniques are used for the detection of heart disease. We also applied sampling techniques for handling unbalanced datasets. Various machine learning methods are used to predict the overall risk. The framingham_heart_disease dataset is public available on the Kaggle. This dataset is used in our experiments. The end goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset contains 15 features that give patient information. By applying machine learning techniques, we achieved 99% accuracy in heart disease detection.

# CHAPTER 3
# SYSTEM ANALYSIS

System analysis is the act, process, or profession of studying an activity (as a procedure, a business, or a physiological function) typically by mathematical means in order to define its goals or purposes and to discover operations and procedure for accomplishing them most efficiently. It is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components. System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem-solving technique that improves the system and ensures thatall the components of the system work efficiently to accomplish their purpose.

## 3.1 EXISTING SYSTEM

- Medical data is often distributed among various organizations and hospitals, and it is not feasible to move this data to a centralized location due to privacy and ownership concerns.
- It is desirable to have algorithms that can perform computations on the distributed data without moving it.
- This can be achieved by decomposing the computation into localized computations that can be performed locally within each site using their physical databases.
- The challenge is to develop general algorithms that can handle different data distributions and participating nodes.

## DISADVANTAGES:

- These information systems are segregated on the basis of the type of data stored and where they are used. They work separately from each other, because they are localized stand-alone systems.
- In order to improve the quality of healthcare services by granting shared access to data, healthcare systems are opting for the integration of data sources.

## 3.2 PROBLEM STATEMENT

- Expert choice system in light of AI classifiers and the use of fake fluffy rationale is successfully finding the HD therefore, the proportion of death diminishes and The Cleveland heart illness informational index was utilized by different analysts also for the distinguishing proof issue of HD.
- The machine learning prescient models need appropriate information for preparing and testing.
- The presentation of AI model can be expanded whenever adjusted dataset is use for preparing and testing of the model.
- Moreover, the model prescient abilities can improve by utilizing appropriate and related highlights from the information.
- Hence, information adjusting and highlight determination is altogether significant for model execution improvement.
- In writing different analysis strategies have been proposed by different analysts, anyway these strategies are most certainly not successfully analysis HD.

## 3.3 PROPOSED SYSTEM

- This paper proposes a methodology and algorithm for mining distributed medical data sources.
- The data sources are located at different sites such as hospitals and clinics.
- Global computation must be decomposed into local computations due to data distribution constraints.
- Each distributed data source is represented by an agent in the proposed methodology.
- The global association rule computation is performed by the agent by exchanging minimal summaries or travelling to all the sites.

### ADVANTAGES:

- The advantages of EHRs in disease diagnosis and prediction, they are still not widely adopted by physicians
- The barriers that restrain physicians from using EHRs is main advantage.

# CHAPTER 4

# REQUIREMENT SPECIFICATION

## 4.1 INTRODUCTION

Requirement analysis determines the requirements of a new system. This project analyses on product and resource requirement, which is required for this successfulsystem. The product requirement includes input and output requirements in term of input to produce the required output. The resource requirements give in brief about the software and hardware that are needed to achieve the required functionality.

## HARDWARE AND SOFTWARE SPECIFICATION

## 4.2 HARDWARE REQUIREMENTS

| | |
|---|---|
| • Hard Disk | 80GB and above |
| • RAM | 8GB and above |
| • Processor | Intel i3 and above |
| • Speed | 1.1 GHz |

**Table 4.1 Hardware Requirements**

## 4.2 SOFTWARE REQUIREMENTS

| | |
|---|---|
| • Front End | HTML,CSS |
| • Operating System | Windows 7 and above (64 bit) |
| • Scripts | Python Language |
| • Tool | Python3 IDE |

**Table 4.2 Software Requirements**

## 4.4 TECHNOLOGIES USED:

### PYTHON

Python is a popular high-level, interpreted programming language known for its simplicity, readability, and versatility. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is widely used in various fields such as web development, data science, machine learning, scientific computing, and more.

### INTRODUCTION TO PYTHON

Python is a high-level object-oriented programming language that was created by Guido van Rossum. It is also called general-purpose programming language as it is used in almost every domain we can think of as mentioned below:

- Web Development
- Software Development
- Game Development
- AI & ML
- Data Analytics

This list can go on as we go but why python is so much popular let’s see it in the next topic.

### PYTHON PROGRAMMING

Every Programming language serves some purpose or use-case according to a domain. for eg, Javascript is the most popular language amongst web developers as it gives the developer the power to handle applications via different frameworks like react,angular which are used to build beautiful User Interfaces. Similarly, they have pros and cons at the same time. so if we consider python it is general-purpose which means it is widely used in every domain the reason is it’s very simple to understand, scalable because of which the speed of development is so fast. Now you get the idea why besides learning python it doesn’t require any programming background so that’s why it’s popular amongst developers as well. Python has simpler syntax similar to the English language and also the syntax allows developers to write programs with fewer lines of code. Since it is open-source there are many libraries available that make developers' jobs easy ultimately results in high productivity. They can easily focus on business logic and Its demanding skills in the digital era where information is available in large data sets.

## WORKING OF PYTHON PROGRAMMING

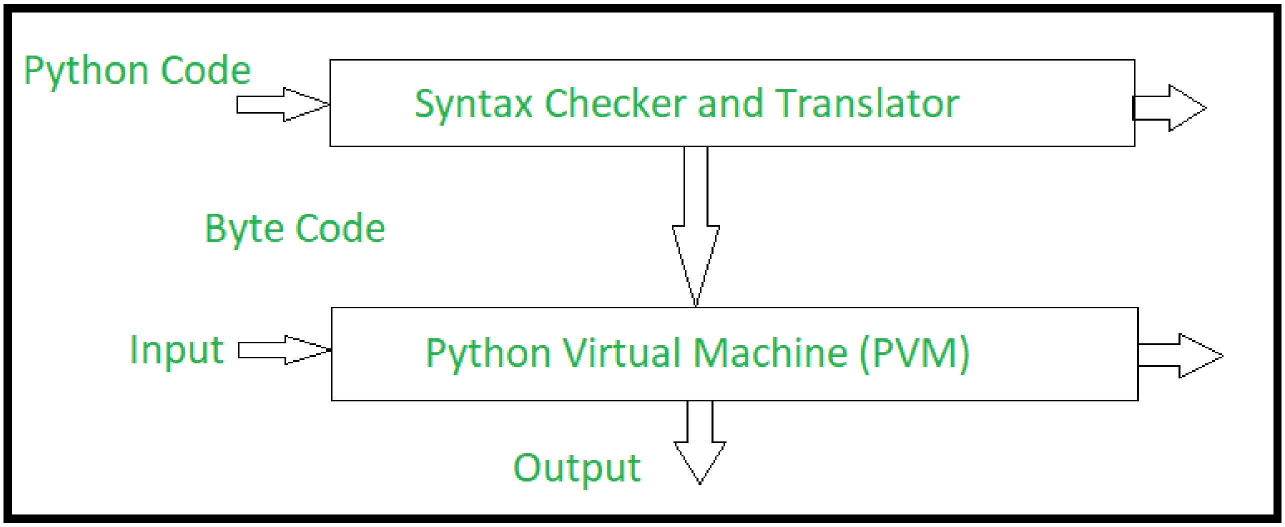


**Figure 4.3 Internal working of Python**

Step 1: The python compiler reads a python source code or instruction. Then it verifies that the instruction is well-formatted, i.e. it checks the syntax of each line. If it encounters an error, it immediately halts the translation and shows an error message.

Step 2: If there is no error, i.e. if the python instruction or source code is well-formatted then the compiler translates it into its equivalent form in an intermediate language called "Byte code".

Step 3: Byte code is then sent to the Python Virtual Machine which is the python interpreter. PVM converts the python byte code into machine-executable code. If an error occurs during this interpretation then the conversion is halted with an error message.

## APPROACH IN PYTHON

**Step-1**: Start with a “Hello World” Program
If you happened to learn some programming languages, then I am sure you are aware of what I am talking about. The “Hello World” program is like a tradition in the developer community. If you want to master any programming language, this should be the very first line of code we should be seeking for.

**Simple Hello World Program in Python:**

```
print("Hello World")
```

**Step-2:** Start learning about variables
Now once we have mastered the “Hello World” program in Python, the next step is to master variables in python. Variables are like containers that are used to store values.

**Variables in Python:**

```
my_var = 100
```

As you can see here, we have created a variable named “my_var” to assign a value 100 to the same.

**Step-3:** Start learning about Data Types and Data Structures
The next outpost is to learn about data types. Here I have seen that there is a lot of confusion between data types and data structures. The important thing to keep in mind here is that data types represent the type of data. For example. in Python, we have something like int, string, float, etc. Those are called data types as they indicate the type of data we are dealing with.

While data structures are responsible for deciding how to store this data in a computer’s memory.

**String data type in Python:**

```
my_str = "ABCD"
```

**Data Structure in Python:**

```
my_dict={1:100,2:200,3:300}
```

This is known as a dictionary data structure in Python.Again this is just the tip of the iceberg. There are lots of data types and data structures in Python. To give a basic idea about data structures in Python, here is the complete list:

- .Lists
- Dictionary
- Sets
- Tuples
- Frozen set

**Step-4**: Start learning about conditionals and loops
In any programming language, conditionals and loops are considered one of the backbone.

Python is no exception for that as well. This is one of the most important concepts that we need to master.

**IF-ELIF-ELSE conditionals:**

```
if(x < 10):

   print("x is less than 10")

elif(x > 10):

   print("x is greater than 10")

else:

   print("Do nothing")
```

As you can see in the above example, we have created what is known as the if-elif-else ladder

**For loop:**

```
for i in "Python":
    print(i)
```

The above code is basically an example of for loop in python.

## INTRODUCTION TO MACHINE LEARNING

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition**,** speech recognition**,** email filtering**,** Facebook auto-tagging**,** recommender system, and many more.

This machine learning tutorial gives you an introduction to machine learning along with the wide range of machine learning techniques such as Supervised**,** Unsupervised**,** and Reinforcement learning. You will learn about regression and classification models, clustering methods, hidden Markov models, and various sequential models.

## WORKING OF MACHINE LEARNING

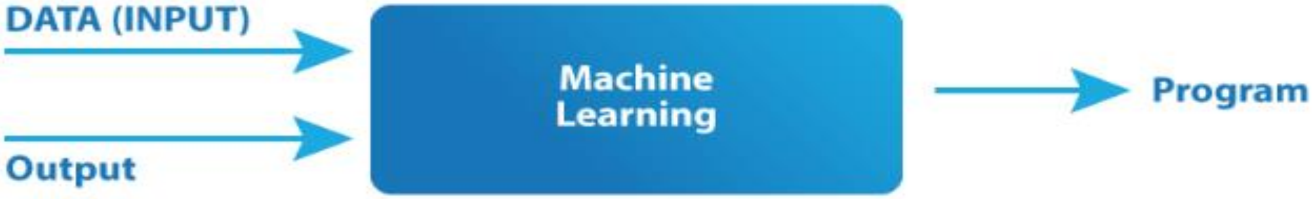


**Figure 4.4 Working of machine learning**

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**.** The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately. Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem

## FEATURES OF MACHINE LEARNING

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

## CLASSIFICATION OF MACHINE LEARNING

At a broad level, machine learning can be classified into three types:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

## SUPERVISED LEARNING:

Supervised learning is a type of machine learning method in which we provide sample label data to the machine learning system in order to train it, and on that basis, it predicts the output. The system creates a model using label data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not. The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is spam filtering**.**

Supervised learning can be grouped further in two categories of algorithms:

- Classification
- Regression

## UNSUPERVISED LEARNING:

Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been label, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns. In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

- Clustering
- Association

## REINFORCEMENT LEARNING:

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

## RANDOM FOREST ALGORITHM

Random Forest is a popular machine learning algorithm that belongs to name suggests, Random Forest is a classifier that contains a number of decision trees the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. On the various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset, instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

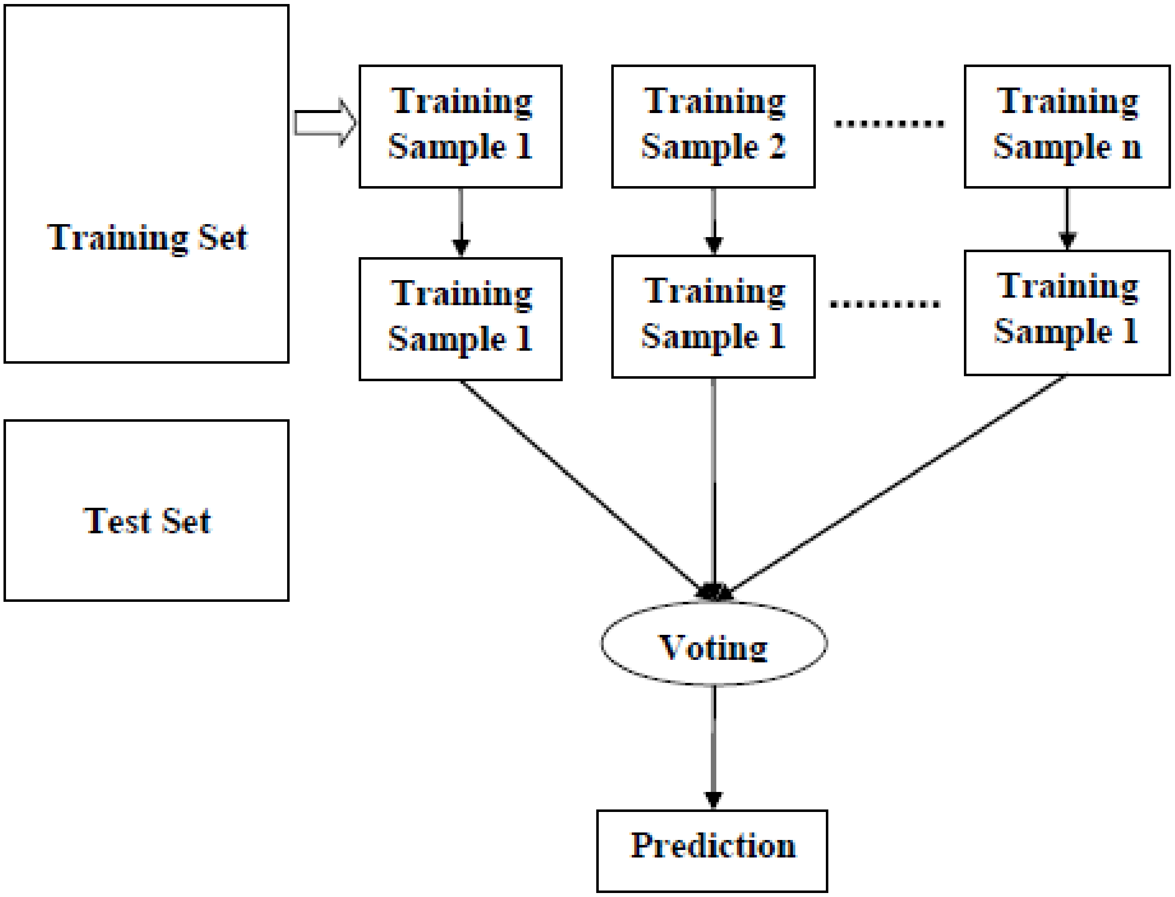


**Figure 4.5 Random Forest Algorithm**

**WORKING OF RANDOM FOREST ALGORITHM:**

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

**Step-1**: Select random K data points from the training set.

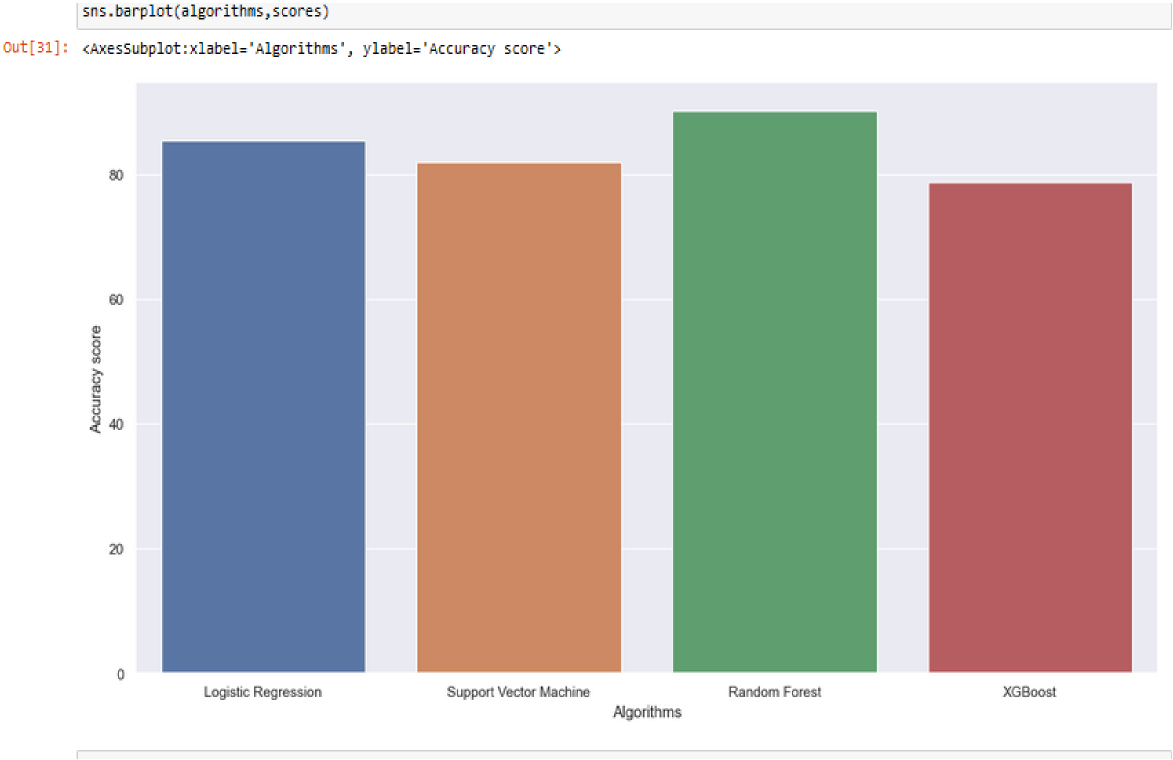
**Step-2:** Build the decision trees associated with the selected data points (Subsets).

**Step-3:** Choose the number N for decision trees that you want to build.

**Step-4:** Repeat Step 1 & Step 2.

**Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Overall, Random Forest algorithm combines multiple decision trees to improve the predictive performance and reduce over fitting. By randomly sampling the training data and features, Random Forest creates a diverse set of decision trees that capture different aspects of the data. By aggregating the predictions of all decision trees, Random Forest provides a robust and accurate prediction.



```
sns.barplot(algorithms,scores)
```

```
Out[31]: <AxesSubplot:xlabel='Algorithms', ylabel='Accuracy score'>
```

**Figure 4.6 Comparison of algorithm**

## ADVANTAGES OF RANDOM FOREST:

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the over fitting issue.

# CHAPTER 5

# SYSTEM DESIGN

System design is the process architecture, components, modules, interfaces and data for a system to satisfy specified requirements. System design could be seen as the application of systems theory to product development of those components and the data go through that system. It is meant to satisfy specific needs and requirements of a business or organization through the engineering of a coherent and well-running system.

## 5.1 ARCHITECTURE DIAGRAM

An architectural diagram is a diagram of a system that is used to abstract the overall outline of the software system and the relationships, constraints, and boundariesbetween components. It provides an overall view of the physical deployment of the software system and its evolution roadmap.

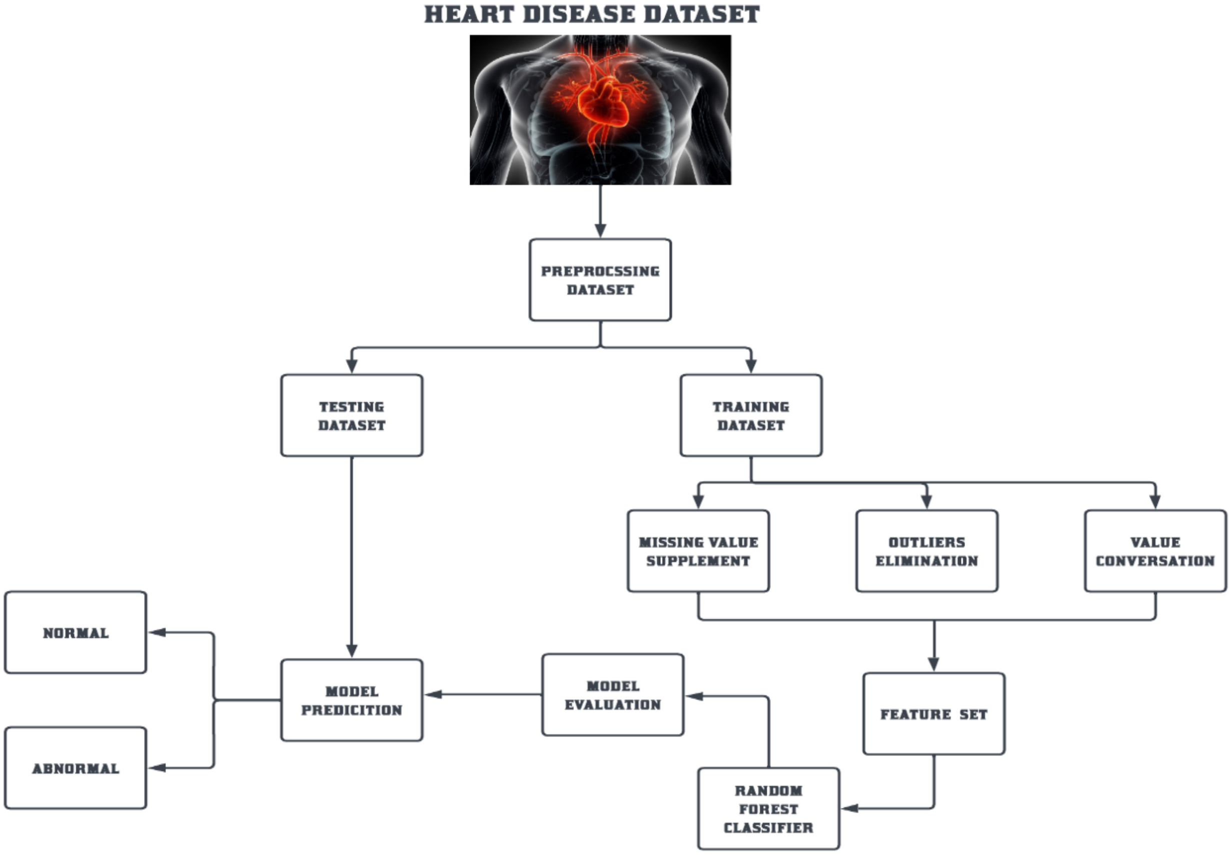


**Figure 5.1 Architecture Diagram**

## 5.2 USE CASE DIAGRAM:

Unified Modeling Language is a standardized general-purpose modeling language in the field of software engineering. The standard is managed and was created by the Object Management Group. UML includes a set of graphic notation techniques to create visual models of software intensive systems. This language is used to specify, visualize, modify, construct and document the artifacts of an object-oriented software intensive system under development.

A Use case Diagram is used to present a graphical overview of the functionality provided by a system in terms of actors, their goals and any dependencies between those use cases. Use case diagram consists of two parts:

**Use case:** A use case describes a sequence of actions that provided something of measurable value to an actor and is drawn as a horizontal ellipse.

**Actor:** An actor is a person, organization or external system that plays a role in one or more interaction with the system.

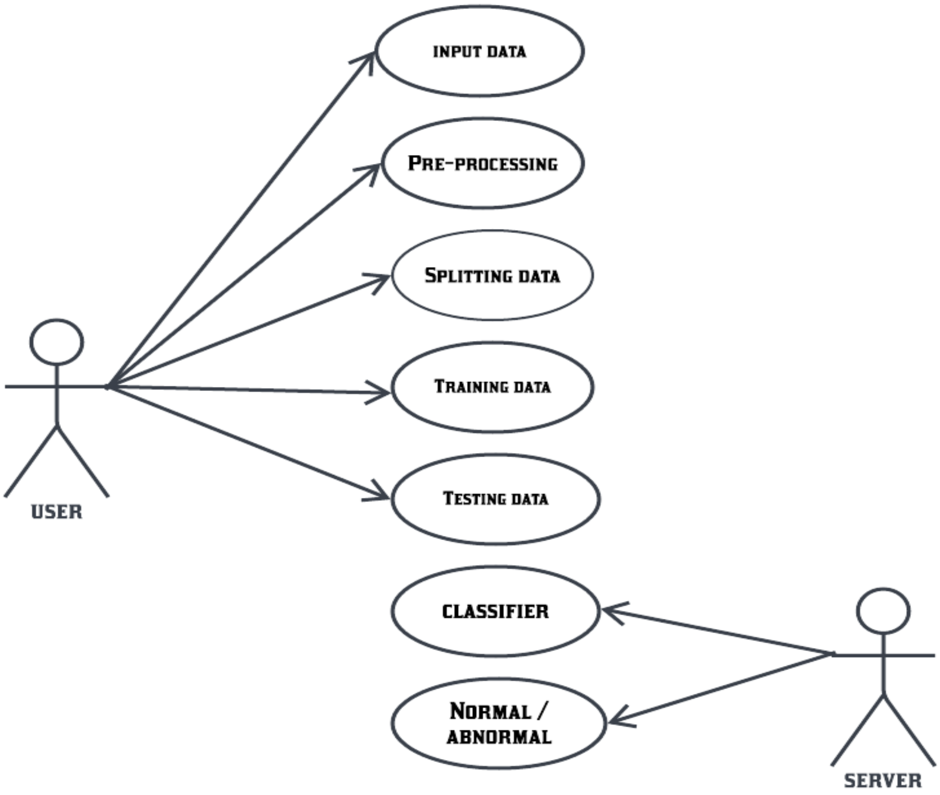


**Figure 5.2 Use Case Diagram**

## 5.3 ACTIVITY DIAGRAM

Activity diagram is a graphical representation of workflows of stepwise activities and actions with support for choice, iteration and concurrency. An  activity diagramshows the overall flow of control.

The most important shape types:

- Rounded rectangles represent activities.
- Diamonds represent decisions.
- Bars represent the start or end of concurrent activities.
- A black circle represents the start of the workflow.
- An encircled circle represents the end of the workflow.

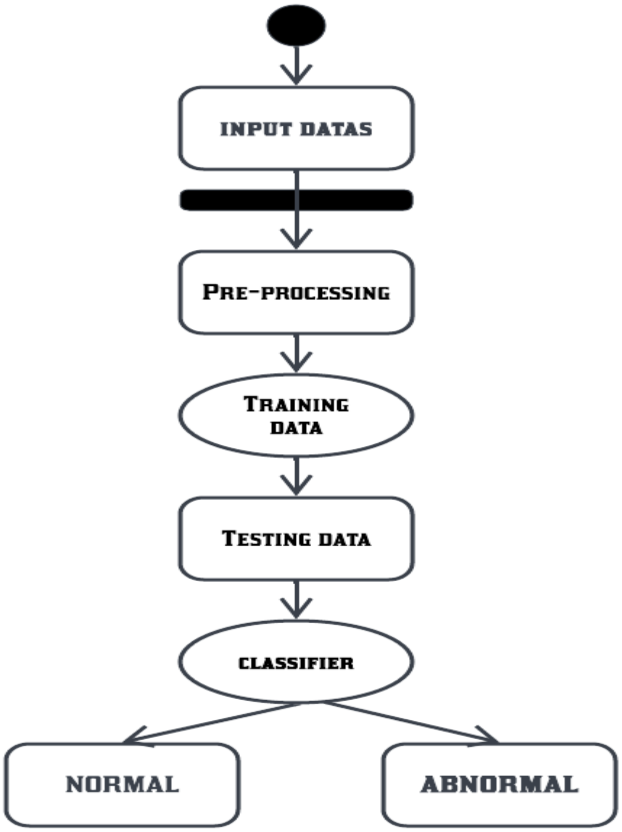


**Figure 5.3 Activity Diagram**

## 5.4 DATA FLOW DIAGRAM

The flow of data of a system or a process is represented by DFD. It also gives insight into the inputs and outputs of each entity and the process itself. DFD does not have control flow and no loops or decision rules are present. Specific operations depending on the type of data can be explained by a flowchart.

It is a graphical tool, useful for communicating with users ,managers and other personnel. it is useful for analyzing existing as well as proposed system.

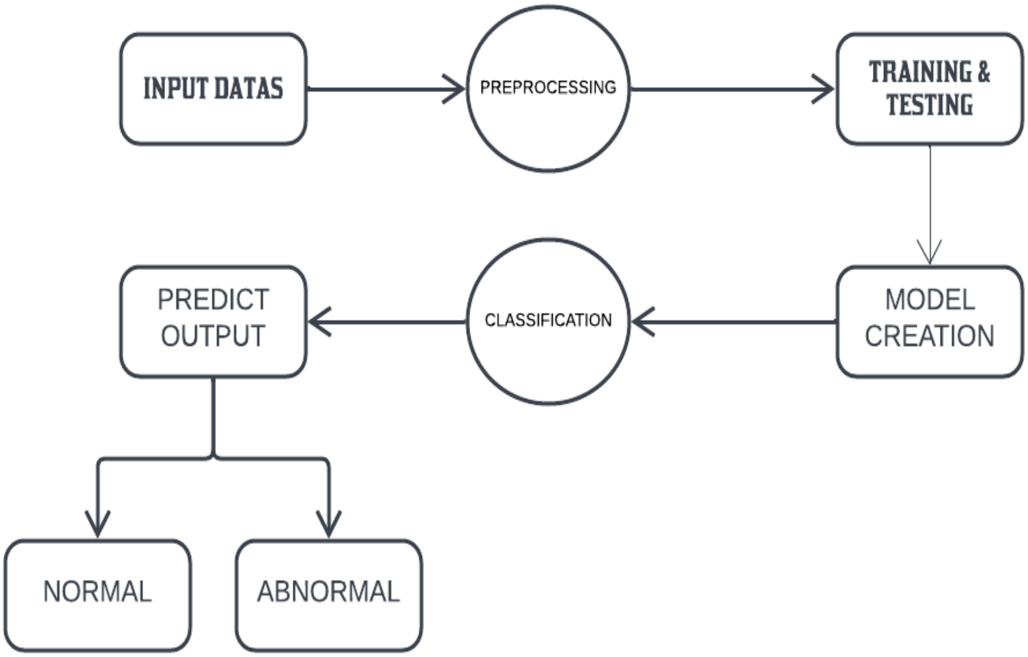


**Figure 5.4 Data Flow Diagram**

# CHAPTER 6
# SYSTEM IMPLEMENTATION

## 6.1 MODULES IMPLEMENTATION

A modular design reduces complexity, facilities change (a critical aspect of software maintainability), and results in easier implementation by encouraging parallel development of different part of system. Software with effective modularity is easier to develop because function may be compartmentalized and interfaces are simplified. Software architecture embodies modularity that is software is divided into separately named and addressable components called modules that are integrated to satisfy problem requirement.

| | |
|---|---|
| **1.** | Dataset Collection |
| **2.** | Data Pre-Processing |
| **3.** | Feature selection and reduction |
| **4.** | Classification model |
| **5.** | Prediction using Random Forest |

**Table 6.1  List of Modules**

## 6.2 MODULE DESCRIPTION

**DATA COLLECTION:**

- Heart diseases Dataset downloaded from Kaggle Website. The dataset have a 14 features column and 300 patient reports.
- The features are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, ECG result, maximum heart rate, ST depression, number of major vessels, thallium stress result and final column is target.
- The value of target is 1 and 0, if target value is 1 the certain person will have chances of affected by heart diseases or if target value is 0 the certain person will not have chances of heart diseases.

**DATA PRE-PROCESSING:**

- Heart disease data is pre-processed after collection of various records.
- The dataset contains a total of patient records, where records are with some missing values.
- Those records have been removed from the dataset and the remaining patient records are used in pre-processing.
- The multiclass variable and binary classification are introduced for the attributes of the given dataset. The multi-class variable is used to check the presence or absence of heart disease.
- In the instance of the patient having heart disease, the value is set to indicating the absence of heart disease in the patient.
- The pre-processing of data is carried out by converting medical records into diagnosis values.
- The results of data pre-processing for patient records indicate that records show the value of establishing the presence of heart disease while the remaining reflected the value of 0 indicating the absence of heart disease and 1 indicating present of heart diseases.

**FEATURE SELECTION AND REDUCTION:**

- From among the attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient.
- The remaining attributes are considered important as they contain vital clinical records.
-

- Clinical records are vital to diagnosis and learning the severity of heart disease.
- As previously mentioned in this experiment, convolutional neural network used, we proposed a Random Forest Algorithm.
- The experiment was repeated with all the ML techniques using all 13 attribute.

## CLASSIFICATION MODEL:

- The clustering of datasets is done on the basis of the variables and criteria of Random Forest (RF) features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance.
- The best performing models are identified from the above results based on their low rate of error.
- The performance is further optimized by choosing the RF cluster with a high rate of error and extraction of its corresponding classifier features.
- The performance of the classifier is evaluated for error optimization on this data set.

## PREDICTION USING RANDOM FOREST:

- The results are generated by applying the classification rule for the dataset.
- The classification rules generated based on the rule after data pre-processing is done.
- After pre-processing, there are four best ML techniques are chosen to train the data's and the results are generated.
- The dataset with RF, XGBoost, Logistic Regression are applied to find out the best classification method.
- The results show that RF are the best algorithm for predict Heart Diseases.
- The RF accuracy rate is high compared to the other algorithms.
- Finally, prediction process has done using trained random forest model.

# CHAPTER 7

# TESTING

## 7.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

## 7.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfied as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

## 7.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

## 7.4 SYSTEM TESTING

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

### 7.5 WHITE BOX TESTING

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

### 7.6 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document. It is a testing in which the software under test is treated. The test provides inputs and responds to outputs without considering how the software works.

# CHAPTER 8
# CONCLUSION AND FUTURE ENHANCEMENT

## 8.1 CONCLUSION

Healthcare around the world is committed to providing quality care to patients via electronic health records. Due to the distributed nature of the EHRs, shared access to health records should be made possible and data integration should be established. Preserving the privacy of patient information is an important consideration when handling medical data. We have developed a privacy-preserving integration model based on association rules for predicting heart disease using patient data collected from horizontally distributed databases. Our model allows the sharing of data summaries (useful information) to be used to predict heart disease. These summaries are not accompanied by private patient information. Our approach is the first to use association rules metrics formally distributed medical datasets to generate weighted rules, which are further generalized using independent test datasets rather than using specific rules for each local model.

## 8.2 FUTURE ENHANCEMENT

In future work, we will utilize different highlights determination calculations, improvement strategies to additional expansion and the exhibition of a prescient framework for HD conclusion. The controlling and treatment of infection is important after determination subsequently. We will predict the accurate treatment and recuperation of infections in future for basic sickness (for example, heart, bosom, Parkinson, diabetes).

# APPENDIX- 1

## SAMPLE CODING

### app.py:

```
from flask import Flask, render_template, request
import pickle
import numpy as np

filename = 'heart.pkl'
model = pickle.load(open(filename, 'rb'))

app = Flask(__name__)

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

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

        age = int(request.form['age'])
        sex = request.form.get('sex')
        cp = request.form.get('cp')
        trestbps = int(request.form['trestbps'])
        chol = int(request.form['chol'])
        fbs = request.form.get('fbs')
        restecg = int(request.form['restecg'])
        thalach = int(request.form['thalach'])
    exang = request.form.get('exang')
    oldpeak = float(request.form['oldpeak'])
    slope = request.form.get('slope')
    ca = int(request.form['ca'])
    thal = request.form.get('thal')

    data = np.array([[age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal]])
    my_prediction = model.predict(data)

    return render_template('result.html', prediction=my_prediction)

if __name__ == '__main__':
        app.run(host='0.0.0.0',port=700)
```

## Train-RF..iynb:

```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 1,
   "id": "3b90bcf0",
   "metadata": {},
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "['.ipynb_checkpoints', 'app.py', 'heart.csv', 'heart.pkl', 'static', 'templates', 'Train - RF.ipynb', 'Untitled.ipynb']\n"
     ]
    }
   ],
   "source": [
    "import numpy as np\n",
    "import pandas as pd\n",
    "import matplotlib.pyplot as plt\n",
    "import seaborn as sns\n",
    "\n",
    "%matplotlib inline\n",
    "\n",
    "import os\n",
    "print(os.listdir())\n",
    "\n",
    "import warnings\n",
    "warnings.filterwarnings('ignore')"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 2,
   "id": "0e7f6e85",
   "metadata": {},
   "outputs": [],
   "source": [
    "dataset = pd.read_csv(\"heart.csv\")"
   ]
  },
  {
   "cell_type": "code",
"execution_count": 3,
"id": "44d4ba20",
"metadata": {},
"outputs": [
 {
  "data": {
   "text/html": [
    "<div>\n",
    "<style scoped>\n",
    "    .dataframe tbody tr th:only-of-type {\n",
    "        vertical-align: middle;\n",
    "    }\n",
    "\n",
    "    .dataframe tbody tr th {\n",
    "        vertical-align: top;\n",
    "    }\n",
    "\n",
    "    .dataframe thead th {\n",
    "        text-align: right;\n",
    "    }\n",
    "</style>\n",
    "<table border=\"1\" class=\"dataframe\">\n",
    "  <thead>\n",
    "    <tr style=\"text-align: right;\">\n",
    "      <th></th>\n",
    "      <th>age</th>\n",
    "      <th>sex</th>\n",
    "      <th>cp</th>\n",
    "      <th>trestbps</th>\n",
    "      <th>chol</th>\n",
"      <th>fbs</th>\n",
"      <th>restecg</th>\n",
"      <th>thalach</th>\n",
"      <th>exang</th>\n",
"      <th>oldpeak</th>\n",
"      <th>slope</th>\n",
"      <th>ca</th>\n",
"      <th>thal</th>\n",
"      <th>target</th>\n",
"    </tr>\n",
"  </thead>\n",
"  <tbody>\n",
"    <tr>\n",
"      <th>count</th>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"      <td>303.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"    <th>mean</th>\n",
"    <td>54.366337</td>\n",
"    <td>0.683168</td>\n",
"    <td>0.966997</td>\n",
"    <td>131.623762</td>\n",
"    <td>246.264026</td>\n",
"    <td>0.148515</td>\n",
"    <td>0.528053</td>\n",
"    <td>149.646865</td>\n",
"    <td>0.326733</td>\n",
"    <td>1.039604</td>\n",
"    <td>1.399340</td>\n",
"    <td>0.729373</td>\n",
"    <td>2.313531</td>\n",
"    <td>0.544554</td>\n",
"  </tr>\n",
"  <tr>\n",
"    <th>std</th>\n",
"    <td>9.082101</td>\n",
"    <td>0.466011</td>\n",
"    <td>1.032052</td>\n",
"    <td>17.538143</td>\n",
"    <td>51.830751</td>\n",
"    <td>0.356198</td>\n",
"    <td>0.525860</td>\n",
"    <td>22.905161</td>\n",
"    <td>0.469794</td>\n",
"    <td>1.161075</td>\n",
"    <td>0.616226</td>\n",
"    <td>1.022606</td>\n",
"      <td>0.612277</td>\n",
"      <td>0.498835</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>min</th>\n",
"      <td>29.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>94.000000</td>\n",
"      <td>126.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>71.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>25%</th>\n",
"      <td>47.500000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>120.000000</td>\n",
"      <td>211.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>133.500000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>2.000000</td>\n",
"      <td>0.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>50%</th>\n",
"      <td>55.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>130.000000</td>\n",
"      <td>240.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>153.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>0.800000</td>\n",
"      <td>1.000000</td>\n",
"      <td>0.000000</td>\n",
"      <td>2.000000</td>\n",
"      <td>1.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>75%</th>\n",
"      <td>61.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>2.000000</td>\n",
"      <td>140.000000</td>\n",
"      <td>274.500000</td>\n",
"      <td>0.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>166.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>1.600000</td>\n",
"      <td>2.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>3.000000</td>\n",
"      <td>1.000000</td>\n",
"    </tr>\n",
"    <tr>\n",
"      <th>max</th>\n",
"      <td>77.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>3.000000</td>\n",
"      <td>200.000000</td>\n",
"      <td>564.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>2.000000</td>\n",
"      <td>202.000000</td>\n",
"      <td>1.000000</td>\n",
"      <td>6.200000</td>\n",
"      <td>2.000000</td>\n",
"      <td>4.000000</td>\n",
"      <td>3.000000</td>\n",
"      <td>1.000000</td>\n",
"    </tr>\n",
"  </tbody>\n",
"</table>\n",
 "</div>"
],
"text/plain": [
 "            age         sex          cp    trestbps        chol         fbs  \\\n",
 "count  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000   \n",
 "mean    54.366337    0.683168    0.966997  131.623762  246.264026    0.148515   \n",
 "std      9.082101    0.466011    1.032052   17.538143   51.830751    0.356198   \n",
 "min     29.000000    0.000000    0.000000   94.000000  126.000000    0.000000   \n",
 "25%     47.500000    0.000000    0.000000  120.000000  211.000000    0.000000   \n",
 "50%     55.000000    1.000000    1.000000  130.000000  240.000000    0.000000   \n",
 "75%     61.000000    1.000000    2.000000  140.000000  274.500000    0.000000   \n",
 "max     77.000000    1.000000    3.000000  200.000000  564.000000    1.000000   \n",
 "\n",
 "          restecg     thalach       exang     oldpeak       slope          ca  \\\n",
 "count  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000   \n",
 "mean     0.528053  149.646865    0.326733    1.039604    1.399340    0.729373   \n",
 "std      0.525860   22.905161    0.469794    1.161075    0.616226    1.022606   \n",
 "min      0.000000   71.000000    0.000000    0.000000    0.000000    0.000000   \n",
 "25%      0.000000  133.500000    0.000000    0.000000    1.000000    0.000000   \n",
 "50%      1.000000  153.000000    0.000000    0.800000    1.000000    0.000000   \n",
 "75%      1.000000  166.000000    1.000000    1.600000    2.000000    1.000000   \n",
 "max      2.000000  202.000000    1.000000    6.200000    2.000000    4.000000   \n",
 "\n",
 "             thal      target  \n",
 "count  303.000000  303.000000  \n",
 "mean     2.313531    0.544554  \n",
 "std      0.612277    0.498835  \n",
 "min      0.000000    0.000000  \n",
 "25%      2.000000    0.000000  \n",
 "50%      2.000000    1.000000  \n",
     "75%      3.000000    1.000000  \n",
     "max      3.000000    1.000000  "
    ]
   },
   "execution_count": 3,
   "metadata": {},
   "output_type": "execute_result"
  }
 ],
 "source": [
  "dataset.describe()"
 ]
},
{
 "cell_type": "code",
 "execution_count": 4,
 "id": "bc24c140",
 "metadata": {},
 "outputs": [
  {
   "name": "stdout",
   "output_type": "stream",
   "text": [
    "<class 'pandas.core.frame.DataFrame'>\n",
    "RangeIndex: 303 entries, 0 to 302\n",
    "Data columns (total 14 columns):\n",
    " #   Column    Non-Null Count  Dtype  \n",
    "---  ------    --------------  -----  \n",
    " 0   age       303 non-null    int64  \n",
    " 1   sex       303 non-null    int64  \n",
    " 2   cp        303 non-null    int64  \n",
    " 3   trestbps  303 non-null    int64  \n",
    " 4   chol      303 non-null    int64  \n",
    " 5   fbs       303 non-null    int64  \n",
    " 6   restecg   303 non-null    int64  \n",
    " 7   thalach   303 non-null    int64  \n",
    " 8   exang     303 non-null    int64  \n",
    " 9   oldpeak   303 non-null    float64\n",
    " 10  slope     303 non-null    int64  \n",
    " 11  ca        303 non-null    int64  \n",
    " 12  thal      303 non-null    int64  \n",
    " 13  target    303 non-null    int64  \n",
    "dtypes: float64(1), int64(13)\n",
    "memory usage: 33.3 KB\n"
   ]
  }
 ],
 "source": [
  "dataset.info()"
 ]
},
{
 "cell_type": "code",
 "execution_count": 5,
 "id": "0cf3262c",
 "metadata": {},
 "outputs": [
  {
   "name": "stdout",
   "output_type": "stream",
    "text": [
     "age:\t\t\tage\n",
     "sex:\t\t\t1: male, 0: female\n",
     "cp:\t\t\tchest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic\n",
     "trestbps:\t\t\tresting blood pressure\n",
     "chol:\t\t\t serum cholestoral in mg/dl\n",
     "fbs:\t\t\tfasting blood sugar > 120 mg/dl\n",
     "restecg:\t\t\tresting electrocardiographic results (values 0,1,2)\n",
     "thalach:\t\t\t maximum heart rate achieved\n",
     "exang:\t\t\texercise induced angina\n",
     "oldpeak:\t\t\toldpeak = ST depression induced by exercise relative to rest\n",
     "slope:\t\t\tthe slope of the peak exercise ST segment\n",
     "ca:\t\t\tnumber of major vessels (0-3) colored by flourosopy\n",
     "thal:\t\t\tthal: 3 = normal; 6 = fixed defect; 7 = reversable defect\n"
    ]
   }
  ],
  "source": [
   "info = [\"age\",\"1: male, 0: female\",\"chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic\",\"resting blood pressure\",\" serum cholestoral in mg/dl\",\"fasting blood sugar > 120 mg/dl\",\"resting electrocardiographic results (values 0,1,2)\",\" maximum heart rate achieved\",\"exercise induced angina\",\"oldpeak = ST depression induced by exercise relative to rest\",\"the slope of the peak exercise ST segment\",\"number of major vessels (0-3) colored by flourosopy\",\"thal: 3 = normal; 6 = fixed defect; 7 = reversable defect\"]\n",
   "\n",
   "\n",
   "\n",
   "for i in range(len(info)):\n",
   "    print(dataset.columns[i]+\":\\t\\t\\t\"+info[i])"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 6,
   "id": "0760ddcd",
   "metadata": {},
   "outputs": [
    {
     "data": {
      "text/plain": [
       "count    303.000000\n",
       "mean       0.544554\n",
       "std        0.498835\n",
       "min        0.000000\n",
       "25%        0.000000\n",
       "50%        1.000000\n",
       "75%        1.000000\n",
       "max        1.000000\n",
       "Name: target, dtype: float64"
      ]
     },
     "execution_count": 6,
     "metadata": {},
     "output_type": "execute_result"
    }
   ],
   "source": [
    "dataset[\"target\"].describe()"
   ]
```

```
},
{
 "cell_type": "code",
 "execution_count": 7,
 "id": "2d8b974b",
 "metadata": {},
 "outputs": [
  {
   "data": {
    "text/plain": [
     "array([1, 0], dtype=int64)"
    ]
   },
   "execution_count": 7,
   "metadata": {},
   "output_type": "execute_result"
  }
 ],
 "source": [
  "dataset[\"target\"].unique()"
 ]
},
{
 "cell_type": "code",
 "execution_count": 8,
 "id": "72089ba9",
 "metadata": {},
 "outputs": [
  {
   "name": "stdout",
   "output_type": "stream",
   "text": [
    "target      1.000000\n",
    "exang       0.436757\n",
    "cp          0.433798\n",
    "oldpeak     0.430696\n",
    "thalach     0.421741\n",
    "ca          0.391724\n",
    "slope       0.345877\n",
    "thal        0.344029\n",
    "sex         0.280937\n",
    "age         0.225439\n",
    "trestbps    0.144931\n",
    "restecg     0.137230\n",
    "chol        0.085239\n",
    "fbs         0.028046\n",
    "Name: target, dtype: float64\n"
   ]
  }
 ],
 "source": [
  "print(dataset.corr()[\"target\"].abs().sort_values(ascending=False))"
 ]
},
```

# APPENDIX- 2

## SNAP SHOTS

### USER INTERFACE

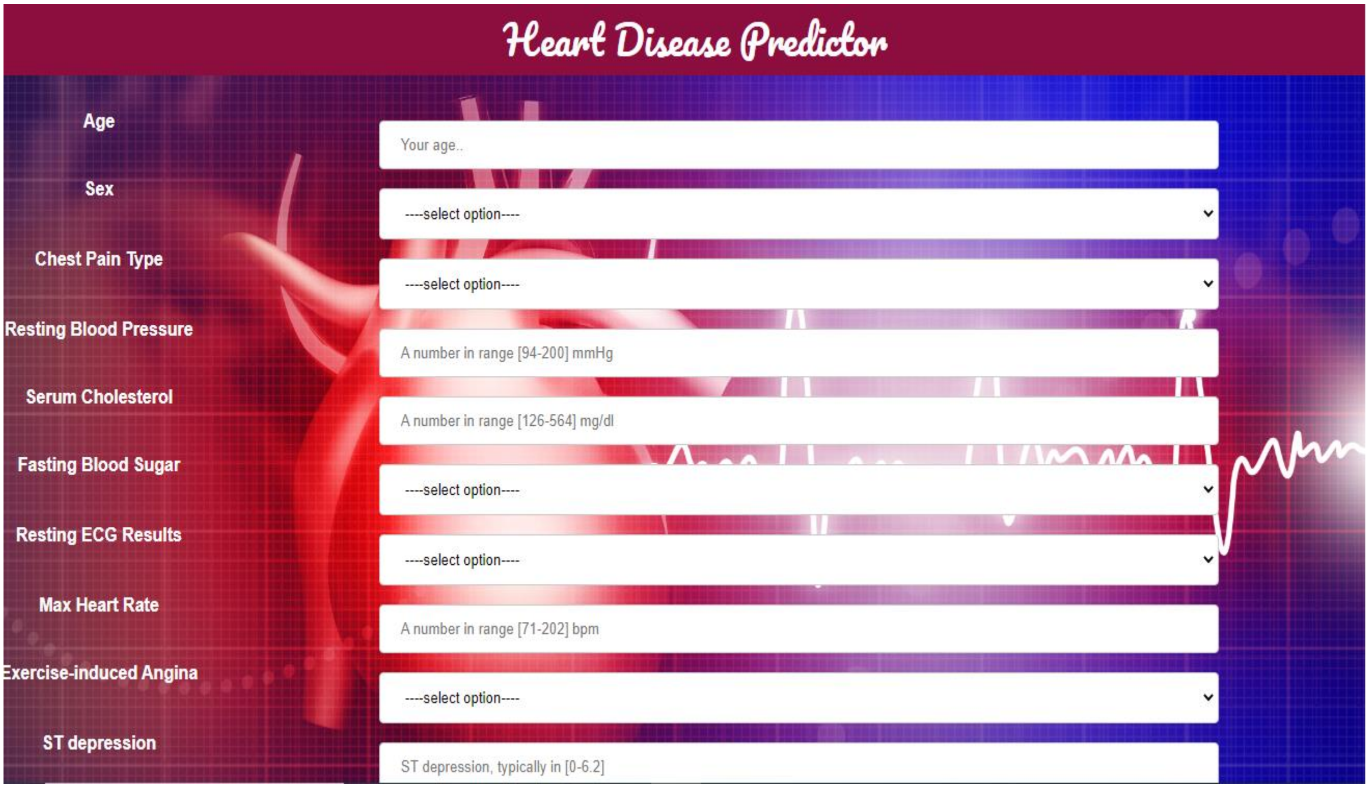
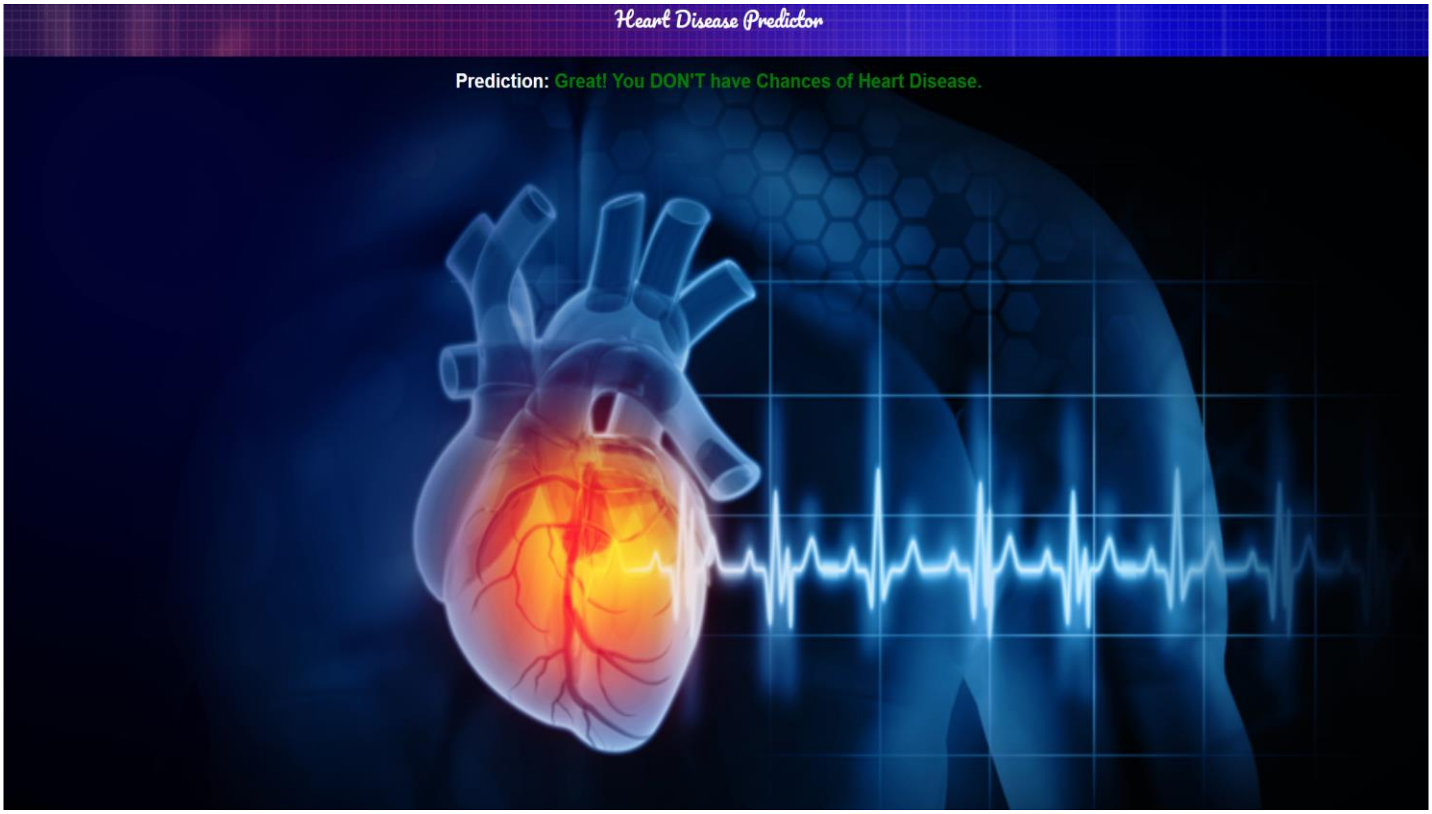


**Figure A.2.1.User Details**

This figure A.2.1 shows the user interface to enter the required details of the user details to know that they have heart disease or not. These are the most important features to accurately predict the heart disease,

## USER INTERFACE - OUTPUT

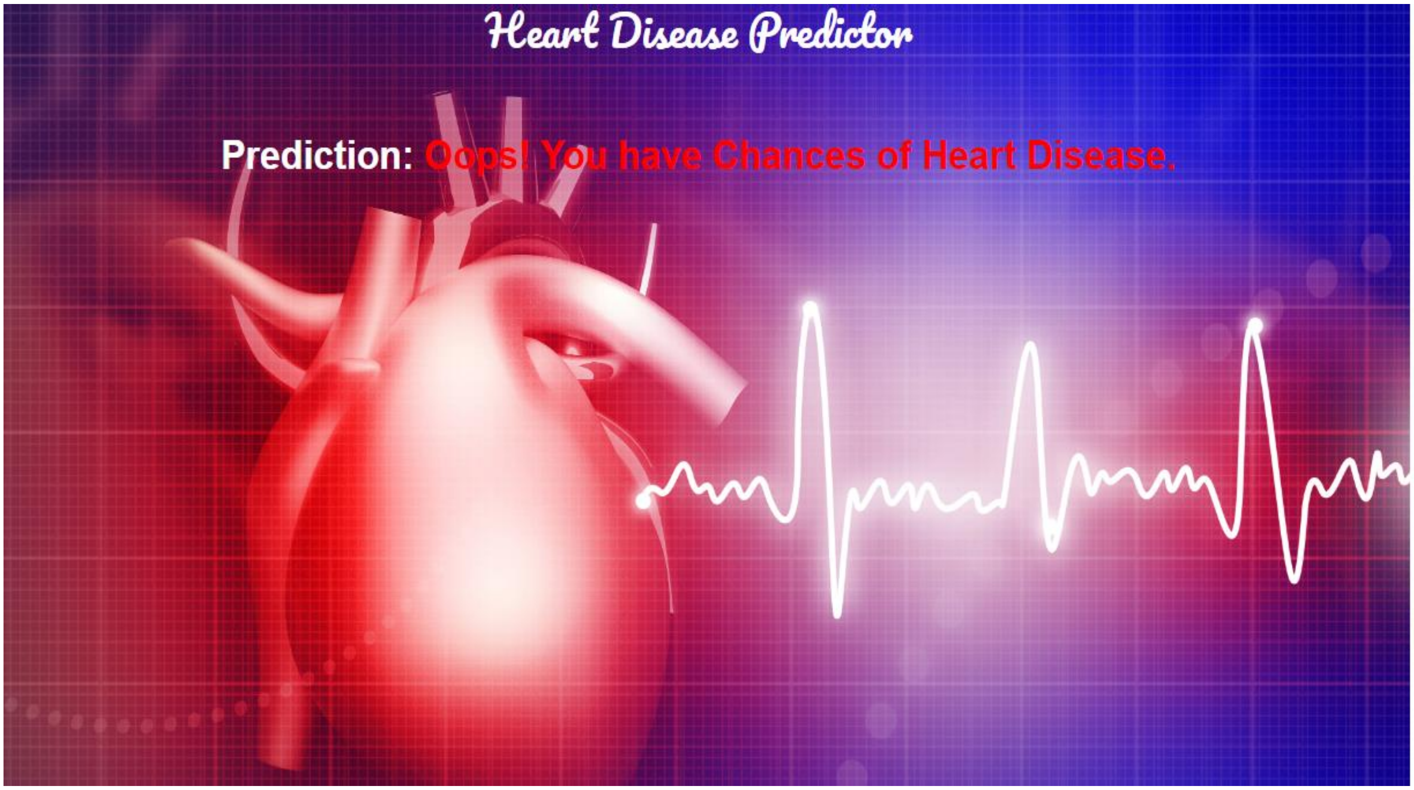


Prediction: Great! You DON'T have Chances of Heart Disease.

**Fig. A.2.2.Prediction Output- Normal**

The above figure A.2.2 shows that the user have no heart disease. It is predicted with the help of training data to identify accurate output.

## USER INTERFACE - OUTPUT



Prediction: Oops! You have Chances of Heart Disease.

**Fig.A.2.3. Prediction Output - Abnormal**

The above figure A.2.2 shows that the user have heart disease. It is predicted with the help of training data to identify accurate output.

# REFERENCES

[1] S. J. Pasha and e. S. Mohamed, ‘‘novel feature reduction (nfr) model with machine learning and data mining algorithms for effective disease risk prediction,’’ ieee access, vol. 8, 2020.

[2] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, ‘‘Classification models for heart disease prediction using feature selection and PCA,’’ Informat. Med. Unlocked, vol. 19, Jan. 2020, Art. no. 100330.

[3] E. Nasarian, M. Abdar, M. A. Fahami, R. Alizadehsani, S. Hussain, M. E. Basiri, M. Zomorodi-Moghadam, X. Zhou, P. Pławiak, U. R. Acharya, R.-S. Tan, and N. Sarrafzadegan, ‘‘Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach,’’ Pattern Recognit. Lett., vol. 133, May 2020

[4] A. Gupta, L. Kumar, R. Jain, and P. Nagrath, ‘‘Heart disease pre-diction using classification (naive bayes),’’ in Proc. 1st Int. Conf. Comput., Commun., Cyber-Secur. (ICS). Singapore: Springer, 2020.

[5] S. Mohan, C. Thirumalai, and G. Srivastava, ‘‘Effective heart disease prediction using hybrid machine learning techniques,’’ IEEE Access, vol. 7,, 2019.

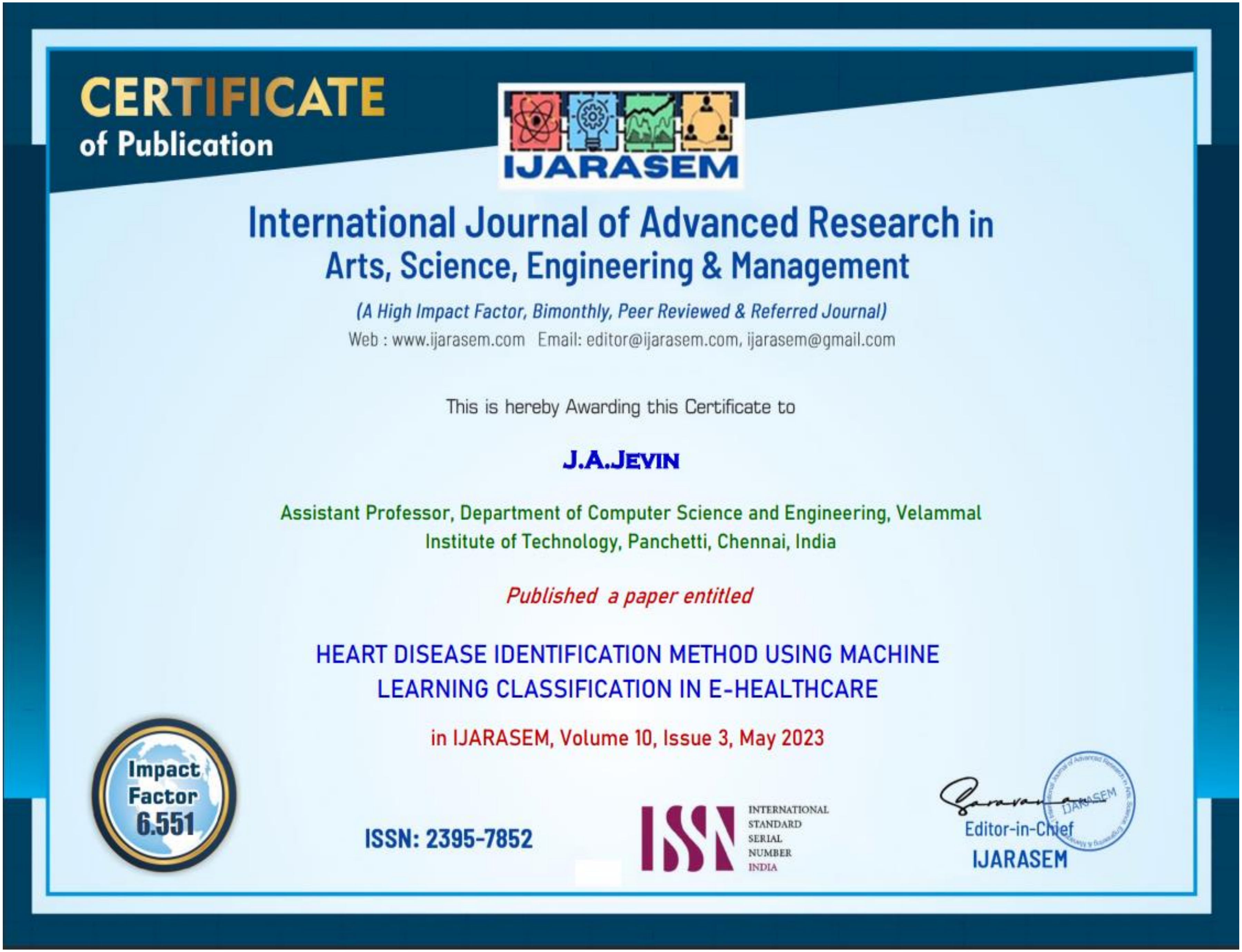
[6] Y. Khan, U. Qamar, N. Yousaf, and A. Khan, ‘‘Machine learning techniques for heart disease datasets: A survey,’’ in Proc. 11th Int. Conf. Mach. Learn. Comput. (ICMLC), Zhuhai, China, 2019.

[7] D. W. Hosmer, S. Lemeshow, and E. D. Cook, Applied Logistic Regression, 2nd ed. New York, NY, USA: Wiley.

[8] S. Goel, A. Deep, S. Srivastava, and A. Tripathi, ‘‘Comparative anal- ysis of various techniques for heart disease prediction,’’ in Proc. 4th Int. Conf. Inf. Syst. Comput. Netw. (ISCON), Mathura, India, Nov. 2019.

[9] R. Atallah and A. Al-Mousa, ‘‘Heart disease detection using machine learning majority voting ensemble method,’’ in Proc. 2nd Int. Conf. new Trends Comput. Sci. (ICTCS), Oct. 2019.

[10] A. Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, ‘‘Machine learning techniques for heart disease prediction,’’ Int. J. Sci. Technol. Res., vol. 8, no. 11,Nov. 2019.



# CERTIFICATE
## of Publication

IJARASEM

International Journal of Advanced Research in Arts, Science, Engineering & Management

*(A High Impact Factor, Bimonthly, Peer Reviewed & Referred Journal)*

Web : www.ijarasem.com Email: editor@ijarasem.com, ijarasem@gmail.com

This is hereby Awarding this Certificate to

**J.A.JEVIN**

Assistant Professor, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

*Published a paper entitled*

HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE

in IJARASEM, Volume 10, Issue 3, May 2023

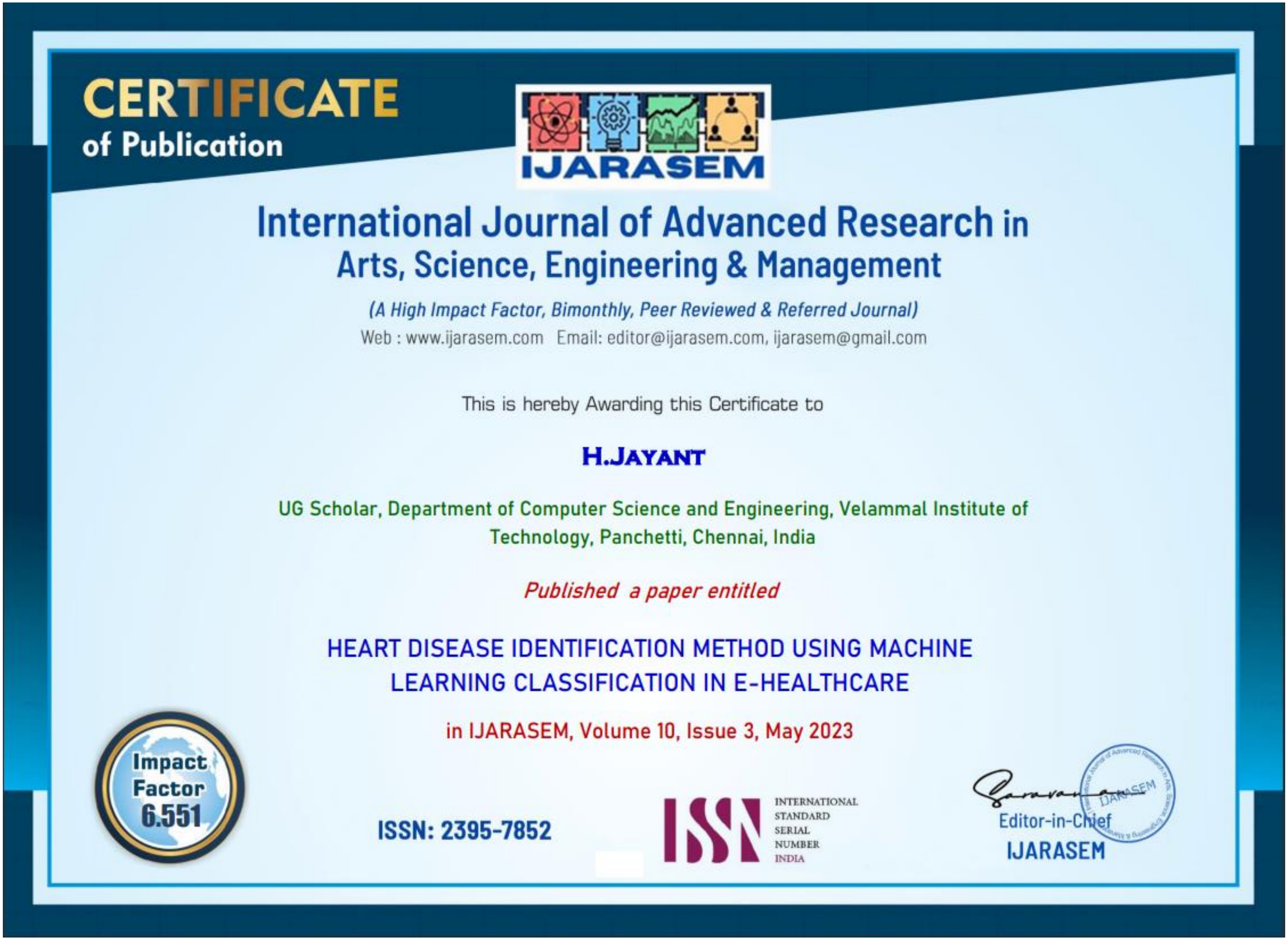
Impact Factor 6.551

ISSN: 2395-7852

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

IJARASEM

Editor-in-Chief
IJARASEM



# CERTIFICATE
of Publication

IJARASEM

## International Journal of Advanced Research in Arts, Science, Engineering & Management

*(A High Impact Factor, Bimonthly, Peer Reviewed & Referred Journal)*

Web : www.ijarasem.com   Email: editor@ijarasem.com, ijarasem@gmail.com

This is hereby Awarding this Certificate to

**H.JAYANT**

UG Scholar, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

*Published  a paper entitled*

HEARTH DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE

in IJARASEM, Volume 10, Issue 3, May 2023

Impact Factor 6.551

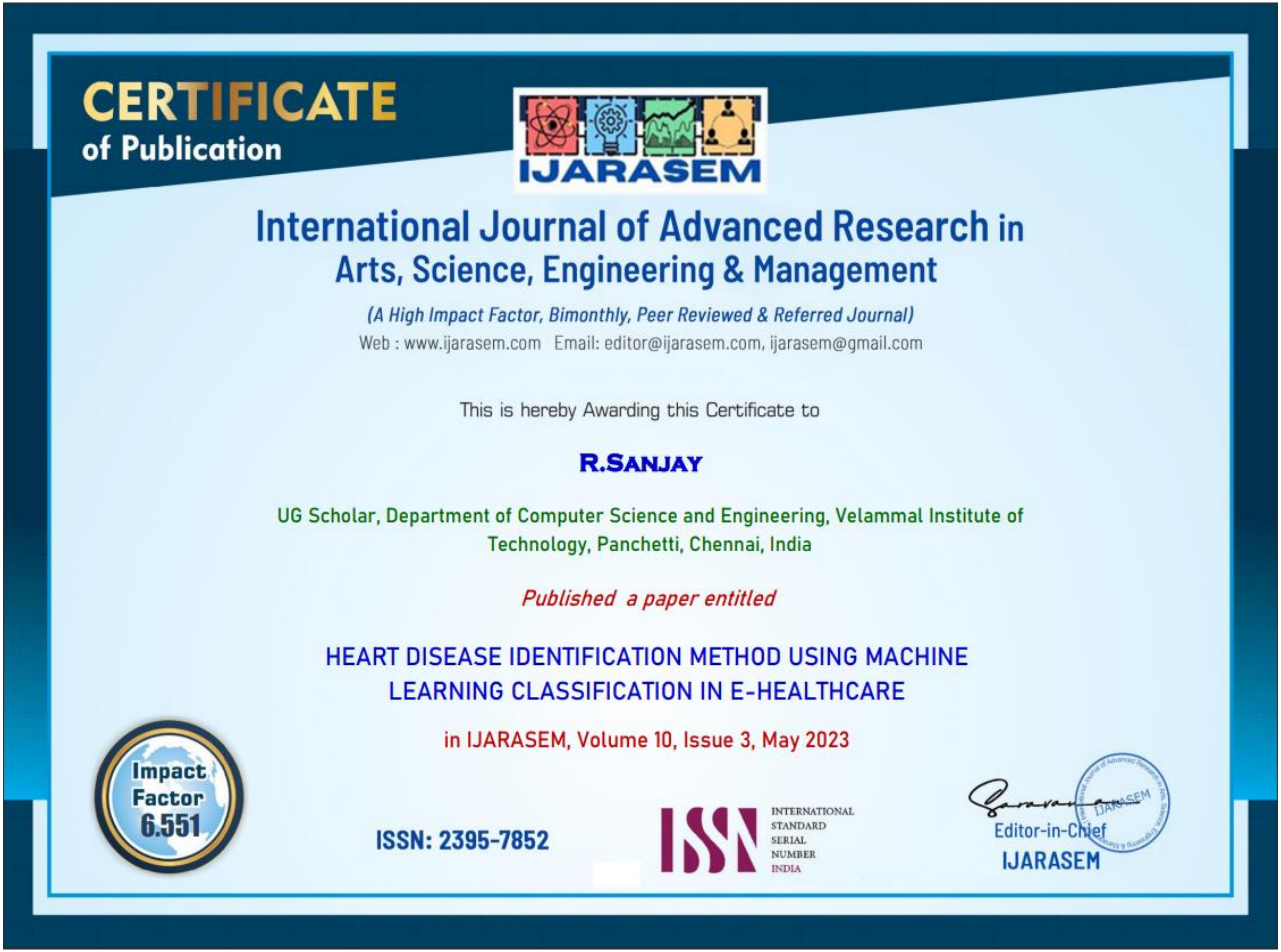
ISSN: 2395-7852

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Saravanan

Editor-in-Chief
IJARASEM

IJARASEM



# CERTIFICATE
of Publication

IJARASEM

## International Journal of Advanced Research in Arts, Science, Engineering & Management

*(A High Impact Factor, Bimonthly, Peer Reviewed & Referred Journal)*

Web : www.ijarasem.com Email: editor@ijarasem.com, ijarasem@gmail.com

This is hereby Awarding this Certificate to

**R.SANJAY**

UG Scholar, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

*Published a paper entitled*

HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE

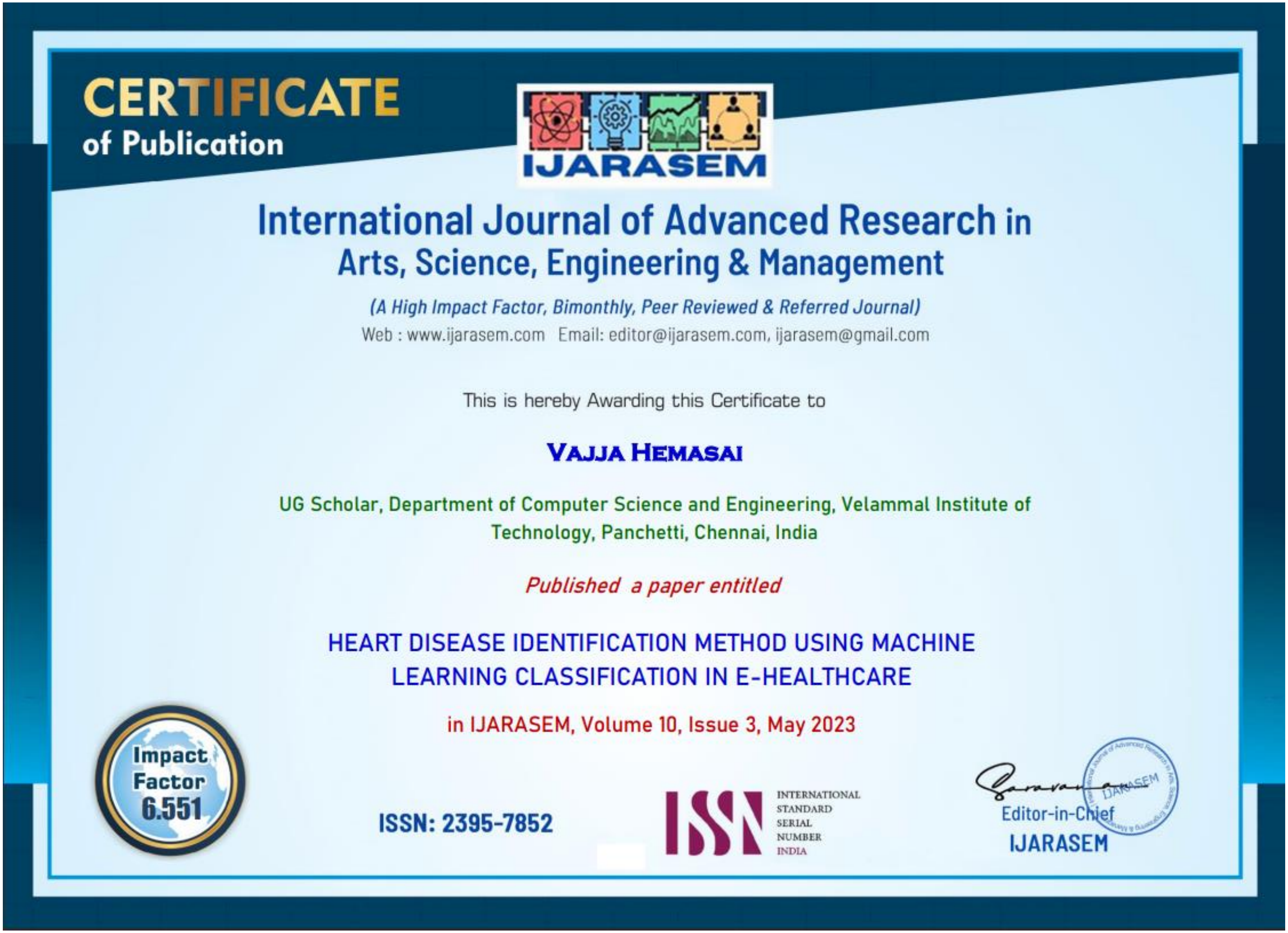
in IJARASEM, Volume 10, Issue 3, May 2023

Impact Factor 6.551

ISSN: 2395-7852

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief
IJARASEM



# CERTIFICATE
of Publication

IJARASEM

## International Journal of Advanced Research in Arts, Science, Engineering & Management

*(A High Impact Factor, Bimonthly, Peer Reviewed & Referred Journal)*

Web : www.ijarasem.com Email: editor@ijarasem.com, ijarasem@gmail.com

This is hereby Awarding this Certificate to

**VAJJA HEMASAI**

UG Scholar, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

*Published a paper entitled*

HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE

in IJARASEM, Volume 10, Issue 3, May 2023

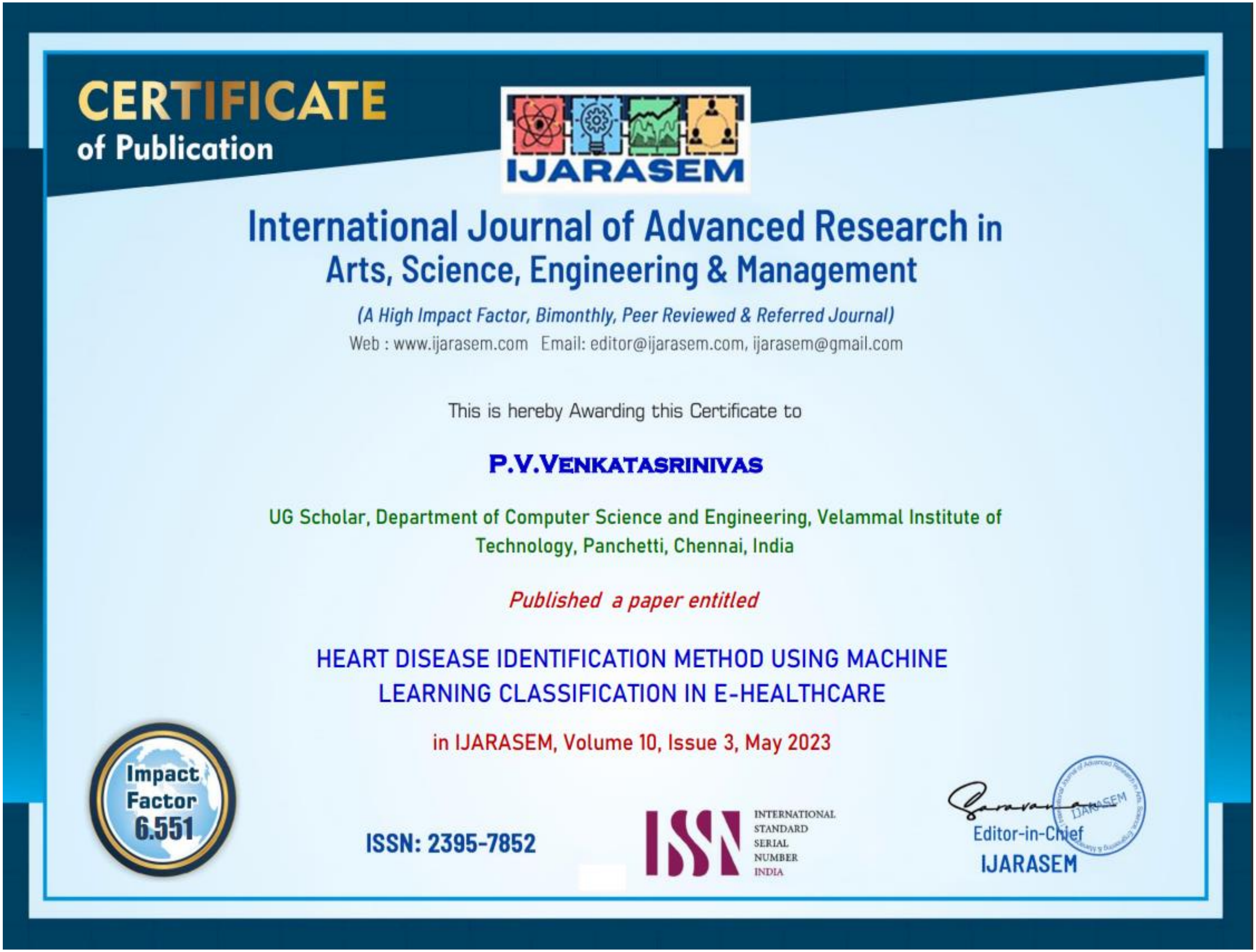
Impact Factor 6.551

ISSN: 2395-7852

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Saravan

Editor-in-Chief
IJARASEM



# CERTIFICATE
of Publication

IJARASEM

## International Journal of Advanced Research in Arts, Science, Engineering & Management

*(A High Impact Factor, Bimonthly, Peer Reviewed & Referred Journal)*

Web : www.ijarasem.com Email: editor@ijarasem.com, ijarasem@gmail.com

This is hereby Awarding this Certificate to

**P.V.VENKATASRINIVAS**

UG Scholar, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

*Published a paper entitled*

HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNING CLASSIFICATION IN E-HEALTHCARE

in IJARASEM, Volume 10, Issue 3, May 2023

Impact Factor 6.551

ISSN: 2395-7852

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Saravanan

Editor-in-Chief

IJARASEM