HYBRID FRAMEWORK TO DETECT HEART DISEASE AND ADVISING PREVENTIVE MEASURES TO PREVENT HEART ATTACK USING MACHINE LEARNING

Thesis Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

in

Computer Applications

By

Rahul Kumar Jha
Registration Number: 41800904

Supervised By

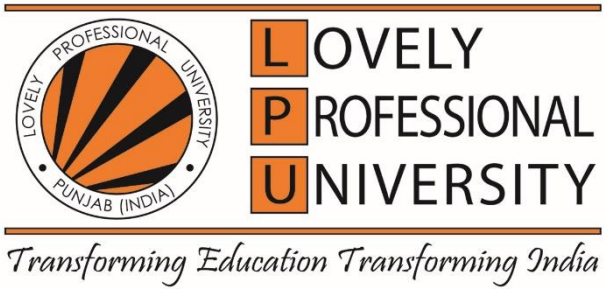
Dr. Santosh Kumar Henge (24372)

Associate Professor
Faculty of Technology and Sciences
School of Computer Applications
Lovely Professional University
Phagwara, Punjab, India

Co-Supervised by

Dr. Ashok Sharma

Assistant Professor
Jammu University
Jammu, India

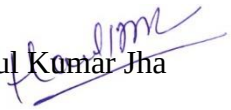


LOVELY PROFESSIONAL UNIVERSITY, PUNJAB
2024

DECLARATION

The work presented in the thesis titled "*Hybrid Framework To Detect Heart Disease And Advising Preventive Measures To Prevent Heart Attack Using Machine Learning*" submitted to the Faculty of Technology and Sciences, Lovely Professional University, Punjab, India, for the award of the degree of Doctor of Philosophy in the subject of Computer Application has been carried out by me under the supervision of Dr. Santosh Kumar Henge, School of Computer Applications, Lovely Professional University, Phagwara and co-supervised by Dr. Ashok Sharma, Department of Computer Science and IT, Bhaderwah Campus, University of Jammu, Jammu.

The thesis documents original piece of work and has not been submitted to any other place fully or partially for the award of any other degree. All the ideas and references have been duly acknowledged.



Rahul Kumar Jha

Research Scholar

Reg. No: 41800904

CERTIFICATE

It is to certify that Mr. Rahul Kumar Jha, Ph.D. Scholar, Department of Computer Application has carried out research work entitled " *Hybrid Framework to Detect Heart Disease and Advising Preventive Measures to Prevent Heart Attack Using Machine Learning* " for the award of degree of Doctor of Philosophy in Computer Application, Faculty of Technology and Sciences, Lovely Professional University, Punjab, India.

Further certify that:

I. The thesis personifies the original effort of the candidate and is not copied from any other sources.

II. The candidate has worked under my supervision throughout the period required under statutes.

III. The candidate has put in the required attendance in the department during the period of research.

IV. The candidate has not submitted this research work to any other institution for any other degree/diploma.

Supervised By

Dr. Santosh Kumar Henge

School of Computer Applications,

Lovely Professional University,

Phagwara

Co-Supervised By

Dr. Ashok Sharma

Department of Computer Science and IT,

Bhaderwah Campus,

University of Jammu, Jammu

ABSTRACT

Over the past decade, heart disease has emerged as one of the most widespread illnesses and has risen to the top of the mortality rate rankings. Unfortunately, heart attacks have impacted a large percentage, specifically 80%, of the population. Heart attack is a severe condition that requires special handling and cure but is often irreversible. Besides its criticality, this disease is also in the list of highly paid treatments and that is the reason that even there are treatments available but due to high cost, people below middle class are unable to avail the facility and many time they even lose their life due to non-availability of good treatment/medicine as well. Myocardial Infarction or heart attack occurs due to deposit of plaque in the inner layer of the artery and starts rupturing and preventing the flow of blood. Due to this, an organ called Ischemia when doesn't get enough blood and oxygen, starts damaging and cause heart attack. Heart attack has severed ill-effect and along with making heart weaker, a patient with heart attack symptoms also tends to live in threat and depression.

Diagnosis of HD within time could save life of millions. Due to the criticality and high cost of this disease, this has been a demanding research topic since decades and numerous distinguished researchers worldwide are enthusiastic about developing a solution to forecast and avert the adverse consequences of this ailment. Despite having many frameworks available which focus on heart disease prediction, there have been many new frameworks designed for disease prediction using different algorithms and dataset using various methods. In addition to forecasting heart disease, many present-day researchers are concentrating on preventing heart attacks with the aspiration of discovering a technique to predict heart attacks, which could be a significant lifesaving advancement for the general population. Artificial Intelligence (AI) with its extensive choice of ML methods, made a notable difference in contributing powerful solution for the prediction of heart disease as well as spotting other dangerous illnesses. Advancement of NN family along with hybrid frameworks has pushed the boundaries to unprecedented level and with the incorporation of data pre-processing techniques

and other supportive methods has made the system capable of human-like prediction power.

Given the significant threat, we introduce a new research study that focuses entirely on predicting heart disease and determining the likelihood of a heart attack. The goal is to enable individuals to take preventative measures before the ailment strikes and causes harm. The idea came from the situation being growing day after day wherein any person whether healthy or sick is fringed by heart attack and feel helpless in this panic situation. There must be a solution which can help him or alert them of this worst situation so that they can take some preventive measure to avoid such situation. The research has based this idea and to commence it lot of study has been done and information was gathered, later based on the gathered information related to key area a framework has been designed to build a system which can fulfil the requirement and come up an optimal solution. The research work has been conducted using hybrid approach powered by neural network supported by genetic algorithm and fuzzy inference system and tested over very popular publicly available Cleveland dataset for HD available in UCI repository [24] having 303 patient records with 76 attributes. Research has been conducted in phases wherein first model has been trained using various traditional ML techniques like SVM, K-NN, LR, DT, RF, NB and compared for the quantitative analysis, thereafter hybrid model has been tuned for finding the prediction having the combination of neural network with genetic algorithm. The result was compared with traditional methods based on the model ACC. In the later part, stage 2, model has been built using FIS that could find the probability of heart attack in the patient. During the research work all the dataset attributes has been analysed and criticality and priority has been defined which has been then used as the baseline for further experiment work used in the measurement of probability of the risk with the help of 13000+ fuzzy rules created based on collected information. Whole system has been integrated together and an API endpoint has been exposed for testing purpose to test the model accuracy leaving the future scope of machine integration with trained model via API endpoint which could be used by general people in their daily routine life.

The entire research work is mainly covered in four objectives wherein objective one mainly covers the information gathering phase including literature review and

comparative analysis of existing work, objective 2 and 3 focus on creation of desired framework, model training, executing experiments and so on, and at last objective 4 includes the testing work and exposing API for further testing. All the objectives has been discussed in seven chapters wherein chapter 1 starts with introduction, chapter 2 covers the literature review which provides the detailed overview of existing related work and studies in heart disease prediction and heart attack probability calculation, chapter 3 covers the research framework, in chapter 4 and 5 discussion about the data collection, experiments and results can be overviewed, chapter 6 discusses the test cases and API integration, and thereafter concluding remarks, summarizing the findings and their implications and avenues for future research are explored in chapter 7 and finally the thesis has been concluded with bibliography and ending words.

ACKNOWLEDGEMENT

It is with immense pleasure and deep gratitude that I extend my appreciation to my esteemed supervisor and Co-Supervisor Dr. Santosh Kumar Henge and Dr. Ashok Sharma respectively, who's with their expert guidance, uncompromising standards of accuracy, unceasing interest, deep understanding of the subject and critical reviews made this road towards a fruitful outcome, as a result of which my study became a rewarding journey. My word of gratitude to every member of the expert panel, faculty members, department head and staff of Block 38, Research Cell at Lovely Professional University. They were always accessible and willing to help whenever I need it.

I am profoundly obliged to my parents, my family and friends and specially my brother Dr. Pranay Jha for always helping and standing in need and for all the enthusiasm they shown during the research work, not only I was curious about the end result but along with my mentors, they were also one of them who were very keen to know about it. I want to thank all for their love, support, understanding, constant inspiration, and prayer.

At last, my immense gratitude to God for always making me faith in me which could lead to the completion of my research work.

Rahul Kumar Jha

ABBREVIATIONS

ACC	Accuracy
AI	Artificial Intelligence
ANFIS	Adaptive Neural Fuzzy Inference System
ANN	Artificial Neural Network
AUC	Area Under Curve
AWS	Amazon Web Services
BP	Blood Pressure
BPM	Beat Per Minute
Chi2	Chi-Square
CHF	Coronary Heart Failure
CHOL	Cholesterol
CNN	Convolutional Neural Network
COA	Centroid of Area
COG	Center of Gravity
COS	Center of Sums
CV	Cross-Validation
CVDs	The Cardiovascular Disease
DNN	Deep Neural Network
ESC	The European Society of Cardiology
FIS	Fuzzy Inference System
FN	False Negative
FP	False Positive

FPR	False Positive Rate
GA	Genetic Algorithm
GANN	Genetic Algorithm with Neural Network
GNFIS	Genetical Neural Fuzzy Inference System
GSO	Galactic Swarm Optimization
HD	Heart Disease
HF	Heart Failure
HR	Heart Rate
HRV	Heart Rate Variability
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IaC	Infrastructure-as-Code
KNN	K-Nearest Neighbor
LDA	Latent Dirichlet Allocation
LDL	Low-Density Lipoprotein
LOOCV	Leave One Out Cross-validation
LR	Logistic Regression
MF	Membership Function
MI	Myocardial Infarction
MKL	Multiple Kernel Learning
ML	Machine Learning
MRMR	Minimum Redundancy Maximum Relevance
MSSO	Modified Salp Swarm Optimization
NB	Naive Bayes
NHA	National Heart Association

NN	Neural Network
PCA	Principal Component Analysis
ReLu	Rectified Linear Unit
RF	Random Forest
RNN	Recurrent Neural Network
Max	Maximum
Min	Minimum
ROC	Receiver Operator Characteristic
SENS	Sensitivity
SME	Subject Matter Expert
SPEC	Specificity
SPO2	Blood Oxygen Saturation
SVM	Support Vector Machine
TEMP	Body Temperature
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UCI	University of California, Irvine
WHO	World Health Organization

Table of Content

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Introduction

Now a days, individuals live a busy and stressful life and have lot of mental and physical pressure due to their workload. They owe low immunity and this compulsive living give birth to many kinds of diseases like hypertension, diabetes, migraine, depression, mental disturbance, back pain and many other critical diseases which affect human life a lot. Among them, heart diseases stand out as a significant cause of mortality in both males and females and this is not only the case of India but most countries in the world have same situation.

The Cardiovascular Disease (CVDs) or Heart Disease (HD) has been a critical one all over the globe. A WHO survey estimated more than 17.9 million demises from HD that is approximately 31% of world deaths, among these, 5:4 (around 85% of total) were because of HD or stroke and mostly common in people above 60s age group [1]. HD is one of the major morbidities worldwide. There are many medical parameters like BP, body temperature, sugar level, cholesterol level, smoke habit and etc. which can impact heart and cause disease. According to WHO "*around 1.1 billion people are addicted to smoking worldwide, out of which, 7 million people die every year*". This disease is very common these days and being the number 1 cause of death globally can be notice in almost all part of the world with mostly all age groups. According to "European Society of Cardiology", "*approximately 50% of heart disease people die within initial 1-2 years*". According to them, 3% of health-care budget are consumed in cardiovascular disease [2]. There are several reasons for heart disease, few of them are genetic disorder, body medical parameters along with certain lifestyle habit which may contribute to HD.

1.2 Myocardial Infarction (MI) or Heart Attack

Myocardial Infarction, a primary cause of death worldwide, occurs post build-up of plaque in the inner surface of the coronary artery, results in blockage of blood flow. It

is commonly known as heart attack. The heart gets damaged due to narrowing down of heart arteries that supply blood to heart muscles. In Cardiovascular disease, deviations can cause malfunction of the heart for example it is unable to fulfil the circulatory demands of the body. Like other muscles in the body, heart also requires a volume of blood supply so that muscles can perform better the functioning of pumping blood to rest part of the body. When one or more coronary arteries get narrowed, it creates difficulties for adequate volume of blood to reach the heart and hence can cause the heart muscle to swollen. Continuous narrowing of the arteries may stress the heart and provoke symptoms. Heart failure is a multifaceted medical illness that impedes the heart from meeting the circulatory needs of the body. This results in the ventricles' impaired ability to transport blood, causing the heart to be incapable of pumping an adequate amount of blood to other regions of the body and hence unable to function normally, and may result in failure.

1.2.1 Cause of disease

In the context of heart disease (HD), a prevalent occurrence involves the constriction or narrowing of the coronary arteries, which are responsible for supplying blood to the heart muscles. This ailment causes the arteries to constrict, thereby hindering the proper supply of fresh blood and reducing the overall blood flow to the heart and hence starts malfunctioning due to not receiving enough oxygen and nutrition. The major causes for failure are cholesterol, high blood pressure, stress, smoking, blood sugar etc. Below image (*figure 1.1*) shows the cause of coronary artery disease wherein the first (A) shows normal artery in heart indicating person is healthy and in second part narrowing arteries can be noticed.

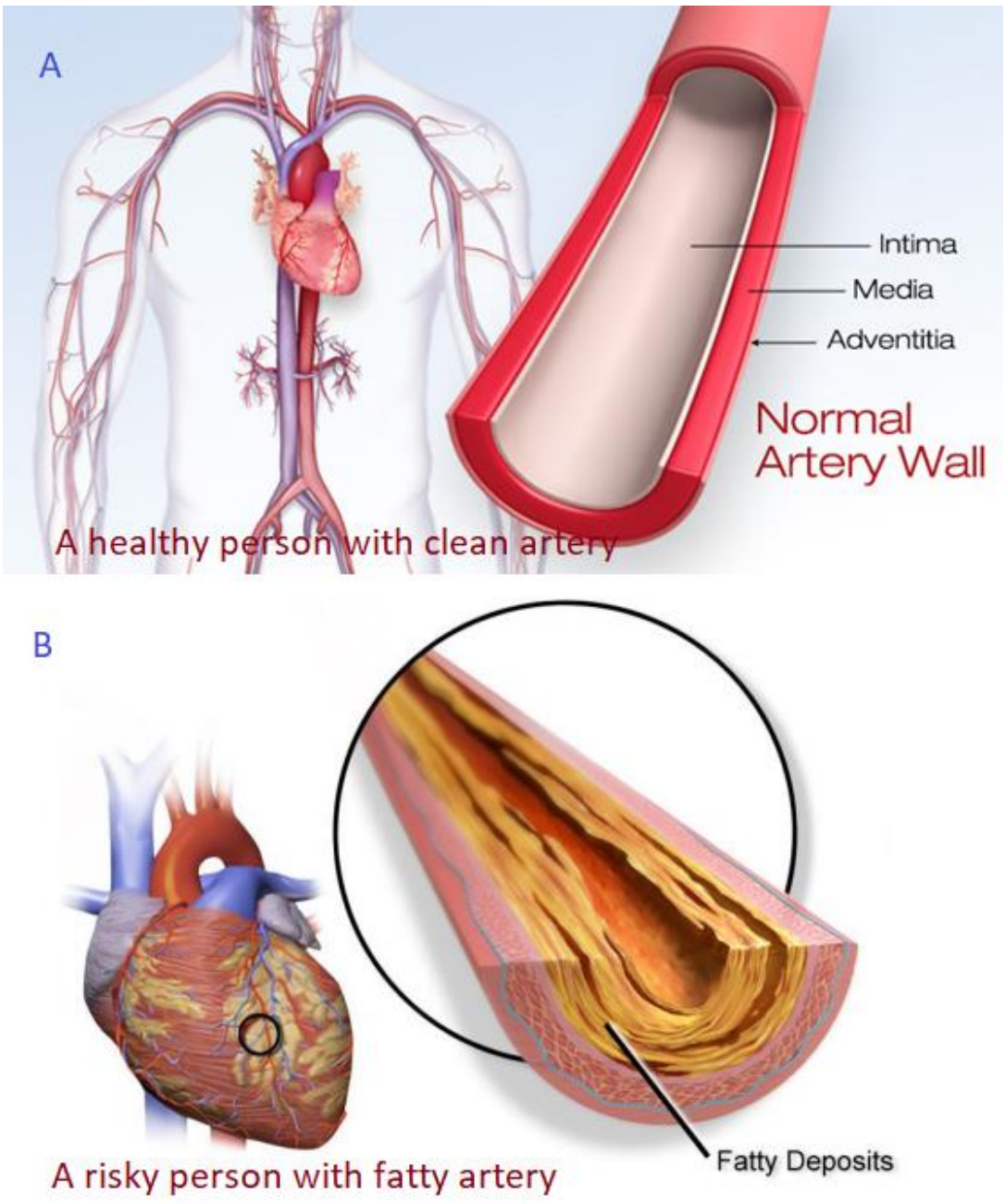


Figure 1.1: Figure showing fatty deposited narrowing arteries [32]

In above figure, it can be clearly notice that when artery is clean, it can help in smooth flowing of blood through it but if a plaque developed inside the artery (some kind of fatty deposit), then blood flow will be blocked which will cause deficiency of blood in heart and other parts as well that later can cause heart attack too.

1.2.2 Medical parameters causing Heart Attack

There are many medical parameters which cause heart attack, few of them are listed below. These parameters have different thresholds, based on that they impact the human body and create symptoms which might causes heart attack, for example heart rate. Let say if a person has normal heart rate, his health condition is considered as healthy but on the other hand if he suffering from any disease or doing some intense physical activity and due to this if somehow his heart rate starts increasing and reach to certain level of threshold, he might get in panic situation and could be the candidate of stroke. Weight gain is also a symptom of heart failure. We get alerted to worsening heart failure when we notice a weight gain of more 2-3 pounds in 24-hour or 5+ pounds in a week [35]. Below we have listed few medical parameters which could be responsible for creating symptom that can cause heart attack.

- Heart rate (HR)
- Blood pressure (BP)
- Blood temperature (Temp)
- Cholesterol (Chol)
- Blood sugar
- Blood Oxygen Saturation (SpO2)
- Smoking
- Family history of heart disease

1.2.2.1 Heart Rate (HR)

Heart rate act as an indicator in human body and it is very closely related to heart related diseases. Any ups and down in heart rate can cause serious issue in heart and even result in heart failure. This can be taken as an indicator for any healthy heart. An arrhythmia (An abnormal rate of muscle contractions in the heart), a disease that causes the heart to beat irregularly either too quickly or slow. Another abnormality called Tachycardia which symptoms as greater heart rate higher than 100 BPM in a resting state [4]. When

the heart rate is between 150 – 200 BPM or more, this condition is known as supraventricular tachycardia (SVT). In this heart rate occasionally beats faster or slower [36].

1.2.2.2 Blood Pressure (BP)

Like heart rate, this is also a crucial indicator of measuring human heart health. BP at both high and low end can impact human heart and lead to sever consequences. There are several factors which can result in abnormal BP, few of them are stress, any kind of nausea like smoke or use of alcohol, excessive physical work like exercise or running. All these should be monitored periodically and should be avoid for keeping heart healthy. Moreover, as age goes up, people amass plaques in the arteries and due to this the flexible walls become stiff and results in tough to pump blood [49] and hence consequently heart starts failing.

1.2.2.3 Body Temperature (Temp)

This plays an important role in maintaining human health. Lowering or hike in body temperature can also lead to CVD.

1.2.2.4 Cholesterol (Chol)

Cholesterol is an important part for our body and helps body in building new cells, insulate nerves, and produce hormones but as there is some good area for this component, there is a bad face as well and that is excessive cholesterol can harm our body till the extend that it can cause heart disease as well [50]. When the cholesterol level in the body is high, it starts depositing in the wall of the artery, the stage is known as atherosclerosis. The blood circulation to heart muscle gets slower or even blocked due to narrowing down of arteries which can cause insufficient flow of oxygen and blood in the heart and might lead to a stage where a person can suffer from stroke or even lead to heart failure.

1.2.2.5 Blood Sugar

If a person having diabetes, then he is very much eligible for heart disease as well, the more is the level of sugar in the body, the more likely you are to have heart disease.

According to the data collected by National Heart Association in 2012, 65% people with symptoms of diabetes will die from HD or stroke [51]. With the span of time, elevated blood glucose levels can inflict damage upon both the vasculature and the neural regulatory mechanisms governing cardiac function. Blood sugar is also associated with other medical parameters like high BP and high LDL cholesterol [52][93].

1.2.2.6 Blood Oxygen Saturation (SpO2)

SpO2, the measure of oxygen in body which is bound to cell's haemoglobin, is an important parameter and can impact heart on a serious note as it is directly linked with body metabolism. Low SpO2 level in the blood can result in failure of organ functions and that can harm body seriously such as necrosis and loss of function. Normally the values remain between 95 to 100%.

1.2.2.7 Smoking

Smoking is one of the critical medical parameters which impact heart of patients and if a person is chain smoker, then he might get impacted with HD [47, 48].

1.2.2.8 Heart disease family history

Like other medical parameters which put direct effect on heart, family history does not impact directly but yes, this also plays an important role. A person with HD family history is at a heightened likelihood of encountering same disease. He also might suffer from the disease if not now then may be at later age. Family history of heart disease works same way as diabetes, a genetic issue which get transferred in offspring from parents or fore parents [94].

1.2.3 Disease Symptoms

Below are the few symptoms that could be useful in prediction the disease [56]:

- Discomfort or pain in the chest radiating other parts: This is the commonly reported indication of a heart attack. It may manifest as tightness, pressure, squeezing, fullness, or aching in the chest area, and can radiate to other parts of the body i.e. shoulder, back, arms, neck or jaw.

- Breathing difficulties: A person experiencing a heart attack may find it hard to breathe, feeling short of breath or like they are breathing faster than usual.
- Profuse sweating: Even when not engaging in physical activity, a heart attack can cause sudden cold sweats.
- Nausea or vomiting: During a heart attack, a person may feel nauseous or actually vomit.
- Feeling faint or dizzy: Light headedness or dizziness can also be a symptom of a heart attack.
- Unusual fatigue: A person experiencing a heart attack may feel more tired than usual, even if they haven't been doing anything strenuous.

1.2.4 Challenges in Treatment

Heart attack diagnosis is very important, but due to lack of understanding of heart failure characteristics is quite difficult. The disease is so serious that a single minute delay in diagnosis can cause death so it is useless to say that it needs immediate attention. Its criticality is not hidden from anyone and needs special attention to find the suitable solution to diagnose and prevent this disease. With the advancement of AI as an option to search a way, many techniques came into existence and lead into gearing the experiments using ML and the concept of data mining and stood as an alternative medicinal option.

Many researchers from various corners of the globe are dedicatedly working to discover a solution for this challenge and to mitigate these challenges, machine learning came into existence. With the help of ML, a system can be built that will help in not only diagnosis heart disease within time but also in suggesting few precautions that can help patient to overcome the chances of heart attack.

1.3 Introduction to Machine Learning

Technology is growing very rapidly and everything is going towards Artificial Intelligence. Every stack wants to automate their work and introduce some intelligence in their existing system so that they can provide solutions for different problem

automatically by learning the behaviour and problem. In last decade AI and ML has revolutionized the research and changed the perception to think for technology. It has showed its presence in every field and contributed a lot in development of many applications to provide solutions in all domains. Development of models related to the health-related issues has been the most active research areas and use of computer-based diagnosis and prediction method with the help of Artificial Intelligence remains noticeable. Many experiments have been carried out to build a clinical solution for the forecasting many critical illness like cancer, heart, diabetes, Alzheimer, mental health disorder etc. where Machine Learning (ML) with data mining has come up with desired solutions. Artificial Neural Network (ANN) with its architecture has demonstrated higher efficiency. In this research work, many ML techniques has been used for experiments and results were compared to find optimal solution that develop personalized risk assessments based on their individual medical history and lifestyle factors that should be able to show better model accuracy and should be able to find the likeliness of heart attack using the medical parameters included in the proposed work, few of them are discussed in below sections. Machine learning (ML) algorithms can analyse enormous volumes of data to recognize patterns and build predictive models for different heart disease and the solution would take into account the latest medical research and best practices to ensure the better quality of care.

1.3.1 K-Nearest Neighbors (KNN)

A ML classical technique based upon the supervised-learning works on the comparison of nearest neighbouring point (k points), very slow in nature and expensive as compared to other traditional algorithms, a lazy learner non-parametric algorithm that in spite of learning from the training set shows output based on stored dataset, can work on both types of problems i.e. classification and regression, but popularly can be seen for classification once. In training phase, it only stores the dataset and in testing phase, data is classified into a category that is much similar to the stored data. It formulates with K= (1...n), where K represents numbers 1 to n [42]. It is known as instance-based learning and is one of the most popular ML methods. It doesn't contain any stage for model training. Instead, it calculates the approximate covered distance within different

local input vectors and decision is based on the measurement of distance between class levels, top levels are picked after applying sorting on measured distance and outcome is predicted and with that reason performance can be boosted by normalizing the dataset. The K parameter affects performance and accuracy. Below, *figure 1.2* represents the KNN classifier.

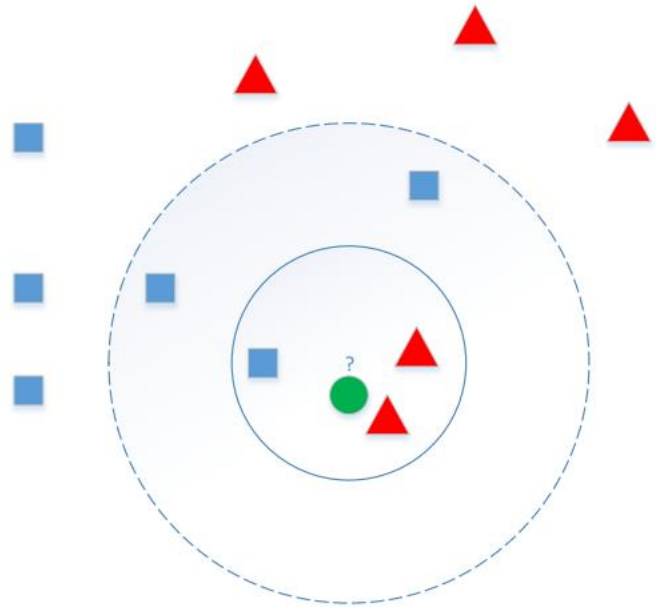


Figure 1.2: Class representation in KNN

1.3.2 Support Vector Machine (SVM)

This method falls under supervised-learning ML techniques having objective to find an optimal solution (creating decision boundaries between class levels) in n-dimensional plane, are useful for training linear as well as non-linear classification model [38] mostly used in binary classification. SVM is uses for regression and classification problems of the supervised learning. SVM follows kernel concept that is mainly a mathematical approach which transform the input data accordingly into the desired format and find the optimum solution. Below *figure 1.3* is the pictorial representation of SVM having hyperplane between two class A and B.

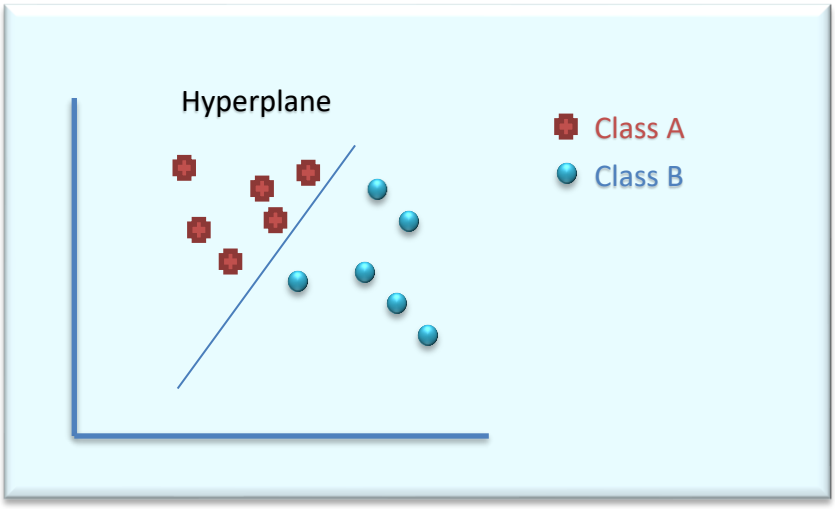


Figure 1.3: SVM showing decision boundary between class A and class B [38]

1.3.3 Decision Tree (DT)

A tree based very simple node-based structure that works on branching system and decisions are based on if-else condition where class levels are represented as a leaf node. Decisions are made on the basis of tree leaf and get divided into two parts – left or right, conditions are applied and best solution that match the conditions are sent as output [39]. Decision tree has been pictorial represented in the *figure 1.4*.

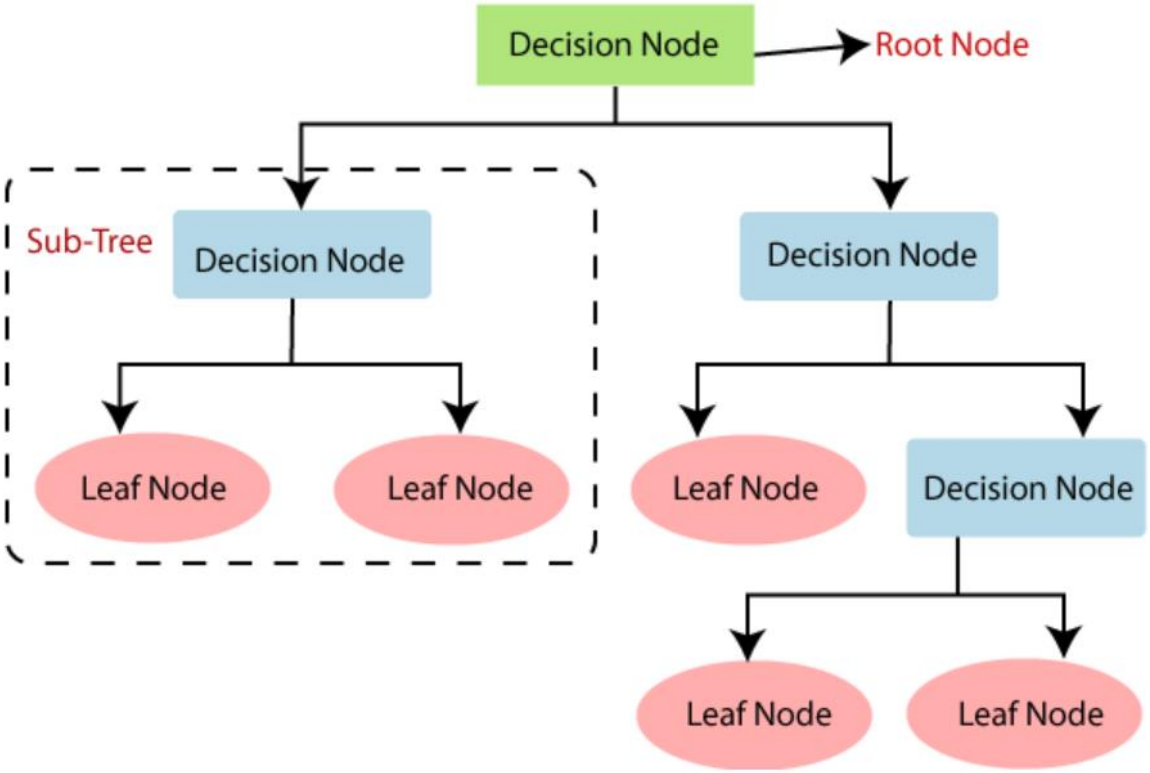


Figure 1.4: Decision Tree showing different level of nodes [40]

1.3.4 Random Forest (RF)

Structure of RF is mostly similar to DT and it follows same principle, you can understand this as the larger version of decision tree where it build a number of decision trees and later club them together to build a large tree to obtain higher and stable accuracy, *figure 1.5* is the pictorial representation of RF. This can be made more sophisticated by applying randomness to the model. Adding to this, this has some additional features like it can work on large dataset having large number of features without reduction in features, is useful in solving both classifications as well regression problems and can be used for time series problems as well [41]. It learns and train model using data available in dataset and can perform better than other methods but on the other hand, it bound itself with this restriction and is unable to predict future value which somehow emerged as its major drawback when we deal with time-series problems.

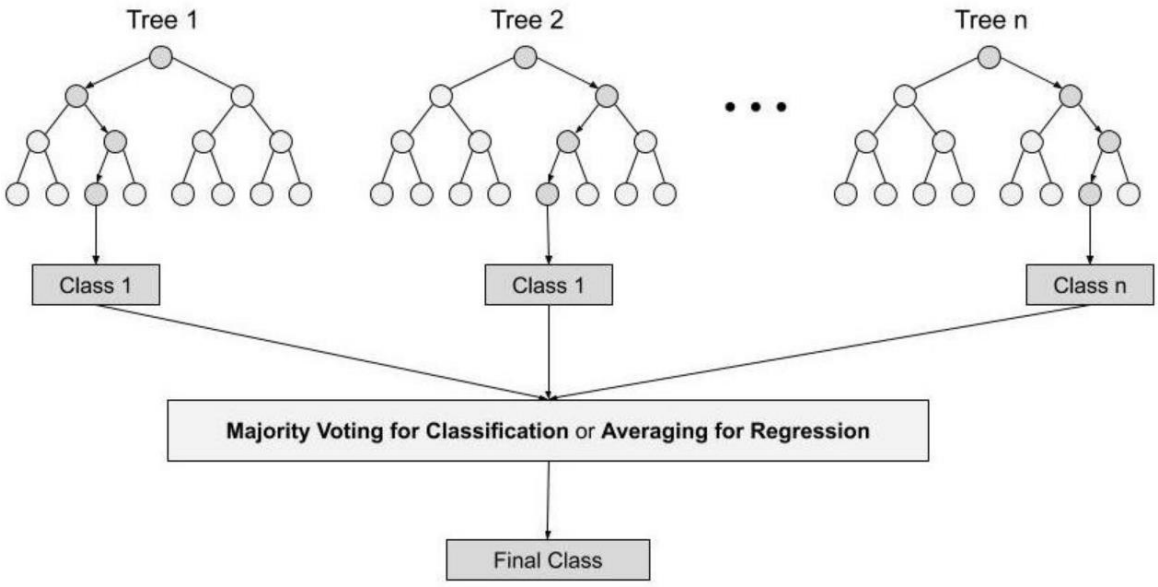


Figure 1.5: Diagram showing Random Forest (RF) [45]

1.3.5 Logistic Regression (LR)

The LR model, a popular employed technique in supervised learning, proves highly effective for tackling classification issues, be they binary or encompassing multiple classes [43]. This statistical model is harnessed to gauge the likelihood of a binary outcome variable by relying on one or more predictor variables, operates through the application of a sigmoid-shaped function, commonly referred to as the logistic function. This mathematical transformation serves to convert input values into probabilities that fall within the range of 0 to 1. The model's formulation manifests as a linear combination of predictor variables, each accompanied by its corresponding coefficient that play a pivotal role in indicating the alteration in the log-odds of the outcome variable for every one-unit increment in the predictor variable [46]. This intrinsic property makes logistic regression an invaluable tool for understanding the relationships between predictor variables and the probability of a particular outcome occurring. Through this approach, logistic regression excels in discerning the likelihood of an event transpiring, making it indispensable in a myriad of domains. Its adaptability to both binary and multi-class classification scenarios, along with its interpretable coefficients, renders it a cornerstone in the toolkit of machine learning practitioners. The significance of logistic regression is further accentuated by its capacity to offer insights into how individual predictors impact on the likelihood of a given outcome.

This elucidation, in turn, fosters a deeper comprehension of the underlying mechanisms governing complex phenomena. As a result, logistic regression stands as a robust and versatile technique for a wide array of classification tasks, forming an integral component of predictive modelling in diverse scientific and practical domains. Output of LR is depicted in *equation 1.1* and *figure 1.6*. It creates a s-curve like graph and be shown by below equation.

$$\log\left[\frac{y}{1-y}\right] = \omega^0 + w^1x^1 + \omega^2x^2 + \cdots \cdot w^nx^n \qquad (1.1)$$

Equation 1.1 shows the logistic regression where *y* is the probability of the dependent variable taking the value 0-1, x^1, x^2, x^3, .. x^n are independent variables and w^0, w^1,..w^n are coefficient to be estimated from data.

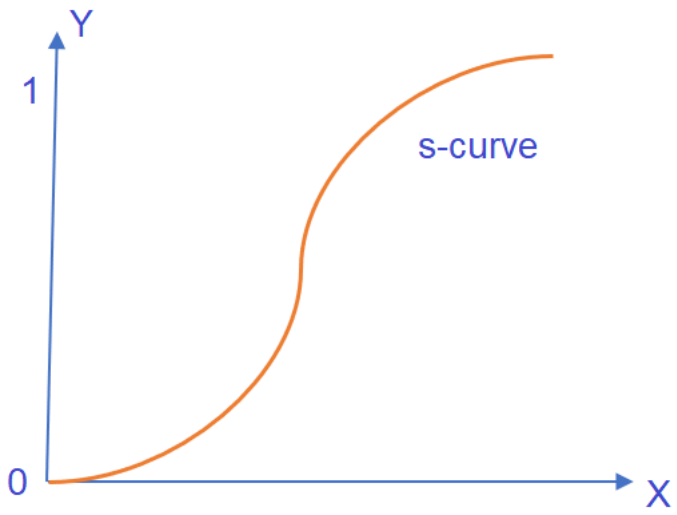


Figure 1.6: Graph showing Logistic Regression (LR)

1.3.6 Naive Bayes (NB)

NB belongs to the simple classification family based on "*Bayesian theorem*", a bunch of algorithms which follows common principle with a theory that all features are independent of others present in the dataset; can learn efficiently and quickly using supervised learning model and performs better as compared to other models with different set of datasets, most importantly with the large dataset and prefer categorical input over numeric one. One of the major drawbacks of this classifiers are that their

output is not reliable and practical in nature even if they are easy to train and faster in result due to the reason that for the category of data in test dataset that is not present in training set, NB predict no probability, other drawback is impossibility of separation of independent feature of a class which is not tangible in real scenario. NB has many varieties of algorithm but the most common one is Gaussian [44]. Below formula can be used as the base line to predict the output.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.2}$$

In equation (1.2), the probability of A given the occurrence of B is denoted as *P(A|B)*, while the probability of B given the occurrence of A is denoted as *P(B|A)*. *P(A)* and *P(B)* represent the marginal probability of A and B, respectively.

1.3.7 Neural Network

The concept is mainly emerged from the neuron functionality inside human body. Neural network or NN basically reflect the behaviour of the human brain and are very capable in making human like decisions. It allows computer programs to recognize patterns, make classification related decision and solve common problems. This is the future of artificial intelligence and with the help of ML it is solving problems from different domains. Various types of neural networks exist, including ANN, DNN, CNN, RNN and etc. Now a days many hybrid systems also emerging like GANN which combine neural network with other algorithms like in GANN, neural network is combined with genetic algorithm.

1.3.7.1 Structure of Neural network

The neural network (NN) architecture comprises three pivotal layers - input, hidden, and output. Within this structure, each neuron in a given layer establishes connections with neurons in the subsequent layer which are characterized by weights, which are computed in tandem with an activation function, augmented by a constant term known as bias [53]. The initial phase involves the reception of input data by the neurons situated within the first layer. Subsequently, neurons in the ensuing layers compute their values through a weighted summation of the neurons in the preceding layer. This

dynamic interaction of weights and activations drives the network's ability to discern intricate patterns and relationships within the data. It's this hierarchical arrangement that endows neural networks with their formidable computational ability, making them adept at modelling complex phenomena across various domains. This structural composition and interconnectedness are crucial in enabling the network to learn and adapt from the provided data. Through a process of iterative adjustments to the weights and biases, the network refines its ability to capture the underlying features and dependencies in the data which is what underpins the efficacy of neural networks in tasks ranging from image recognition to natural language processing, forming the cornerstone of modern machine learning methodologies. The neural network can have multiple hidden layers in the case of a multi-layer NN. F*igure 1.7* shows the layer structure in NN.

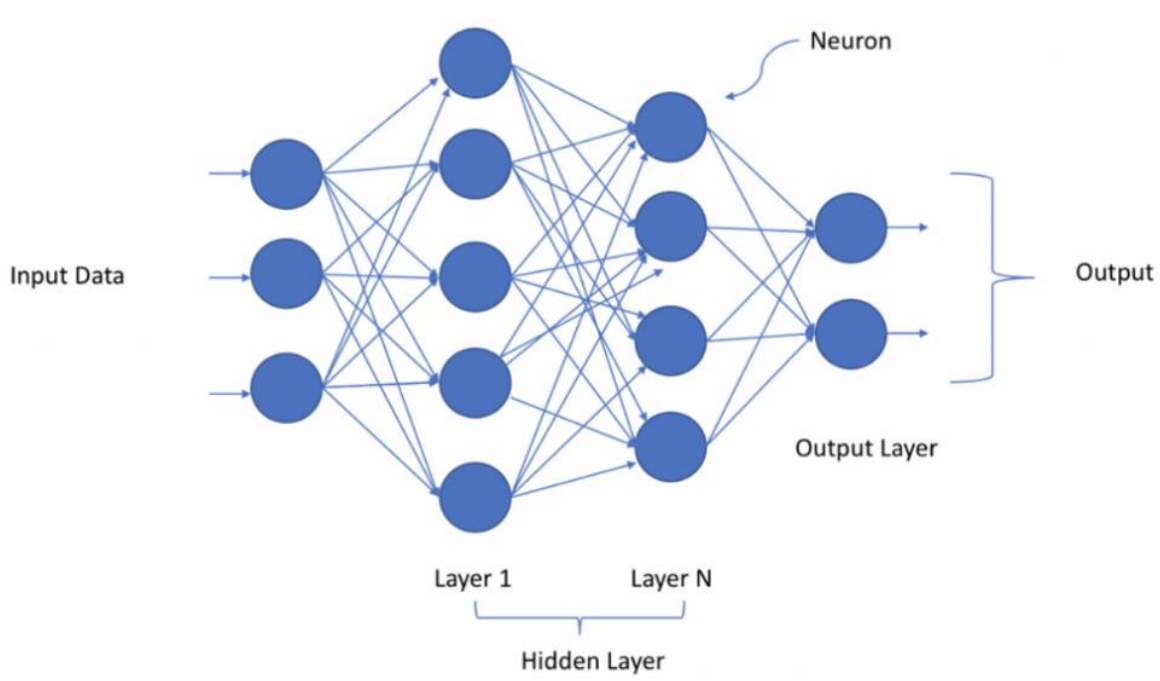


Figure 1.7: A neural network showing neurons fitted into layers [53]

1.3.7.2 Components of neural network

In neural network there are four components which creates neural network – neuron, weight, bias, activation function, optimizer. Let’s see these one by one.

Neuron: This is the fundamental element of NN or we can say that neurons are the most required component to build a neural network.

Weight: This is second most important component in NN wherein all the neurons of two consecutive layers are connected to each other with an associated weight attached to it. This could be defined as the impact of input of current layer on the output for next neuron. *Figure 1.8* shows the formation of new NN node after applying weighted sum to neurons or node in current layer.

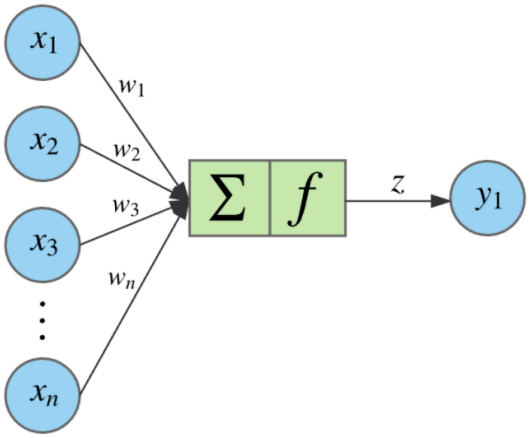


Figure 1.8: A NN node after applying weighted sum to get new node

Bias: This could be understood as a constant value inject in the network serving to fine-tune the activation function within the neural network. This augmentation plays a vital role in regulating the behaviour of the activation function, exerting a discernible influence on the network's overall computational dynamics.

Activation function: The activation function mainly helps in deciding if the neuron can participate in the network or not or simply say that it decides the importance of neuron for the proposed network. For that it follows simple formula that if the incoming neurons outcome is greater than the threshold, the output is passed else ignored. There are several types of activation function which is applied on the basis of requirement. Few of them are describes below:

ReLu: This function, also recognized as the “Rectified Linear Unit (ReLU)”, finds its predominant application within the hidden layers of neural networks. Its characteristic trait lies in confining output values within the range of 0 to x. Specifically, if the input proves to be positive, the function faithfully reproduces the same value as output. Conversely, in the event of a non-positive input, it yields an output of zero. This property imparts a distinctive non-linearity to the activation function, endowing it with

the capability to facilitate complex learning processes within the network. It is among popular activation functions, trains very fast and generates better performance than other activation functions [54]. In this research work ReLu function has been used at hidden layers in the experiments. Below equation explains the ReLu function.

$$h = \max(0, a), where\ a = Wx + b \tag{1.3}$$

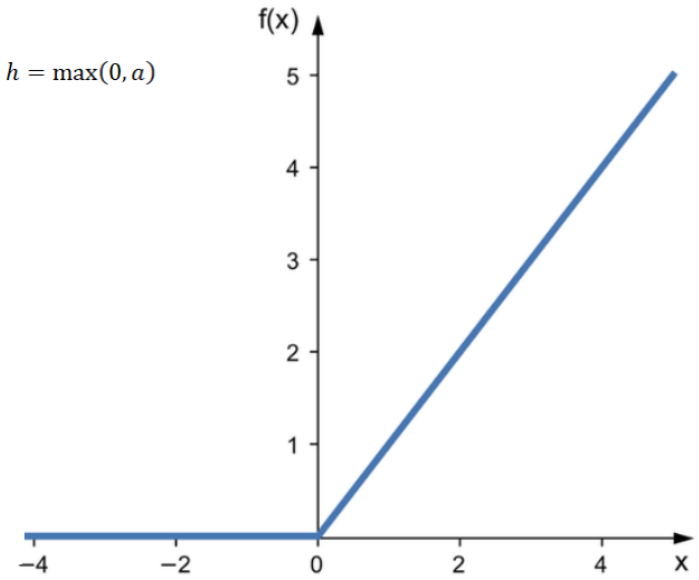


Figure 1.9: ReLu activation function

Sigmoid: This function, akin to a squashing operation, restricts the output within the range of 0 and 1, exhibiting a distinctive S-shaped curve, predominantly employed in models aimed at predicting probabilities as their output. Logistic sigmoid function is used when output is monotonic or binary classification and softmax function if output is multiclass classification. In this research work sigmoid function has been used at output for binary classification. Below, *equation 1.4* and *figure 1.10* enlighten the same.

$$f(x) = \frac{1}{1+e^{-x}} \tag{1.4}$$

where e is the mathematical constant Euler's number, x is input variable. *f(x)* is calculated as when *x* is positive number, e^{-x} approaches 0, and result approaches 1, whereas if x is negative, e^{-x} approaches infinity, and result approaches 0.

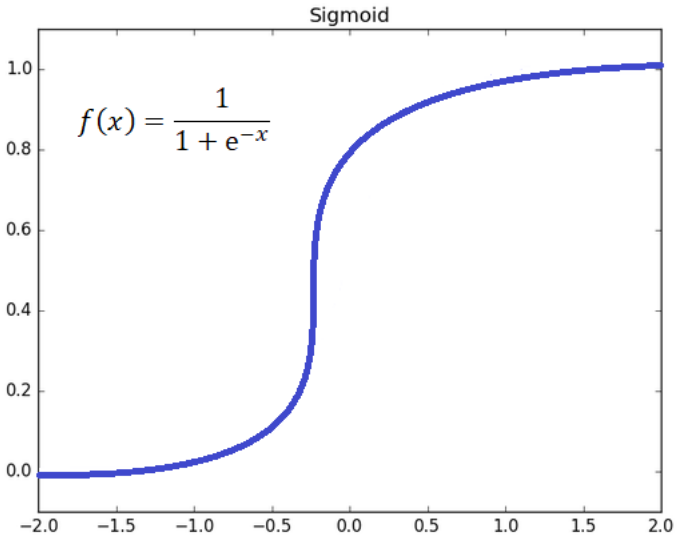


Figure 1.10: Sigmoid activation function [55]

Tanh: The Tanh function shares similarities with the sigmoid function, yet it differs in that its range spans from -1 to 1. The function exhibits a symmetric S-shaped curve around the origin and is steeper than the sigmoid function, making it more responsive to changes in the input, generally used in the hidden layers of neural networks to incorporate non-linearity and increase the model's complexity. *Figure 1.11* explains the tanh functions.

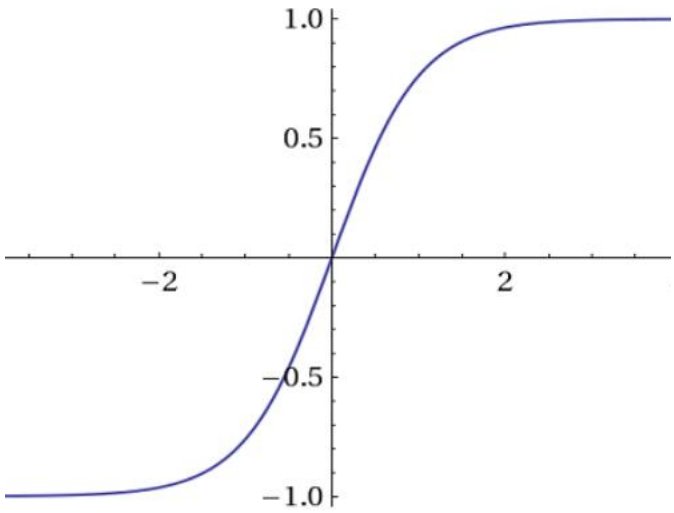


Figure 1.11: Tanh activation function [54]

1.3.8 Genetic Algorithm

The system is built on the *"Darwin's theory of evolution"*, works on "random-based classical evolutionary algorithm". This algorithm helps in generating new offspring from existing parents by apply certain techniques like mutation and crossover until model produce the best solution. It assigns a fitness score to each solution, operating under the premise *"higher is the fitness value, better is the solution".*

- *Key terms:* Genes, chromosome, population, parent, offspring, matt-pooling, crossover, mutation
- *How it works:* The evolution mechanism is iterative in nature and starts from initial population that is *"randomly generated individuals"* and execute till the defined generation (number of iteration) or till the best fit solution is found. The fittest individuals are then selected to undergo "*crossover and mutation*" to generate offspring for the subsequent cohort. Crossover involves exchanging genetic material between two individuals to create new solutions, while mutation involves introducing random changes in the offspring's genetic makeup. This process continues for multiple generations until a satisfactory solution is found or a predetermined stopping criterion is met. Ideally, the algorithm ends up with the condition – "either a maximum number of generations is produced, or a satisfactory fitness level has been reached for the population" [57-59].
- *Operations in GA:* GA induced with three major operations – Selection, Crossover and Mutation.
    - *Selection:* This operation mainly deals in selecting best fit individual from current parent so that new offspring can be generated for iterating in next generation. To do this, fitness of each individual is measured based on the accuracy and individual with highest fitness values are selected for next stage.
    - *Crossover:* Very next operation in generating new set of parents is to apply crossover among the selected set of individual or say parents. This is done randomly so that biasness can be removed from the generation

and best solution can be generated [57-59]. *Figure 1.12* depicts the process of crossover where genes in current population is exchanging themselves to produce new offspring. Generated offspring will undergo another process of mutation to generate final one which will be used in next iteration.

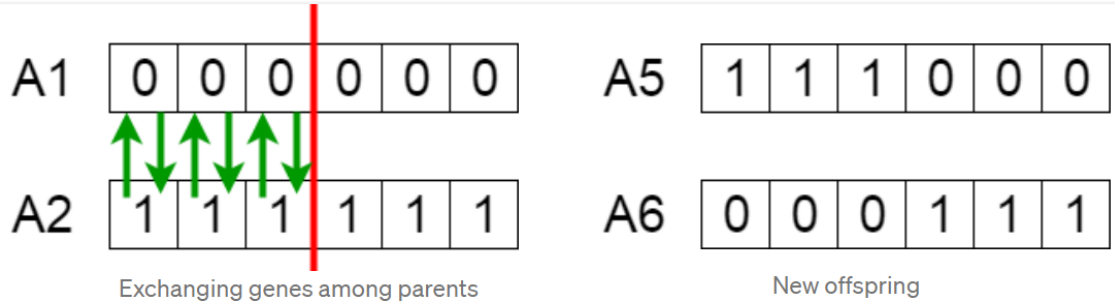


Figure 1.12: Process of crossover for generation of new offspring [57]

- *Mutation:* The very next step is mutation process which is mainly done to mix random individual in newly generation population set [57-59]. *Figure 1.13* shows the mutation process by replacing genes randomly to remove biasness from the generated population.

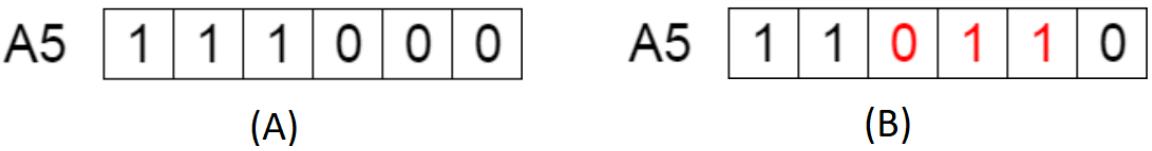


Figure 1.13: Process of mutation showing before (A) and after (B) stage [57]

1.3.9 Fuzzy Inference System (FIS)

The need of emergence of the FIS came from idea to enable the model with the human like decision making capability where input is not crisp rather it works on some fuzzy logic something having linguistic value, for example, if car reaches closer to another car running then start breaking to lower down the speed. This kind of problem can be handled using FIS. We can understand this as *"the process of interpretation of values in the input vector and converting the fuzzy values into the output vector using fuzzy rules, sets and membership functions processing through the inference engine"* [60].

Many researchers did lot of work on this system by combining many available systems with FL to create a new system like with inference system, NN and GA and so on.

1.3.9.1 Components of Fuzzy Inference System

There are mainly seven components or say functional blocks from which FIS is composed of, *figure 1.14* represent the same and are following:

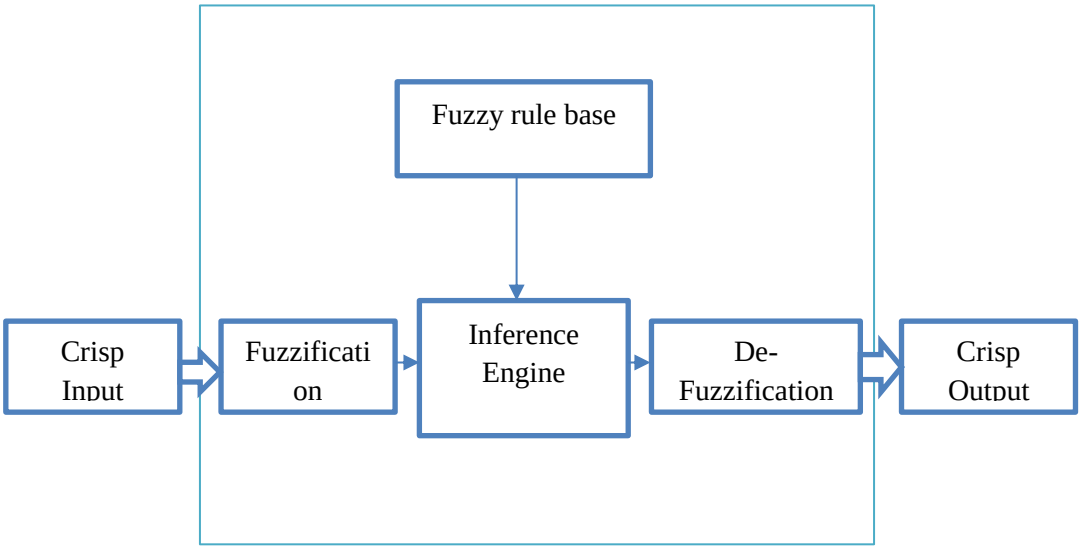


Figure 1.14: Functional blocks of FIS

Input: This mainly corresponds to the input values passed to the FIS which is being process and converted into desired formats for further use.

Knowledge base: This refers to the collection of fuzzy rules (If-then conditions) and membership functions that define how inputs are mapped to outputs. It contains the linguistic variables, fuzzy sets, and the rules that govern the behaviour of the system. The knowledge base is what allows the FIS to process inputs that are expressed in imprecise, linguistic terms. It adds decision making capability to the system based on a set of rules that capture expert knowledge. This can be further extended as:

- **Linguistic Variables:** These are variables that use linguistic terms (like "low", "medium", "high") to represent values. For example, in a disease prediction system, a linguistic variable could be "BP", and it could have fuzzy terms like "low", "medium", and "high".

- **Fuzzy Sets:** These are sets that define how each linguistic term relates to a range of values. For example, the fuzzy set "medium" might include BP from 110 to 130.
- **Fuzzy Rules:** These are statements that define how inputs relate to outputs. They are usually in the form of "If [antecedent], then [consequent]". For instance, "If BP is high, then probability of disease is low to medium".

Membership function: A membership function μ_A enable us to represent the fuzziness of the set with the help of graph on *x*-axis and *y*-axis in [0, 1] interval respectively wherein membership degree is represented by $\mu_A(x)$ of *x* to a membership value lies into “0” and “1”, indicating the degree to which the input relates to a particular fuzzy set. Same has been shown in equation 1.5 where *A* delineates the fuzzy set operating within a universe of discourse denoted by X. Each element within this universe is assigned a membership degree in set *A*, a value spanning the interval from 0 to 1. This value, known as the membership grade [61], epitomizes the extent to which an element is deemed a part of set *A*. In practical terms, this means that membership functions provide a quantitative representation of how much an element belongs to a particular fuzzy set.

$$\mu A: X \rightarrow [0, 1] \tag{1.5}$$

Features of membership function: A membership function is basically divided into three parts – Core, Support and Boundary explained below [62], *figure 1.15* represent the same in pictorial form.

- *Core:* This comprise of those elements for which membership value =0
- *Support:* This is the set of all the elements for which membership values >= 0
- *Boundary:* This is the boundary i.e. 0 < membership value <1

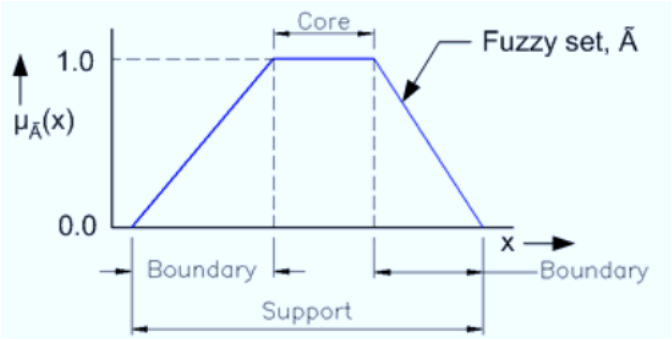


Figure 1.15: Pictorial representation of features of membership function [62]

There are mainly four MFs available, those are following:

- Triangular
- Trapezoidal
- Gaussian
- Sigmoidal

Triangular function: This has **p**, **q** as a lower and upper limit with a value **m**, where **p** < **m** < **q**. There are four cases exist in triangular function, equation 1.5 shows the same.

1) Value remains 0 if $x \leq p$ or $x \geq q$
2) Value can be calculated as $\frac{x-p}{m-p}$ for p < x ≤ m
3) Value can be calculated as $\frac{q-x}{q-m}$ for m < x < q
4) Value remains 0 if r ≤ x

$$f(x;p,q,r) = \begin{Bmatrix} 0, & x \leq 0 \\ \frac{x-p}{q-p}, & p \leq x \leq q \\ \frac{r-x}{r-q}, & q \leq x \leq r \\ 0, & r \leq x \end{Bmatrix} \quad (1.5)$$

Trapezoidal function: The function of the vector is determined by four scalar parameters, namely p, q, r, and s. The curve's feet are represented by p and s, while the shoulders are represented by q and r., equation 1.6 represent the same.

$$f(x;p,q,r,s) = \begin{cases} 0, & x \leq 0 \\ \frac{x-p}{q-p}, & p \leq x \leq q \\ 1, & q \leq x \leq r \\ \frac{r-x}{r-q}, & r \leq x \leq s \\ 0, & s \leq x \end{cases} \quad (1.6)$$

Gaussian function: Gaussian, the bell curve distribution, consisting of two parameters *"mean and standard deviation"* that controls the *"width of curve"*; function is symmetric around the mean and covers the highest value at the mean itself, and leading to decreasing the values when moving away from the mean, following the bell-shaped curve. The below given *equation 1.7* represents a continuous function that is dependent on a vector "x", with parameters "p" and "q". The parameter p denotes the standard deviation, which determines the width of the distribution such that, larger the value of p, the wider the distribution, and vice versa. The parameter q is the mean and determines the location of the peak.

$$f(x;p,q) = e^{-\frac{(x-q)^2}{2p^2}} \quad (1.7)$$

Sigmoid function: Sigmoid is the function, dependent on vector x, determined by two parameters, p and q, given in equation 1.8. It takes two parameters p and q, and the input x is transformed based on these parameters. As the input x increases, the output approaches 1, when x decreases, the value approaches 0. The parameter q controls the midpoint of the curve and p controls the steepness of the curve, i.e., larger value of p results in a steeper curve. The function's output always lie between 0 and 1, making it suitable for binary classification problems.

$$f(x;\mathrm{p},\mathrm{q}) = \frac{1}{1+\mathrm{e}^{-\mathrm{p}(x-\mathrm{q})}} \quad (1.8)$$

Fuzzification: This is the method of translating crisp or numerical inputs into fuzzy linguistic variables or fuzzy sets in fuzzy logic. It is a crucial step in fuzzy inference systems that allows for the representation of imprecise and uncertain knowledge in a more natural and flexible way. Fuzzification involves assigning membership values to the input variables based on their degree of belongingness to different fuzzy sets. This helps in capturing the vagueness and ambiguity in human language and perception, which is often difficult to model using traditional crisp logic.

Inference System: This is the method of translating fuzzified input to the fuzzy output using fuzzy rules. It uses logical rules, expert knowledge, and a knowledge base to make decisions or recommendations. The technology mimics the cognitive process of a knowledgeable individual in a specific domain and can draw conclusions from known facts and evidence.

Defuzzification: The process of defuzzification involves utilizing various defuzzifier methods to convert the fuzzy output attained from the fuzzification process into a crisp value like “maxima method”, “weighted average method”, “center of gravity method”, “center of sums” and “centroid of area method” etc.

1.3.9.2 Types of Inference System

Three type of inference system has been discovered so far which is being used in FIS, they are:

- Mamdani system
- Sugeno system
- Tsukamoto system

1.3.9.2.1 *Mamdani system*

Mamdani-type FIS was proposed in 1977 by “Ebhasim Mamdani” [60]. In this method, post aggregation process, fuzzified input get converted into fuzzy output sets which later need to go through the defuzzification process for converting into crisp output. In this method there is two antecedents (x_1 *and* x_2) and one consequent y and they are all fuzzy sets.

In Mamdani there are two cases for 2-input Mamdani system.

- Max-Min inference method (here we take truncated membership function)
- Max product inference method (we take scaled membership function)

Inference means to reach on a conclusion based on some evidence associated with a logic.

Max-Min inference method: In this method we check for rule connector ("and", "or"). If it is and, then pick minimum value and if it is or, then pick maximum value. Later we truncate the triangle and take the trapezoidal part.

Max product inference method: In this method we do same as above but we don't truncate the triangle, rather we scale it down and take shorten triangle instead of trapezoidal.

1.3.9.2.2 *Sugeno system*

Takagi, Sugeno and *Kang* in 1985, proposed the system [60]. Here also input is provided as fuzzy set, lowest point is picked but it returns crisp output so no need of Defuzzification is required and direct crisp result can be utilized for further use. It is better than Mamdani as it gets crisp output hence no Defuzzification is required. We can directly apply weighted average of defuzzification method. In this system there are two antecedents (input) and one consequent (output) but the difference is that antecedents are fuzzy set and consequent is the function of two input in the antecedents.

If p1 is M1, p2 is M2, then q = f(p1, p2) where p1 and p2 are input, q is output and M1, M2 are fuzzy sets and q is a crisp output where q = f(p1, p2) is crisp function in the consequent.

IF p is M and q is N then r = f(p, q)

1.3.9.2.3 *Tsukamoto method*

In this method there is two antecedents and one consequent and they are all fuzzy sets. The distinction is that the consequent's membership function is monotonic, implying that "*each fuzzy rule's consequent is a fuzzy set with a monotonic membership function*".

Monotonic function is also called shoulder function and is the one whose successive value is increasing, decreasing or constant.

If p1 is M1, p2 is M2, then q = N where p1 and p2 are input, q is output and M1, M2 and N are fuzzy sets. We first calculate values of p1 and p2 based on the membership values and then calculated membership values wq1 and wq2 based on minimum or maximum of p1 and p2 based on the rules. Later we calculated crisp values of q1 and q2 by extending wq1 and wq2. Now when we have crisp q1 and q2 we calculated weighted average of q1 and q2 that is q*.

1.3.9.3 Working of FIS

The Fuzzy Inference System (FIS) predominantly navigates through four sequential stages subsequent to receiving input “Knowledge Base”, “Fuzzification”, “Inference Engine”, and “Defuzzification” [63].

- *Knowledge base:* The fuzzy rules creation is a critical stage in FIS, which comprises antecedents and consequents that depend on the number of inputs. Subsequently, corresponding membership functions are generated for each input by utilizing the fuzzy rules and fuzzy sets, and for this there are many methods available like triangular, trapezoidal, gaussian or sigmoid.
- *Fuzzification:* In this step crisp input is converted into linguistic output which later is used by inference engine for further processing.
- *Inference engine:* In this stage fuzzified value is taken in and membership degree is being calculated. The output is fuzzy again. This is done using one of three available methods i.e., Mamdani, Sugeno or Tsukamoto.
- *Defuzzification:* This is the last stage in this system where input from last stage defuzzified and are converted into crisp output. There are various methods available for this task as well like center of gravity, center of sum, centroid of area etc.

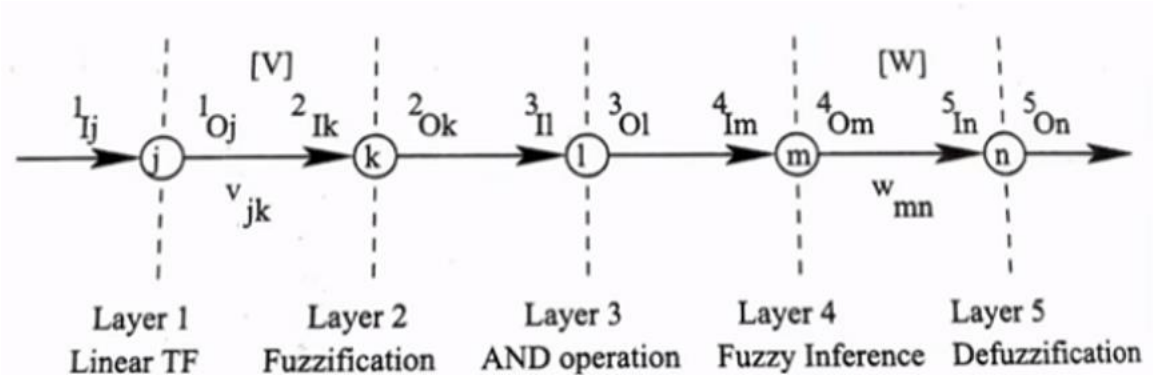


Figure 1.16: Layer wise placement of input/output neurons in Neural Network

The visual depiction in *figure 1.16* illustrates the arrangement of input and output neurons across various layers of the Neural Network (NN). Moving on to *figure 1.17*, we observe the graphical representation of the "*Adaptive Neuro Fuzzy Inference System (ANFIS)"*, organized into five distinct layers: "Input", "Membership", "Fuzzification", "Normalization", and "Defuzzification". The inference system in ANFIS predominantly employs the Sugeno method. This system essentially amalgamates the strengths of "Artificial Neural Networks" and "Fuzzy Logic Systems", utilizing the capabilities of ANN to generate precise fuzzy rules alongside appropriate membership functions [64].

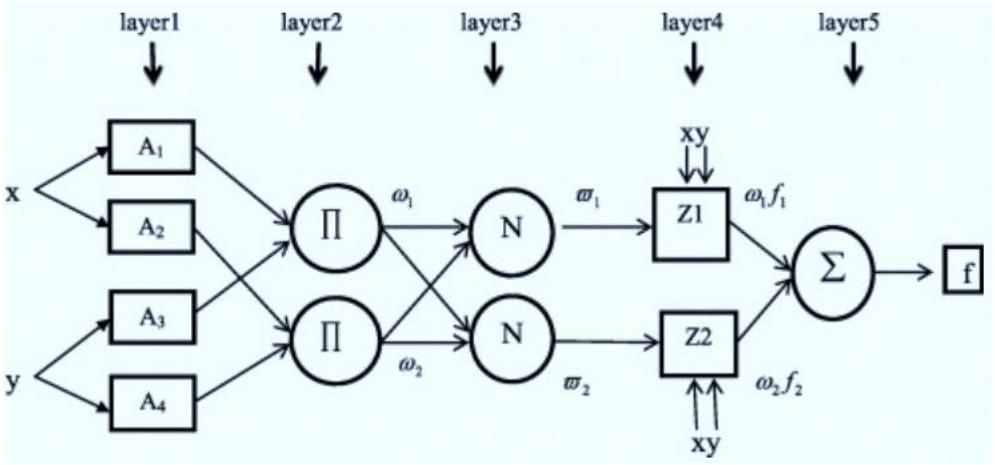


Figure 1.17: Five layered structure showing ANFIS

In *figure 1.17* displayed above, the input variables x and y are introduced into the ANFIS architecture. These variables traverse through distinct layers within the system. The initial layer, denoted as the Membership Layer, is where the process initiates. Following this, the second layer is dedicated to Fuzzification. The third layer, known

as Normalization, comes next. Proceeding further, the fourth layer specializes in Defuzzification. Finally, the fifth and concluding layer operates as the Output Layer, delivering the ultimate outcomes of ANFIS computation.

1.4 Motivation towards the proposed work

An average of 31% of death materializes due to heart disease worldwide out of which 80% are because of heart attack only [1], it shows the criticality of this disease and needs special attention to find a suitable solution to diagnose and prevent this disease before the situation gets out of control.

- *Motivation:* Before we picked this research topic, initial investigation was done on the selected area to analyse the root cause of the disease and how people are affected from heart attack. We tried to talk to many people around us to know that what they think about heart attack and how they feel when someone talks about this syndrome. We observed that in fact people were not ready to talk about the disease in normal scenario. People are afraid of this disease as they even don't know that when they will be in trouble due to this, any harsh situation can lead to stroke regardless of age boundary. We have observed people getting heart stroke at every age regardless of they follow good health chart, take good food, do regular exercise and etc. In fact, people do everything they can to avoid this disease but somehow most of us get into this. I personally faced situation when one of colleague husband got heart stroke while arguing with Police and die on the spot. This disease is so critical that today many renowned researchers are concluded their studies on to finding solution for different-different variant of HD. We have focused on to predict heart disease and finding the heart attack probabilities.

1.5 Problem Statement

Heart disease is a major contributor to global mortality rates, and timely detection is crucial for managing and preventing this illness. Despite the various solutions provided in recent years, there are still challenges and limitations that need to be overcome. The objective of this research is to introduce a ML framework that incorporates a "Neural

Fuzzy Inference System" and "Genetic Algorithm" for predicting the likelihood of heart attacks in patients.

- Current methods of predicting the disease rely on manual assessment of patient data by healthcare professionals, which usually is time-taking and can cause error. To address this challenge, many ML embedded automated solutions has been proposed in past and more is required to handle the weak area in the provided ones.
- There is a lack of a comprehensive ML framework that uses these techniques to predict the disease in patients, hence, there is a need of an effective method with good accuracy for the prediction of disease.
- Imbalanced datasets or choosing inadequate number of features may lead to inaccurate predictions. In recent work, it has been noticed that 12-14 features have been used for experiment purpose and it leave a scope to do a in depth feature analysis so that more features can be included for study purpose and look around the significance with the identified problem.
- Currently available prediction models might not account for individualized risk factors, such as "family history" of other related disease for example diabetes or heart disease and lifestyle habit like intake of cigarette or alcohol, which may impact the accuracy of predictions. There is a need to study such parameters as well.
- Prediction models should be generalized well to diverse populations, potentially leading to inaccurate predictions for certain groups based on age or gender.
- Integration of the ML framework into existing healthcare systems may pose technical and resource-related challenges. There is a need of exposing trained model for external use so that masses can take advantage of the framework. Exposing framework via APIs could be a good idea as it can be utilized over web platform and solve the purpose.

1.6 Novelty in the proposed work

This study is focused on developing a comprehensive framework aimed at predicting the likelihood of a heart attack. The approach combines the adaptive learning capabilities of a neural network (NN) with an automated decision-making mechanism. The resulting hybrid model, known as NN-GA-FIS, integrates a genetic algorithm (GA) and a fuzzy logic control inference system (FIS). This unified model is specifically designed for cardiovascular disease (CD) prediction. The framework operates in two key stages. The initial stage employs NN-GA, utilizing hyperparameter tuning-based mutation techniques to generate the most effective models. In the subsequent stage, NN is combined with FIS, incorporating supportive inference rules derived from fuzzy knowledge representation for precise decision-making. The integrated NN-GA-FIS model is devised to provide enhanced predictions. Within this framework, a neural fuzzy inference system (NFIS) plays a pivotal role. It encapsulates the training data derived from multi-dimensional functions. An error computing module is incorporated to refine learning instructions based on error measurements. At the beginning, the MFs are defined, and their parameters are dynamically adjusted as needed for optimal operations. The methodology is rigorously tested using sample test cases from the "Cleveland" heart disease dataset. It encompasses both dependable and non-dependable parameters, along with contributing factors and data matrices. Notably, the research incorporates an extensive set of over 13,000 fuzzification rules, enabling precise decision-making. Various normalization processes and planting techniques are employed to ensure the feasibility of calculating heart attack probabilities. As a result, the model achieves an impressive accuracy rate ranging between 94 to 96 percent. This research extends its potential applications beyond prediction, offering the possibility of developing an adaptive advisory system. This could be further augmented through the integration of hardware peripheral circuit devices, potentially revolutionizing preventative healthcare practices. The comprehensive framework, integrating cutting-edge technologies and methodologies, presents a significant advancement in the field of cardiovascular disease prediction.

1.7 Scope of the proposed work

Proposed work will be used to build a framework, which can predict the probability of heart attack with the help of affected medical parameters considered in the proposed work. Later extending the work, alert could be generated and preventive measures could be suggested to patient based on the affected medical parameters to prevent and overcome this disease. The scope of work includes the following tasks:

- *Overview:* The primary objective of this research work is to construct a machine learning framework capable of forecasting the likelihood of a heart attack in individuals. This is accomplished through the integration of a neural fuzzy inference system, a sophisticated computational model that harnesses the power of adaptive learning algorithms. By merging these technologies, the research aims to significantly enhance the accuracy and efficiency of heart attack predictions. The system will comprise of a neural network uses the facility of genetic algorithm passes through a fuzzy inference system to perform the desired prediction. Research framework has been discussed in chapter 3.
- *Data Collection and Pre-processing:* In this phase, a very popular Cleveland heart disease dataset from UCI repository will be used that contains feature such as age, gender, blood pressure, cholesterol levels, heart rate etc. which is very useful in heart disease prediction. Data will be pre-processed to ensure consistency and accuracy of the model. Data collection and processing has been deeply discussed in chapter 4.
- *Model Development:* During this phase, we will develop a neural fuzzy inference system will be developed that uses a genetic algorithm to optimize the model performance. The model will be trained using the pre-processed data collected in the previous phase. Chapter 5 discusses the experiment in details.
- *Testing and Validation:* During this phase, model will be tested for performance and prediction ability will be validated using same dataset applying splitting technique of the dataset in 80:20 train-test ratio.

- *Deployment and Maintenance:* During this phase, ML model will be prepared for deployment, end point will be exposed using a user-friendly interface that can be accessed by healthcare professionals and patients.
- *Limitation:* The proposed research work comes with some advancement in healthcare of patients but it also has some limitations. The framework is designed to predict the probability of heart attack in patients using a neural fuzzy inference system and genetic algorithm, however, the scope of work excludes the development of a mobile application or integration with device or an existing healthcare system. As a result, the ML framework will be accessible only through a web interface and will not be readily available on mobile devices.
- *Benefits to the society:* Early detection and prevention of heart disease in patients can lead to improved health outcomes, minimizing the risk of mishappen and quality of life. From our believe, this system might prove to be a lifesaver tool which will be preferred by masses.

1.8 Thesis Structure

In this thesis, a research work has been proposed which focus on building a framework that will help in prediction of heart attack and advising preventive measures to patients so that they can take inevitable precautions. The research work has been mainly covered in four objectives and seven chapters wherein chapter 1 starts with introduction and overview of the research work, common terms related to heart disease, cause, prevention and etc., common terms related to machine learning and other techniques that has been used in the proposed research work, chapter 2 covers the literature review which provides the detailed overview of existing related work and studies in heart disease prediction and heart attack probability calculation, chapter 3 covers the detail discussion on research objectives and proposed research framework wherein along with these details we also covered a detailed discussed on how we will complete the research objectives using the proposed framework, in chapter 4 and 5 discussion about the data collection, experiments and results can be find wherein we discussed the dataset in detail, obstacles and steps in data collection, and the experiment done with the result

obtained with different variants, chapter 6 discuss the test cases and API integration, conclusion and future work is covered in chapter 7 and finally we concluded the thesis with bibliography and ending words.

1.9 Summary

In this chapter we documented the basic details on the heart disease, its symptoms, medical parameters that can cause heart attack and its prevention. We also discussed the basic of ML algorithms like “SVM”, “KNN”, “LR”, “DT”, “RF”, “NB” with the structuring and working of “NN”, “GA” and “FIS” which are being used in the experiments. In this chapter we also discussed about the motivation and scope of the proposed framework, why we choose it and how it can prove to be a lifesaver for society.

Chapter 2: Literature Review

2.1 Introduction

For any research to carry out, there are several stages which has to be completed one-by-one. In this thesis, so far in chapter 1, we have covered the basics of heart disease and ML techniques. After having the basic idea on the subject, we have arrived to the most important stage of reviewing the existing literature that involves gathering relevant and valid information from credible sources, which can be condensed into a document and serves as the foundation for the study's rationale, which we can say, is mandatory for completion of any research work. Literature review helps in understanding the existing knowledge and also taking out the gaps and limitation of the studies, based on which new study can be proposed keeping old one as the baseline. A good literature review help in building a strong base for research work and solve many purposes like identifying relevant literature, assessing the quality of sources, synthesizing the information and presenting a summary of findings, establishing the research problem and developing relevant questions or hypotheses, determining the appropriate research methodology, identifying gaps or inconsistencies in the existing literature and providing insights into the current state of research and also to suggesting directions for future research. In this chapter we have discussed the already existing knowledge shared by many renowned researchers. We have reviewed many literatures which were similar to the work we have proposed in the thesis to get an idea on their thoughts on this subject and selection of techniques and toolset. Deep analysis has been done after going through the existing studies which helped in building a roadmap, framing the research objectives, selecting the material, methods and tools to accomplish the proposed research work. The chosen research area is very critical in terms of social welfare. Lot of work has been done for the prevention of heart disease and after including AI and ML, the level of research and technique to provide feasible and optimal solution has been increased by the time. Many classical ML techniques like “SVM”, “KNN”, “LR”, “DT”, “RF”, “NB” and others along with advance techniques

from neural network family like ANN, DNN or hybrid solution has helped researchers a lot in finding the finest solution for the prediction of HD and providing suggestive measures to precure it as well. In next few sections we have discussed the literature reviews and comparative analysis of many existing works done till date.

2.2 Existing Studies

The heart disease prediction became a motivating research topic for years and many researches have been accomplished to forecast the disease likelihood with variety of ML methods putting various parameters in mind i.e., heart rate variability (HRV), systolic and diastolic Coronary Heart Failure (CHF), predicting re-hospitalization, destabilization etc. These researches helped both specialized and non-specialized doctors in diagnosing HD and taking suitable decision advising preventive measures. Various popular ML techniques i.e., SVM, KNN, DT, LR, RF, ANN, DNN, Fuzzy logics, Multi-layer perceptron and many more have been experimented to record the observation with popular "Hungarian", "Cleveland", "Long-beach-VA", "Switzerland" and "Z-Alizadeh Sani" dataset, among which "Cleveland" and "Z-Alizadeh" with 303 sample are very popular and preferred in most research works. Apart from above classic algorithms, many hybrid-intelligent systems have also been experimented to improve the performance of model and enhancing the decision-making capabilities with objective to take advantage and eliminate disadvantage from constituent models.

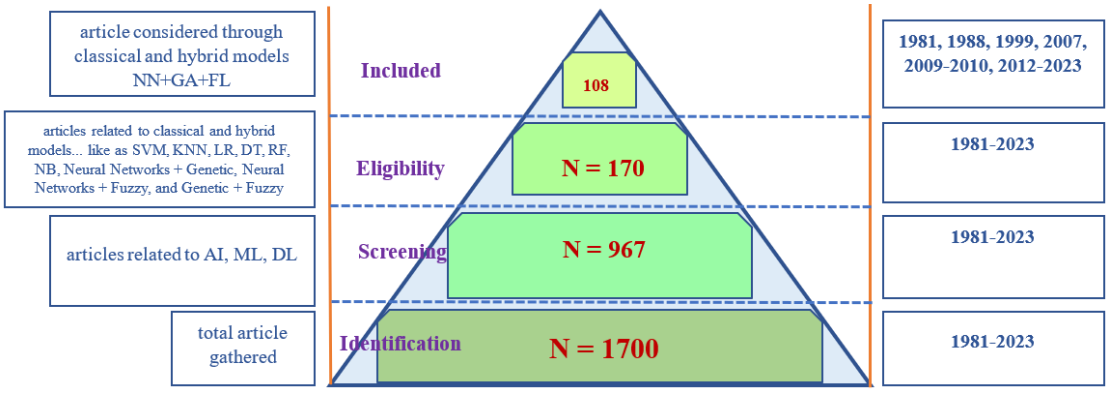


Figure 2.1: Literature search details using PRISMA framework

Figure 2.1 depict the systematic literature search flow using PRISMA framework. This is divided into four steps starting with identification of relevant papers, screening,

eligibility according to research and at last including them for reviews. We follow the systematic literature review flow wherein literature review started with searching the literature in popular databases like SCOPUS, ScienceDirect, Web of Science etc. with the help of keywords associated with the research idea like in our case they are “heart disease”, “heart attack”, “ANFIS” were entered into database to fetch the article list. The total number of 1700 articles were listed on the initial search in the range of year 1981 - 2023. After we get the paper list, they undergo narrowing down with first level of screening to filter the relevant papers based on the research subject area AI, ML and NN which left 967 relevant papers. At third and fourth stage, eligibility check and inclusion of literature was done which filtered those papers which were not accessible or were duplicate/similar in nature or were out of scope i.e., not related to classical or hybrid system, were not related to fuzzy inference or genetic algorithm. So, after removing all such paper we left with 108 papers. Below figure 2.2 shows the year wise selection of research papers.

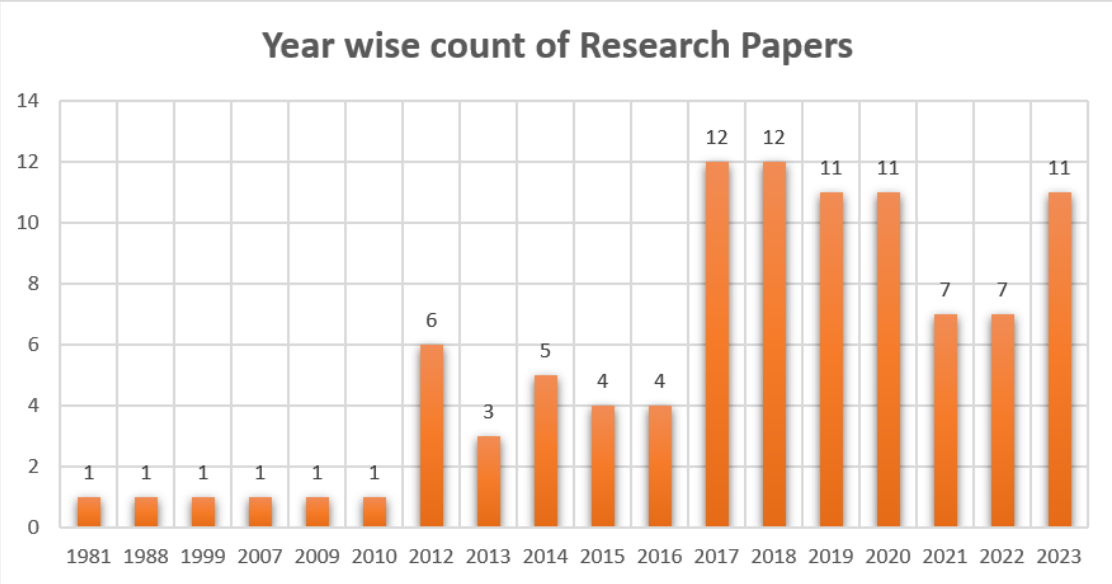


Figure 2.2: Year wise selection of research papers

Shadman Nashif and Amin Ul Haq et al. anticipated a ML system for envisaging heart disease using different algorithms. In a study by Nashif et al., they introduced a cloud-based system that employed the support vector machine technique within the WEKA framework. This system was designed for the real-time tracking of critical parameters

in patients, including temperature, blood pressure, and heart rate [2]. Their model demonstrated remarkable performance metrics, achieving an accuracy of 97.53%, a sensitivity of 97.50%, and a specificity of 94.94% when subjected to 10-fold cross-validation on the UCI Heart Disease database. This indicates a high level of reliability and effectiveness in predicting heart-related conditions using their proposed methodology. On the other hand, Haq et al. evaluated SVM, K-NN, LR, DT, RF, ANN, and NB algorithms on the Cleveland HD dataset from UCI repository with three feature selection methods (mRMR, LASSO, and Relief) using K-fold validation [3]. Their results showed that applying feature reduction techniques can enhance accuracy and reduce execution time. Among all, SVM with mRMR outperformed with 88% accuracy with reduced features. ANN with 16 hidden neurons with Relief feature selection algorithm outperformed in terms of sensitivity with 100%. In another study by Rahul Kumar Jha et. al. presented a result comparison of classification algorithms i.e., SVM, KNN, DT, RF, DNN and NB. Experiment has been conducted over Cleveland dataset on Rapid minor ML tool. Study recorded an accuracy above 93% with DNN which outperformed other ones [5]. In the study conducted by W. Mehrdad Aghamohammadi et al., they introduced a novel hybrid system that integrates Adaptive Neuro-Fuzzy Inference System (ANFIS) with K-fold cross validation and genetic algorithms for the prediction of Heart Disease (HD). This approach was rigorously assessed using the renowned Cleveland dataset from UCI, which encompasses 14 crucial features [7]. The model's performance was meticulously evaluated based on key metrics including accuracy (ACC), sensitivity (SENS), and specificity (SPEC), yielding impressive results of 84.43%, 91.15%, and 79.16% respectively. These findings underscore the promising potential of their hybrid system in effectively forecasting heart disease. Another study by Oluwarotimi Williams Samuel et al. presented a hybrid system that combines “Fuzzy_AHP” and “ANN” to forecast the risk of heart failure [9]. The experiment with a hybrid system combining ANFIS supported by GA algorithm for the prediction of HD using Cleveland dataset for model training was presented by Negar Ziasabounchi and performance was recorded 92.30% accuracy [8]. Author A.V. Senthil Kumar proposed a hybrid system for predicting HD that combines a Fuzzy-Inference-System (FIS) and a neural network (NN) [10]. This system was trained on the UCI Cleveland dataset using a hybrid learning procedure with MATLAB. The model

consists of 5 layers, with the input variables in the first layer, and the internal computation in the second, third, and fourth layers, and the output of the model in the fifth layer. The model achieved an accuracy of 91.83%. Zeinab Arabasadi et al. proposed a study which presented experiment with ANN and GA. Dataset named Z-Alizadeh Sani from UCI has been used in the experiments. The neurons in neural network were structured as 22 X 5 X 1 in input, hidden and output layer comprising of single hidden layer. Weights were generated using GA. Experiment recorded an accuracy of 93.85% [12]. The authors Kaan and Ahmet in their study proposed a recurrent fuzzy neural system equipped with genetic algorithm (GARFNN) to diagnose HD using Cleveland dataset for heart publicly available in UCI repository [13]. The results of the experiment indicated that the GARFNN system outperformed the ANN-Fuzzy_AHP system, with a test set accuracy of 97.78% compared to 91.1% for ANN-Fuzzy_AHP. Another experiment presented by G. S. G. Thippa Reddy et al., a hybrid system integrating fuzzy logic and genetic algorithms (AGAFL), rough set feature selection, and fuzzy rule-based classification was employed to predict heart disease [14]. The experiment was carried out using the UCI dataset in multiple stages. The first step involved feature selection utilizing rough set logic, which was subsequently refined and trained using both traditional and adaptive models proposed in the study. The results revealed an accuracy (ACC), sensitivity (SENS), and specificity (SPEC) of 90%, 91%, and 90% respectively for the proposed (AGAFL) model, surpassing the performance of other models explored in the experiment. This underscores the efficacy of the hybrid AGAFL system in heart disease prediction. Akgul M et al. [15] presented a comparison study between ANN and hybrid GANN and measured the performance over Cleveland dataset. Results were outstanding wherein ANN generated the accuracy 85.02% and GANN outperformed with 95.82% accuracy. The author Sneha Nikam with others proposed a study which performed experiment using neural fuzzy system and generic algorithm. Experiment was done to advance the efficacy of the model [17]. The combination proved to improvise the model efficiency and reduction of error rate. Author Yoichi Hayashi conducted experiment to study the HB levels during the treatment of Anaemia using rule-extraction during pre-dialysis in view of the importance of variations in Hb levels [28]. The authors Yaowei Li, Lina Zhao, Yang Zhang, Yao Zhang, Liuxin Zhan, Li Zhang and Chengyu Liu proposed a study that

combines few CNNs like AlexNet, DenseNet, and SE_Inception_v4 with distance distribution matrix (DDM) to distinguish patient between normal and heart failure category using entropy calculation experimented over "MIT-BIH RR Interval" Databases [29]. "FuzzyGMEn-generated DDM" and "Inception_v4" model outperformed with 81.85% accuracy which is highest of other proposed combinations. The authors Santhanam et al. presented a study concluding an aim to diagnose HD. In their experiments they used hybrid genetic-fuzzy system comprising of GA and fuzzy logic. They achieved satisfactory results [30]. A comparison study on the automation of heart disease prediction using MATLAB was presented by MAbushariah et al. [31]. Different methods, such as the "Multilayer Perceptron (MLP)" structure on the ANN and ANFIS approach, were evaluated in the experiment. Results indicated that the ANN model outdid the ANFIS model with a score of over 87%. The author Azam Davari Dolatabadi et. al.in their study proposed a system for diagnosis of heart-disease using HRV signals. In the study, they experimented HRV signals extracted from ECG and for that they used SVM and PCA to cut the size of pulled out features and to optimize the accuracy. They achieved an accuracy of 99.2% in the experiment [80]. Zerina Masetic and Abdulhamit Subasi in their study "Congestive heart failure detection using random forest classifier" proposed a study and examined five classifiers "KNN, SVM, DT, RF and ANN" for prediction of heart failure with autoregressive (AR) Burg feature extraction method and found best one based on the accuracy [81]. Experiment showed that random forest classifier outperformed with 100% accuracy. The author Moloud Abdar et. al. in their study "A new machine learning technique for an accurate diagnosis of coronary artery disease" proposed a work to predict HD using N2_Genetic optimizer. In their experiment, "Z-Alizadeh Sani" dataset has been used from UCI heart disease repository [82]. They used genetic algorithm and PSO-algorithms following by cross-validation method. Experiment revealed expected result with the 93.08% accuracy and 91.51% F1-score. Shashikant et al. in their study "Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter" proposed a study which compares the performance of ML techniques like LR, DT and RF model to predict the probability of heart attack in smokers [83]. Experiment was performed on the dataset provided by MITU Skillogies Pune, India. Result shows that random forest has outperformed with an ACC, SENS and SPEC of 93.61%, 92.11%

and 95.03% respectively. Sultan Noman Qasem and Monirah Alsaidan in their study "A New Hybrid Intelligent System for Prediction of Medical Diseases" proposed a hybrid system, which includes NN-GSO based model to analyse and compare the results experimented on various critical diseases like Heart, diabetes, breast cancer, thyroid, liver and etc. with the algorithms including "Particle Swarm Optimization based Neural Network" and "Genetic Algorithm based Neural Network" [84]. NN-GSO algorithm outperform for most of the disease but NN-GA give better results for heart. Mohammad Shafenoor Amin et. al. proposed a study "Identification of significant features and data mining techniques in predicting heart disease" to predict HD using seven classification techniques "DT, NB, KNN, SVM, LR, Vote and ANN" and applied with feature selection method over Cleveland and UCI Statlog heart disease dataset [85]. Experiment showed that Vote gives better accuracy for disease prediction. The authors Rajesh Nichenametla, T. Maneesha, Shaik Hafeez and Hari Krishna in their study "Prediction of Heart Disease Using Machine Learning Algorithms" proposed the approach using NB and DT ML algorithms to find the optimal among themselves in predicting heart disease [87]. Naibe Bayes demonstrated best result with large set of data and decision tree show better result for small dataset. The authors Shadman Nashif et. al. in their research "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System" proposed a study which has worked on a cloud-based prediction system using ML SVM-approach and K-fold cross validation method [86]. Experiment showed an ACC of 97.53%, SENS and SPEC of 97.50% and 94.94% respectively. In this experiment, a neural network was trained using weights computed by the "Fuzzy_AHP" system, resulting in 91.10% accuracy. Diman Hassan et al. [89] proposes an approach with a *"pre-trained DNN with PCA and LR"* for feature extraction, dimensionality reduction and model prediction and generated 91% accuracy. Mrs. K. Uma Maheswari and Ms. J. Jasmine [90] presented a study using LR and ANN to find the prediction of HD over Cleveland dataset and obtained an accuracy of 84%. In a study led by C.V. Aravinda et al. [91], a prediction model for heart disease was introduced employing deep learning algorithms, including Neural Networks (NN), alongside random forest and decision tree classifiers. The study demonstrated a commendable accuracy rate of 90% for the developed model. This suggests promising potential for the application of deep learning techniques in

heart disease prediction. Abdollahi et. al. [95] introduced a novel method for group learning in illness categorization. Their approach utilized wrapper-based feature reductions to select the most relevant features. In the field of medical diagnostics, there have been significant advancements in using ensemble learning for heart disease predictions. These developments have proven to be highly accurate and have resulted in reduced treatment costs compared to traditional techniques. The study's findings revealed the crucial role of Thallium Scan and vascular occlusion variables in accurately distinguishing between individuals with heart disease and those who are healthy, achieving an impressive probability of 97.57%. V. K. Sudha and D. Kumar [96] in the study "Hybrid CNN and LSTM Network for Heart Disease Prediction" introduced an innovative approach for predicting heart disease by integrating Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks. This hybrid model outperformed conventional machine learning techniques in terms of accuracy. The combined CNN and LSTM architecture was employed to classify the heart disease dataset into two categories: normal and abnormal. To assess the effectiveness of this novel method, a comprehensive comparison was conducted against several established machine learning algorithms, including Support Vector Machines (SVM), Naive Bayes, and Decision Tree. The results revealed a remarkable accuracy rate of 89%, which was further validated through k-fold cross-validation. A study proposed by Aymen A. Altae and Abdolvahab Ehsani Rad [97] explores the performance of various classifiers, namely extreme learning machines, enhanced fast learning networks, support vector machines, and decision trees, for the prediction of heart disease and also demonstrates a hybrid model by combining the Particle Swarm Optimization Algorithm (PSOA) with the learning classifiers. The hybrid diagnostic method achieves an ACC of 93% (PSO-ELM), 96% (PSO-EFLN), 91% (PSO-SVM), and 93% (PSO-DT) based on ten runs using a 10-fold cross-validation (CV) approach. M. Raihan et. al. [98] have devised a pragmatic approach to prognosticate the risk of myocardial infarction by harnessing the capabilities of smartphones. This groundbreaking method empowers the masses to undergo cardiologist evaluations, thereby mitigating the occurrence of abrupt cardiovascular fatalities. Their pioneering endeavour entails the development of an Android-based prototype software that seamlessly amalgamates comprehensive clinical data derived from an extensive cohort

of 787 patients. This dataset is meticulously scrutinized and correlated with an array of perilous determinants encompassing blood pressure, glucose levels, serum cholesterol concentrations, smoking behaviour, family medical history, adiposity, and psychological stress, alongside prodromal symptomatic manifestations. Subsequently, the risks are judiciously stratified into low, medium, and high categories to facilitate a thorough assessment of ischemic heart disease (IHD). Upon diligent comparison and classification of the patients' data, a profound association emerges between the incidence of cardiac events and the juxtaposition of the low-high and medium-high risk categories (p=0.0001 and 0.0001, respectively), affirming the robustness of this innovative diagnostic paradigm. In a comprehensive study proposed by Najmul Hasan et al. [99], the primary objective was to discern the most effective "feature selection approach" for predicting CVD. The investigation entailed a meticulous comparison of various feature selection methods, encompassing filter, wrapper, and embedding techniques. Subsequently, a feature subset was derived through a systematic boolean process-based condition, employing a two-stage retrieval process. To ascertain the optimal predictive analytics model, a range of classifiers including "Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Naive Bayes, and eXtreme Gradient Boosting (XGBoost)" underwent evaluation for their relative accuracy. As a baseline for comparison with all features, an ANN was employed. The findings unveiled the XGBoost classifier, when paired with the wrapper technique, as the most adept in delivering highly accurate predictions for cardiovascular ailments, achieving an outstanding accuracy rate of 73.74%. SVM followed closely with an accuracy of 73.18%, and the ANN demonstrated 73.20% accuracy, further affirming the superiority of the XGBoost approach. In a research study led by Shah et al. [100] "Heart Disease Prediction using Machine Learning Techniques", the main objective was to establish a robust predictive model for cardiovascular disease utilizing cutting-edge machine learning techniques over "Cleveland heart disease dataset" encompassing 303 instances and 17 attributes, obtained from the "UCI machine learning repository". The study accentuates the potential efficacy of ML techniques in forecasting CVD and underscores the significance of judiciously selecting appropriate models and methodologies to attain optimal outcomes. The study included a diverse array of ML classification methods, comprising "naive Bayes, decision tree, random forest, and k-

nearest neighbor (KNN)" algorithms. Remarkably, the findings of this study showcased that the KNN exhibited the higher ACC reaching 90.8%. In their paper titled "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", Senthilkumar Mohan et al. [101] introduced a novel approach named "Hybrid Random Forest with a Linear Model" that aims to enhance the accuracy of cardiovascular disease prediction by leveraging ML techniques. The proposed method combines the power of "Hybrid Random Forest" and a Linear Model, employing ML method to identify significant features from the UCI Cleveland dataset. This study utilized "R Studio" as the platform for implementation, and the "Hybrid Random Forest with a Linear Model (HRFLM)" integrated ANN with back propagation alongside 13 medical features as input. To optimize the prediction model, various feature combinations were introduced, and a range of established classification techniques were employed, including "Naive Bayes (NB), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), Voting Classifier (VOTE), Logistic Regression (LR), Deep Neural Networks (DNN), Generalized Linear Model, and Gradient Boosted Trees". The experimental results demonstrated a significantly improved performance level, with the proposed hybrid model HRFLM achieving an impressive accuracy level of 88.7% for heart disease prediction, surpassing all other models under consideration. This research provides valuable intuitions into the application of advanced ML methods in heart disease prediction, showcasing the potential of hybrid models to enhance accuracy and contribute to more effective and reliable diagnostic tools in healthcare. In the research, "Heart function monitoring, prediction and prevention of Heart Attacks: Using Artificial Neural Networks", D. K. Ravish et al. [102] developed a NN-based predictive model for heart attack diagnosis, utilizing clinical and electrocardiogram (ECG) data to ensure accurate identification of heart attacks and potential abnormalities. The dataset used for experimentation was sourced from the "Physio Net ECG" database. The experiment was conducted using MATLAB, incorporating various ECG features such as "QRS duration", "R-R interval", "P-R interval", and "Q-T interval", in conjunction with "R-wave", "P-wave", and "T-wave" durations. Moreover, relevant clinical data, including "blood pressure", "blood sugar", "heart rate", "cholesterol levels", "age factor", and lifestyle habits like smoking and drinking, were considered for comprehensive analysis. The overall process was divided into three key stages.

Initially, the patient's ECG was acquired using standard “3-lead pre-jelled” electrodes. The acquired electrocardiogram (ECG) signals underwent a rigorous processing phase, involving amplification and meticulous filtering to eliminate any extraneous noise that might have been introduced during the acquisition process. Subsequently, the analog data was efficiently converted into a digital format through an Analog-to-Digital (A/D) converter, a crucial step in addressing potential uncertainties in the data. Moreover, pivotal clinical parameters including mean arterial pressure (MAP), fasting blood sugar (FBS), heart rate (HR), cholesterol levels (CH), and demographic factors such as age and gender were meticulously gathered. These parameters formed the bedrock of the subsequent analysis. This amalgamation of ECG and clinical data then served as the training dataset for the neural network model, facilitating the classification of heart diseases and the prediction of potential cardiac abnormalities. The Artificial Neural Network (ANN) played a pivotal role in training this model. Simultaneously, the genetic algorithm, known for its prowess in evaluating fitness functions, was employed. In this context, it factored in the four clinical parameters utilized in the experiment, ensuring a robust assessment of the model's predictive capabilities. This comprehensive approach demonstrates the potential of advanced machine learning techniques in accurately predicting heart attacks and assessing cardiac health, thereby providing valuable insights into preventive and diagnostic measures in cardiovascular medicine. In the research paper titled “Prediction of Cardiovascular Disease Using Machine Learning Algorithms”, authored by Kumar G Dinesh, K Arumugaraj et al. [103], a comprehensive comparative analysis of popular machine learning methods was conducted using R software. The study utilized a heart disease database consisting of 920 patient records and 76 attributes, out of which 14 relevant attributes were selected for experimentation, obtained from the “Cleveland”, “Hungarian”, “Switzerland”, and “Long Beach VA” datasets available in the “UCI repository”. Data pre-processing techniques, including noise removal, handling missing data, and attribute classification for predictive and decision-making purposes, were meticulously applied to ensure data quality and consistency. Performance evaluation of the predictive models was accomplished through various measures such as classification, accuracy, sensitivity, and specificity analysis. The primary objective of this research was to propose a robust prediction model capable of determining whether individuals have cardiovascular

disease or not, providing crucial awareness and diagnosis in this domain. The ML algorithms, namely “Support Vector Machine”, “Gradient Boosting”, “Random Forest”, “Naive Bayes”, and “Logistic Regression”, were extensively evaluated for their prediction capabilities using the region-specific dataset. Remarkably, the logistic regression algorithm outperformed others, attaining an impressive accuracy of 86.5%, thereby demonstrating its superior potential in predicting cardiovascular disease. This study contributes to the advancement of predictive analytics in cardiovascular health, leveraging machine learning techniques to enhance disease prognosis and decision-making processes. The proposed model showcases promising outcomes, highlighting the significant impact of employing sophisticated algorithms and meticulous data preparation in achieving accurate predictions for cardiovascular disease. The research presented in the study titled “Heart Diseases Prediction based on Stacking Classifiers Model” by Subasish Mohapatra et. al. [104] introduces a novel predictive model for heart disease prognosis. This model employs a stacking approach involving both base-level and meta-level classifiers, effectively amalgamating the strengths of diverse classifiers chosen through an iterative process. The experimentation was conducted on the renowned Cleveland UCI Heart Disease Dataset, utilizing 13 pertinent features for training and assessment purposes. A collection of heterogeneous learners, including RF, MLP, KNN, ET, XGB, SVC, SGD, ADB, CART, and GBM, were synergistically harnessed to produce a robust prediction model. The final model demonstrated a remarkable accuracy rate of 92% in prediction, accompanied by a precision score of 92.6%, a sensitivity measure of 92.6%, and a specificity value of 91%. The research conducted by Chetan Sharma et al. [105] in the study titled “Early Stroke Prediction Using Machine Learning” systematically delves into various contributing factors associated with the onset of strokes. The dataset utilized in this inquiry is meticulously sourced from an openly accessible repository, and a diverse array of classification algorithms is harnessed to prognosticate the likelihood of imminent stroke episodes. Notably, the employment of the random forest algorithm has yielded a remarkable precision rate of 98.94%. In this scholarly work “Machine Learning for Brain Stroke Prediction” authored by Shehzada Mushtaq et al. [106], an advanced algorithm for stroke prediction has been introduced. The study leverages a comprehensive dataset, encompassing critical parameters linked to cerebral stroke, including but not limited to

age, body mass index (BMI), gender, symptoms of heart disease, and smoking status, to construct a robust predictive model. The dataset underwent meticulous preprocessing procedures, addressing missing data, handling categorical variables, and ensuring dataset balance. Various classification algorithms, spanning “Naive Bayes”, “Logistic Regression”, “XgBoost”, “Decision Trees”, “AdaBoost”, “K-Nearest Neighbor”, “Random Forests”, “Voting classifier”, and “Support Vector Machines”, were rigorously assessed in this experimentation. The empirical findings notably highlight the superiority of the Support Vector Machine algorithm, which demonstrated exceptional performance, achieving an impressive accuracy rate of 99.5%. Enhancing the efficiency of heart disease diagnosis is a crucial global healthcare objective, and data mining algorithms have played a vital role in achieving this goal. Gangavarapu Sailasya and Gorli L Aruna Kumari in paper titled "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms" [107] incorporates a range of medical parameters. Leveraging sophisticated machine learning algorithms such as "Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors, Support Vector Machine, and Naive Bayes Classification", five distinct models were meticulously trained to enhance prediction precision. Notably, the Naive Bayes algorithm exhibited superior performance in this regard, achieving an impressive accuracy rate of approximately 82.92%. Asif Newaz et al. [109] in the study “Survival prediction of heart failure patients using machine learning techniques”, conducted a research endeavour aimed at developing a “robust decision-support system” for the likelihood of heart failure in patients, leveraging advanced machine learning techniques. The study utilized a dataset sourced from the "Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad, Pakistan," accessible through the “UCI Repository” [108, 111]. To enhance the accuracy of survival prediction, the researchers employed two distinct feature selection methods - the “Chi-square test” and “Recursive Feature Elimination”. Further, a “5-fold cross-validation” scheme was applied to partition the data into train-test sets for evaluation. An innovative sampling strategy was integrated into the ensemble learning framework, enhancing the robustness and balance of the random forest classifier. This modification enabled the model to effectively address the inherent data imbalance, resulting in a more reliable and widely applicable outcome with significantly improved accuracy. The proposed

model's performance was then compared with several standard classifiers, including RF, SVM, KNN, LR, and AdaBoost. The results demonstrated that the proposed model achieved an impressive accuracy rate of 76.25%, showcasing its efficacy in accurately identifying heart failure patients at risk of survival issues. In their investigation titled "Feature Analysis of Coronary Artery Heart Disease Data Sets", Randa El-Bialy et. al. [112] employed rapid decision tree algorithms, including the pruned C4.5 tree, to mitigate potential challenges arising from missing, erroneous, or inconsistent data encountered during the data collection process, wherein the resultant decision trees were derived from distinct datasets and subsequently juxtaposed. This scrutiny encompassed datasets like Cleveland, Hungarian, V.A. heart disease, and the statlog project heart disease dataset, each characterized by a set of 13 distinct features. The analysis strategically culled common features from these datasets for disease assessment, ultimately yielding a commendable classification accuracy rate of 78.06% with the amalgamated dataset, surpassing alternative approaches. In the research paper titled "Machine Learning Approach for Estimation and Novel Design of Stroke Disease Predictions using Numerical and Categorical Features," Santosh Kumar Satapathy et al. [113] conducted a comprehensive analysis of ML methodologies to determine the optimal model for providing both medical practitioners and patients with insightful prognostications regarding the likelihood of experiencing a cerebral stroke in the future. The experimental phase encompassed the utilization of the "Random Forest Algorithm (RF), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR) Algorithm" to train distinct models, subsequently enabling a comparative assessment of their predictive performance. Notably, among these algorithms, the Decision Tree exhibited the highest accuracy at 94.50%. In the article "Earlier identification of heart disease using enhanced genetic algorithm and fuzzy-weight based support vector machine algorithm", a novel approach using "EGA (Enhanced Genetic Algorithm) based FWSVM (Fuzzy Weight updating Support Vector Machine)" has been introduced by authors G. Sugendran and S. Sujatha [114] to overcome the challenges associated with this task. The algorithm consists of three key phases: preprocessing, feature selection, and classification. By incorporating enhanced genetic algorithms and fuzzy weight updating support vector machines, EGA + FWSVM aims to accurately predict the early stages of heart disease. Through

experimental analysis, the results have demonstrated that the EGA + FWSVM algorithm surpasses existing techniques in terms of heart disease detection performance. This superiority is evident in various evaluation metrics, including accuracy, recall, precision, f-measure, specificity, and time complexity values. Remarkably, the algorithm achieves an accuracy rate of 92.22%. C. M Chethan Malode et. al. [115] presented a study using soft set creator, fuzzy rule generator, SVM classifier and decision maker for the prediction of heart attack in adolescents using dataset having four parameters - BP, sugar, cholesterol and nicotine presence. Observation found that if any one of health parameter (HP) is abnormal then heart attack probability is 0 - 32%, if two are affected then the probability lies between 33 - 55%, if HP increase to three, the probability remains in 56 - 80% and if all parameter is abnormal, probability has the range between 81 and 100%. Research observation also states the percentage of adolescents exhibits abnormality based on the medical parameter i.e., 17% of adolescents exhibits abnormality in cholesterol, 28% in BP, 20% in nicotine level and 30% in diabetes. E. Maraj and S. Kuka [116] in the paper “Prediction of Coronary Heart Disease Using Fuzzy Logic: Case Study in Albania” presents an innovative fuzzy logic-based model for the prediction of coronary heart disease, incorporating an array of seven meticulously chosen input variables and a single output variable. The model was developed and evaluated on a dataset comprising 30 patients from Albania, obtained from the esteemed University Hospital Centre "Mother Teresa" in Tirana. Employing the sophisticated fuzzy logic toolbox, this pioneering research harnesses a comprehensive range of factors including “blood pressure, cholesterol levels, physical activity, diabetes status, age, BMI, and smoking habits” as crucial inputs, while the output variable pertains to disease classification. By diligently selecting appropriate fuzzy sets and meticulously designing membership functions, the model's efficacy is maximized. The defuzzification process employs the meticulous centroid method to extract crisp results. This meticulous research amalgamates scientific rigor with computational analysis, culminating in a profound contribution to the field of coronary heart disease prediction. In the research “Diagnosis of Coronary Artery Disease Using an Artificial Intelligence-Based Decision Support System”, Noor A. S. et al. [117] presented a novel fuzzy system-based framework and conducted an empirical investigation employing diverse machine learning techniques, including FDSS, MLP-

ANN, KNN, RIPPER, across the Cleveland, Hungarian, Long Beach, and Switzerland datasets sourced from the UCI repository. The Weka tool was utilized for predicting heart disease and evaluating performance metrics. Notably, the FDSS method demonstrated the highest accuracy, achieving an impressive 87% accuracy rate. Muhammad, L.J. and Algehyne, E.A. [118] conducted a comprehensive investigation titled "Fuzzy based expert system for diagnosis of coronary artery disease in Nigeria," where they developed an advanced fuzzy-based expert system to predict heart disease automatically. The study involved processing a dataset comprising 506 records with 12 attributes obtained from "General Hospitals in Kano State, Nigeria," which was meticulously prepared, cleaned, and transformed into the Weka readable file format. To transfer the knowledge of human experts to the expert system's knowledge base, an improved C4.5 data mining algorithm was employed. Subsequently, a performance evaluation system was executed, revealing that the expert system exhibited an impressive overall ACC, SENS, and SPEC of 94.55%, 95.35%, and 95.00%, respectively. These exceptional results validate the system's reliability and its capability to efficiently diagnose both negative and positive cases of coronary artery disease (CAD) patients. In the scholarly work titled "A Clinical Decision Support System: Risk Level Prediction of Heart Disease using Weighted Fuzzy Rules" [119], P.K. Anooj introduces an innovative "Clinical Decision Support System (CDSS)" grounded in weighted fuzzy rules for the prediction of heart disease. The CDSS leverages patient-specific clinical data, drawing insights from 14 distinct features extracted from the Cleveland, Hungarian, and Switzerland datasets, each sourced from the UCI repository. Constructed within the MATLAB environment, the proposed method operates through two pivotal phases: the initial phase entails an automated process for the formulation of weighted fuzzy rules, while the subsequent phase involves the establishment of a fuzzy rule-driven decision support framework. In the preliminary phase, techniques encompassing mining, attribute selection, and attribute weighting were harnessed to derive the weighted fuzzy rules. These rules, a vital foundation of the CDSS, guide its predictive capabilities. Moreover, the fuzzy system's architecture was meticulously designed, meticulously aligning with the prescribed weighted fuzzy rules and the designated attributes. Employing the "Mamdani fuzzy inference system" and adopting the centroid of area defuzzification strategy, the experiment incorporated a

comprehensive fuzzification and defuzzification process. Each input fuzzy set was characterized by four distinct membership functions denoting “Very Low (VL), Low (L), Medium (M), and High (H)”, while the output fuzzy set comprises two membership functions, namely “Low (L) and High (H)”. The fuzzification strategy hinges on the utilization of triangular functions for each membership function, further enhancing the system precision. Subsequent to these design considerations, the proposed method underwent training process and the performance was subsequently benchmarked and the results evaluated an accuracy of 0.509901, 0.715054, and 0.364706 on the training set, and 0.623529, 0.469388, and 0.512195 on the testing set for Cleveland, Hungarian, and Switzerland datasets, respectively. These outcomes underscore the efficacy of the weighted fuzzy rule-driven CDSS and its potential to significantly contribute to the domain of heart disease prognosis and diagnosis. In the comprehensive investigation titled "Comparative Study of Fuzzy Rule-Based Classifiers for Medical Applications," Anna Czmil [120] delves into an examination of 16 distinct rule-based fuzzy logic algorithms encompassing “1R-C”, “C45-C”, “C45Rules-C”, and more, across a spectrum of 12 clinical datasets relating to conditions like Heart, Breast Cancer, Diabetes, Appendicitis, and Hepatitis. The study focused on the application of Fuzzy Rule-Based Classifiers and the consequent rule generation. Notably, OIGA-C emerges as the leading performer with an accuracy rate of 86%, while Ripper-C achieves the highest AUC score at 73%. The findings highlight GPR's ability to yield succinct and interpretable rules without compromising classification efficacy. The objective of the study, titled "A Rule-Based Expert System to Assess Coronary Artery Disease Under Uncertainty," [121] was to introduce an innovative Clinical Decision Support System (CDSS) aimed at enhancing the precision of medical analysis compared to current systems. The paper advocates for a Rule-Based Expert System (RBES) incorporating five distinct Belief Rules, capable of prognosticating five distinct stages of CAD (Class A-E), encompassing normal, unstable angina, Non-ST and ST segment elevation myocardial infarction, and silent ischemia respectively. The dataset, sourced from the “National Heart Foundation, Bangladesh”, encompasses attributes “blood pressure, stress levels, blood sugar levels, lipoprotein levels, triglyceride levels, age, body mass index, dietary habits, smoking status, family medical history, and ethnicity”, stratified into five classes denoted as “Physiological”, “Pathological”, “Demographical”,

“Behavioural”, and “Non-Modifiable” risk factors. The ultimate output is derived through a synthesis of all Belief Rule Bases (BRBs), employing Evidential Reasoning (ER). Performance metrics include success, error, failure, and false omission rate. The proposed expert system achieves an average ACC of 89.90%, surpassing other existing CDSSs. Notably, the success rate in predicting “Class C” type heart disease was the highest (94.08%) among the five classes. Conversely, “Class E” prediction demonstrated the lowest success rate (merely 50%), indicating its inherent complexity. In the study titled "A Fuzzy Rule-Based Decision Support System for Cardiovascular Risk Assessment," Gabriella Casalino et al. [122] introduced a novel fuzzy rule-based framework designed to augment expert decision-making in the realm of cardiovascular diseases, a domain of considerable medical diagnostic significance. The evaluation encompassed real-world data derived from the examination of 116 individuals. Preliminary experimental outcomes, spanning both healthy and afflicted subjects, underscore the efficacy of the fuzzy system in emulating expert decisions. Impressively, it demonstrated an accuracy rate of 91%. The research titled “Utilizing Machine Learning Algorithms for the Development and Prognostication of Medication for Cardiac Arrest Early Warning System” introduces a novel method named "Medication for Cardiac Arrest Early Warning System (MCAEWS)" [123]. This innovative system not only facilitates prompt diagnosis by medical practitioners and immediate alerts but also exhibits heightened sensitivity, reduced false positive occurrences, and lowered mortality rates. The study's findings underscore the potential transformative impact of employing sophisticated data-driven methodologies in the realm of healthcare prognostication and early warning systems for cardiac arrest. The empirical data utilized for this study were sourced from “The Emergency Department of the National Taiwan University Hospital (NTUH)” spanning the duration “January 2014 - December 2015”. The dataset having 23 features was meticulously dissected, categorized into CPR and non-CPR groups, and subsequently subjected to in-depth analysis. A series of comprehensive experiments were then conducted employing a diverse array of machine learning algorithms. The outcomes were meticulously tabulated and assessed for robustness and efficacy. Among the diverse algorithms evaluated, the Random Forest Algorithm emerged as the standout performer. It exhibited remarkable prowess, particularly when contrasted with other techniques such as the Logistic Regression

Algorithm, Decision Tree, and Extreme Random Tree and displayed a significantly elevated performance level, characterized by an impressive Area Under the Curve (AUC) score of 0.98, along with an AUP (Area Under Precision-Recall curve) value of 0.23. The author Hameed A.Z. and Ramasamy B. et. al. in “Efficient hybrid algorithm based on genetic with weighted fuzzy rule for developing a decision support system in prediction of heart diseases” [124] expounds upon a clinical decision support system designed for the purpose of computer-aided heart disease determination, encompassing a weighted fuzzy rule approach in conjunction with a genetic algorithm. This comprehensive study leverages datasets from the renowned Cleveland, Hungarian, and Switzerland repositories available within the esteemed UCI database. In order to surmount the challenges associated with feature selection, an astutely devised genetic algorithm is meticulously employed to elicit highly informative features via an intricate stochastic inquiry process. Subsequently, a meticulously crafted weighted fuzzy framework is meticulously constructed, ingeniously incorporating salient attributes meticulously gleaned from the aforementioned datasets. The proposed framework synergistically capitalizes on the manifold advantages bestowed by the fuzzy rule strategy, while concurrently achieving noteworthy strides in learning efficacy through the judicious application of the refined weighted methodology. The empirical findings gleaned from this research evince the predictive prowess of the genetic algorithm and fuzzy system in the domain of heart attack prediction. Remarkably, the experimental results unequivocally ascertain accuracies of 68%, 48%, and 55% when confronted with test data derived from the Cleveland, Hungarian, and Switzerland datasets, respectively. The author Purushottam Sharma and Kanak Saxena [125] in the article "Application of fuzzy logic and genetic algorithm in heart disease risk level prediction" introduces a comprehensive study proposing an advanced identification framework for coronary illness based on a sophisticated weighted fuzzy standard. The study presents a cutting-edge clinical decision support system that leverages weighted fuzzy rules to yield accurate conclusions regarding coronary illness, utilizing insights gleaned from comprehensive clinical data. The heart disease risk level prediction system, which integrates fuzzy logic and genetic algorithms, encompasses two pivotal stages which comprises of "an automated methodology for the generation of refined weighted fuzzy rules" and "the establishment of a robust heart disease risk level prediction system based

on a synergy between fuzzy rules and genetic algorithms". The fuzzy framework is meticulously devised in accordance with meticulously selected weighted fuzzy standards and superior attribute cases. This groundbreaking study introduces an efficient system that adeptly identifies the essential parameters necessary to predict the risk level of patients based on a myriad of pertinent health-related attributes. The primary objective of this research endeavour is to empower non-specialized medical practitioners with the ability to make well-informed decisions regarding the risk level associated with coronary illness. The performance of the proposed framework is rigorously evaluated and meticulously compared, particularly in terms of rule accuracy. The results unequivocally demonstrate the remarkable potential exhibited by the system in accurately predicting the risk level of coronary illness. A comprehensive comparative analysis is conducted, effectively juxtaposing the outcomes derived from various classical methods such as decision trees, naive Bayes, and K-nearest neighbors (KNN), against those generated by the proposed weighted fuzzy rules and fuzzy logic approach, augmented by the utilization of genetic algorithms. Remarkably, the heart disease risk level prediction system, bolstered by the fusion of fuzzy logic and genetic algorithms, triumphs over all other competing methods, achieving an impressive accuracy rate of 88.11%. This notable achievement solidifies its position as the most accurate and superior method among the alternatives. Nagarajan R. and Thirunavukarasu R. [126] in their research paper “A neuro-fuzzy based healthcare framework for disease analysis and prediction” introduced an innovative neuro-fuzzy healthcare framework, which effectively preprocesses healthcare records and conducts disease prediction. The framework adopts a stratified approach that encompasses several essential tasks, including healthcare data preprocessing, normalization achieved through a comprehensive fuzzification process, disease prediction by employing well-defined rules, and de-fuzzification to obtain valuable insights pertaining to the predicted disease. The design of the fuzzy rule base is meticulously optimized to enhance the decision-making process. To assess its effectiveness, the proposed system is thoroughly validated through an experimental setup and benchmarked against fuzzy-based models and linguistic neuro-fuzzy models with feature extraction. Notably, the neuro-fuzzy based method proposed in this study attains an impressive accuracy value of 87.7%, surpassing the performance of existing methodologies. In the study titled “Heart

Disease Prediction using Hybrid Machine Learning Model", conducted by M. Kavitha et al. [127], a novel approach in machine learning has been introduced for the prediction of heart disease, utilizing the Cleveland heart disease dataset comprising 13 distinct features. Throughout the experimentation phase, three machine learning algorithms were applied: RF, DT and a hybrid model amalgamating RF and DT. The outcomes of the experiments unveiled a remarkable 88.7% accuracy rate attained by the heart disease prediction model, with the hybrid model demonstrating superior performance when compared to the other models. In the research paper titled "An Intelligent Hybrid Classification Algorithm Integrating Fuzzy Rule-Based Extraction and Harmony Search Optimization: Medical Diagnosis Applications," Seyed Mohsen Mousavi et al. [128] introduced an advanced classification method that combines fuzzy rule-based techniques with Harmony Search (HS) and heuristic algorithms for intelligent medical dataset classification. This approach involved the application of two orthogonal and triangular fuzzy systems to elucidate clinical dataset attributes. The Harmony Search algorithm was synergistically employed with a heuristic to generate optimal fuzzy rules within the fuzzy rule-based systems. To evaluate the algorithm's efficacy (TFHS and OFHS), nine prevalent medical datasets including "Cleveland", "WBCD", "Spectfheart", "Twonorm", "Saheart", "Pima", "Haberman", "New-Thyroid", and "Heart" were utilized. A nested cross-validation strategy was implemented to strike a balance between model overfitting and underfitting, encompassing both "outer five-fold CV" and "inner CV". Notably, OFHS demonstrated commendable performance, achieving 71% accuracy with the Cleveland dataset and 82% with Saheart. On average, it attained an impressive accuracy of 88%. The author Oumaima Terrada et. al. [129] in "A fuzzy medical diagnostic support system for cardiovascular diseases diagnosis using risk factors" proposed a novel decision tree-based solution to handle a higher number of rules, aiming to enhance the overall efficiency. The experiment was conducted on the Cleveland dataset within the MATLAB environment. The Mamdani inference method, well-known for its broad adoption in various fuzzy systems, was selected, and a triangular membership method was applied to attributes such as BP, BMI, and Triglyceride. The central aim of this study is to develop a robust fuzzy expert system (FES) for the accurate diagnosis of heart diseases. This FES is constructed using a fuzzy logic approach, leveraging key clinical diagnosis parameters and cardiovascular

risk factors. To create the fuzzy rules base, specific criteria were considered, taking into account the severity of heart disease. The medical system underwent rigorous testing and validation, revealing outstanding efficiency and accuracy in the developed approach. In the study titled “Development of an Adaptive Weighted Fuzzy Rule-Based System for Heart Disease Risk Level Assessment”, author Animesh Kumar Paul et al. [130] has devised an automated diagnostic system incorporating advanced techniques like GA and MDMS-PSO to pedict the risk level of heart disease. Five datasets, specifically Cleveland, Hungarian, Switzerland, Long Beach, and Heart, sourced from the UCI repository, each consisting of 14 distinct features, were meticulously employed for experimentation. Data underwent thorough preprocessing, and attribute selection was meticulously executed utilizing statistical methodologies such as Correlation coefficient, R-Squared, and the Weighted Least Squared (WLS) method. Weighted fuzzy rules were systematically constructed based on the selected attributes, employing the power of GA. To further optimize the membership functions (MFs) of the fuzzy system, the innovative MDMS-PSO was deployed. Triangular MFs were thoughtfully incorporated into this process. An ensemble of fuzzy systems emerged from the collective knowledge base, meticulously blending various local fuzzy systems. It's noteworthy that for each dataset, a distinct attribute selection list and MF tuning values based on the selected attributes were meticulously devised. The empirical analyses substantiate the efficiency of this hybrid model in adeptly managing the inherent vagueness in knowledge and the uncertainties associated with decision-making, leading to enhanced accuracy across various publicly accessible heart disease datasets. The ensemble-based fuzzy Decision Support System (DSS) consistently delivered promising classification results when subjected to diverse attribute selection methods. Interestingly, among these methods, the proposed technique demonstrated superior results, particularly excelling when employing the R2 selection method. For R2 selection, this innovative approach attained remarkable accuracies, including 92.31% for the Cleveland dataset, 95.56% for the Hungarian dataset, 89.47% for the Switzerland dataset, 91.80% for the Long Beach dataset, and 92.68% for the Heart dataset. The investigation in the study "Heart Disease Diagnosis Using Tsukamoto Fuzzy Method" conducted by M. Faris Al Hakim et al. [131] endeavours to assess the risk of heart disease utilizing the Tsukamoto method, incorporating eleven input

parameters encompassing cholesterol levels, blood pressure, and ECG readings, among others. Simultaneously, the output encompasses various categories ranging from 'healthy' to 'very large'. The procedural framework encompasses four primary phases, which encompass a thorough literature survey, the design of a fuzzy inference system, the application of the Tsukamoto fuzzy logic, and a comprehensive evaluation. The study ultimately deduced that the Tsukamoto method's fuzzy logic can be applied to gauge the risk of heart disease, albeit with a model performance currently confined to an accuracy metric of 58%. The article titled "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods", Osman Taylan et al. [132] introduced a comprehensive methodology integrating machine learning, neuro-fuzzy, and statistical techniques for the anticipatory assessment of cardiovascular diseases (CVDs). Employing adaptive neuro-fuzzy and statistical methodologies, along with KNN and NB classifiers, the study conducts an in-depth prediction of seventeen CVD risk determinants. The experimentation phase was executed through MATLAB software, leveraging a dataset retrospectively sourced from the medical archives of family medicine and cardiology clinics within a Saudi Arabian university hospital. This dataset encompassed records of 159 patients aged 16 and above, who sought cardiological consultation due to persistent cardiac symptoms over a period spanning from 6 June 2020 to 10 October 2020. The study employed a “Sugeno-type fuzzy rule-based ANFIS” model, integrating Gaussian membership functions and harnessing the learning potential of the BPNN algorithm. Comparative analysis with established machine learning techniques was performed on authentic CVD data gleaned from a hospital environment. This comprehensive inquiry demonstrated that the ANFIS model achieves a notable 96.56% accuracy during the training phase, with SVR closely following at 91.95% prediction accuracy. The study titled “A Genetic-Neuro-Fuzzy inferential model for diagnosis of tuberculosis” presented by Mumini Olatunji Omisore et al. [133] introduced an innovative approach for tuberculosis diagnosis, employing the “Genetic-Neuro-Fuzzy Inferential method (GNFIS)” to establish an intelligent decision support system. The system was designed to aid medical professionals in delivering precise, timely, and cost-effective TB diagnoses. The performance assessment was conducted through a case study involving ten patients at “St. Francis Catholic Hospital Okpara-In-Land” in Delta State, Nigeria.

The methodology integrates diverse computational techniques, encompassing Neural Networks, Genetic Algorithms, Fuzzy Inference Systems, Triangular Membership Functions, and the Centroid of Area (CoA) method for defuzzification. Each genetic chromosome composed of 24 genes, with each gene denoting the connection weight of a diagnostic variable, expressed in a binary format of 1 bit. The obtained results reveal a sensitivity of 60% and an accuracy of 70%, aligning well with the predefined benchmarks set by domain experts. The objective of the research presented in the paper "A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS," authored by Mohammad Ayoub Khan et al. [134] was to discern the fundamental attributes of heart disease prediction by harnessing advanced machine learning techniques with a primary focus on enhancing the precision of prognostication. To accomplish this objective, a novel Internet of Medical Things (IoMT) framework was devised for heart disease diagnosis, employing a fusion of MSSO and an adaptive neuro-fuzzy inference system (ANFIS). The proposed method was experimented using the Cleveland dataset procured from the UCI repository, encompassing a set of 13 distinctive input features. The data preprocessing stage ensued, followed by feature selection through the Levy crow search algorithm and ultimately, classification utilizing the devised MSSO-ANFIS model. The optimization of learning parameters was executed via the MSSO approach contributing to the refinement of ANFIS's predictive capabilities. This proactive methodology facilitated the identification of cardiac conditions via the classification of sensor-derived data, seamlessly orchestrated by the MSSO-ANFIS paradigm. Rigorous simulation and meticulous analysis substantiated the robust efficacy of MSSO-ANFIS in the domain of disease prognosis. The empirical findings unveiled the prowess of the MSSO-ANFIS predictive model, showcasing an impressive accuracy rate of 99.45%, complemented by a precision score of 96.54%, thereby surpassing the efficacy of alternative methodologies. In their investigation titled "A Hybrid Recommendation System for Heart Disease Diagnosis Utilizing Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System," authors Gunasekaran Manogaran, R. Varatharajan, and M. K. Priyan [135] introduced an innovative approach involving MKL and the ANFIS for the purpose of heart disease diagnosis. This advanced methodology was applied to the KEGG Metabolic Reaction Network dataset, exhibiting a two-fold

strategy. Firstly, the MKL approach effectively allocated parameters between subjects with heart disease and those without, and secondly, the outcome of the MKL process was supplied to the ANFIS classifier for the classification of individuals as either having heart disease or being healthy. The assessment of the proposed approach encompassed metrics such as Sensitivity, Specificity, and Mean Square Error (MSE), ensuring its effectiveness. Furthermore, a comparative analysis was performed against various existing deep learning techniques, including “Least Square (LS)” with SVM, General Discriminant Analysis combined with “Least Square Support Vector Machine (GDA with LS-SVM)”, PCA with ANFIS, and LDA with ANFIS. Notably, the outcomes achieved by the proposed MKL with ANFIS method excelled, demonstrating a remarkable sensitivity of 98%. Narges Pirani and Farzad Faraji Khiavi [139] conducted a comparative study titled “Population Attributable Fraction for Cardiovascular Diseases Risk Factors in Selected Countries: A comparative study,” which aimed to assess the etiological burden of cardiovascular disease in Iran, USA, and Spain. The analysis encompassed studies published between the years 2007 and 2015. The researchers calculated the Population Attributable Risk or Fraction (PAF) for blood pressure, diabetes, and high cholesterol, revealing values of 11.37%, 54%, and 60% for blood pressure, 7.32%, 13%, and 18% for diabetes, and 6.85%, 13%, and 20% for high cholesterol in Iran, USA, and Spain, respectively. The results demonstrated that the risk factor for cardiovascular disease was comparatively lower in Iran compared to the other two countries, shedding light on potential disparities in disease burden and risk profiles. Notably, blood pressure emerged as the most salient risk factor for cardiovascular disease, exemplifying its pivotal role in shaping disease outcomes. However, the comparative analysis also revealed a convergence of risk factors for diabetes and cardiovascular diseases across these nations, such as obesity, age over 45, unhealthy dietary patterns, hypertension, stress, and smoking. These findings underscore the complex interplay of various risk determinants contributing to the prevalence and incidence of cardiovascular disease, necessitating tailored preventive strategies and targeted interventions for population-level risk reduction. The study's implications hold practical importance in reinforcing the significance of regular blood pressure monitoring for individuals with hypertension and the imperative of consistent adherence to prescribed medications to avert untoward consequences.

Furthermore, the comprehensive assessment of risk factors for type II diabetes, including genetic predispositions, obesity, advanced age, race, hypertension, low birth weight, stress, dietary practices, and smoking, underscores the multifactorial nature of diabetes, demanding a holistic approach to disease management and prevention [136]. Nonetheless, numerous prospective studies have consistently demonstrated the significant impact of lifestyle modifications pertaining to factors such as obesity, dietary habits, physical activity, stress management, and smoking cessation in the prevention and management of type II diabetes. These lifestyle interventions are deemed paramount in mitigating the risk and progression of this metabolic disorder. In addition, the results of a separate study conducted in Iran in 2012 to explore gender disparities in coronary heart disease risk identified significant associations between blood pressure, diabetes, and cardiovascular disease, illuminating the need for tailored healthcare approaches for diverse population subgroups [137]. Another study in 2011 sought to quantify the population attributable fraction (PAF) for cardiovascular disease risk factors, with hypertension and high blood pressure, alongside diabetes, receiving priority in preventive strategies for cardiovascular health, catering to both women and men [138]. These findings collectively offer critical insights into the intricate web of cardiovascular disease risk factors and underscore the importance of evidence-based interventions to address the escalating burden of cardiovascular diseases worldwide. J. T. Salonen et al. [140] conducted a study titled “Smoking, blood pressure and serum cholesterol as risk factors of acute myocardial infarction and death among men in Eastern Finland,” which investigated the impact of “smoking”, “serum cholesterol”, and “blood pressure” on the risk of “acute myocardial infarction (AMI)” and death due to all causes and CVD in a random sample of men aged 35 to 59 years hailing from the “North Karelia and Kuopio counties of Eastern Finland”. A total of 4034 men were subjected to the study, achieving a high participation rate of 92%. The study's findings revealed that “smoking”, “elevated blood pressure”, and “serum cholesterol” exhibited independent and combined associations with an elevated risk of myocardial infarction among middle-aged men. Furthermore, “serum cholesterol”, “systolic blood pressure”, and “smoking” were identified as significant factors for predicting the risk of subsequent myocardial infarction and CVD-related mortality, with serum cholesterol emerging as the strongest predictor of cardiovascular disease death. This study serves

as an important contribution to understanding the multifactorial nature of myocardial infarction risk and emphasizes the interplay between these modifiable risk factors, which are of paramount importance for preventive interventions and targeted healthcare strategies aimed at reducing the burden of cardiovascular diseases in the population. The author A. Rezaianzadeh et al. [141] conducted an extensive study to investigate the risk factors associated with cardiovascular disease among a population of 10663 individuals residing in “Kharameh, a city in the South of Iran”, during the years 2015–2022. This research was based on data from the "Kharameh cohort study," which forms part of the larger Prospective "Epidemiological Studies in Iran (PERSIAN)" initiative launched in 2014. The study's inclusion criteria involved selecting individuals between the ages of 40 and 70 years, without any pre-existing mental disorders or cardiovascular disease history, and not having reported any prior stroke incidents. Out of a total of 9442 participants meeting the eligibility criteria, observations were recorded over a 4-year period from 2018 to 2021. The study examined various demographic aspects, behavioural habits, and biological parameters, encompassing attributes “age, sex, marital status, education, place of residence, employment status, Body Mass Index (BMI), waist circumference, hip circumference, alcohol consumption, smoking habits, and the medical history of certain diseases”. Statistical analysis of the data was performed utilizing the log-rank test to assess cardiovascular incidence disparities between men and women. Furthermore, simple and multiple “Cox” regression analyses, incorporating “Firth's bias reduction” method, were employed to identify predictors of CVDs. The statistical computations were conducted using "R version 4.1.2" software, employing the "Coxphf" package, and "STATA version 12." The study's findings revealed that “diabetes”, “hypertension”, “age”, “male gender”, and “alcohol consumption” were significant risk factors associated with CVDs. The incidence of CVDs exhibited an upward trend in older age groups, for both men and women. The risk of CVDs was substantially higher in male compared to female. Notably, with advancing age, the risk of developing CVDs escalated, with individuals in the age groups of 50–60 years and 60–70 years facing 2.4 and 3.7 times higher risk, respectively, compared to individuals aged 40–50 years. Additionally, men were 2.3 times more likely to develop CVDs compared to women. These observations underscore the importance of understanding the complex interplay of various risk

factors in cardiovascular disease, thereby facilitating the implementation of targeted preventive strategies and healthcare interventions to mitigate the burden of CVDs in this population. In the study outlined in "Hydrodynamic Theory of Atherosclerosis Formation in Humans - Reaction to Spasm: Cylindrical Cholesterol Plaque as the Precursor of Heart Attacks and Strokes," [142] SERGEY Rusanov introduced a comprehensive Hydrodynamic Theory that addresses intricate aspects concerning the genesis, maturation, and degradation of 'authentic' plaques in humans. This theory encapsulates the etiology, pathogenesis, clinical presentations, taxonomy, complications, as well as treatment and prevention modalities specifically related to cylindrical plaques, which underlie the occurrence of heart attacks and strokes in humans. Table 2.1 list down the existing study findings.

Table 2.1: Existing studies showing similar and relevant studies

Ref	**Author [Journal]**	**Title**	**Year**	**Method**	**Dataset**	**Findings**	**Accuracy**
[2]	Shadman Nashif et al. "World Journal of Engineering and Technology"	"Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitori	2018	SVM using WEKA tool	Cleveland Heart Disease dataset from UCI repository	Cloud-based machine learning systems can be an effective tool for predicting HD and can be potentially use for real-time monitoring of	97%

		ng System"				patients. However, further research is needed to validate the findings and explore the limitations of the systems.	
[3]	Amin Ul Haq et al. "Mobile Information Systems"	"A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms"	2018	SVM, K-NN, LR, DT, RF, ANN and NB, mRMR, LASSO and Relief validating using K-Fold Validation	Cleveland dataset	Applying feature reduction methods on models can reduce the execution time and improve accuracy.	88%

[5]	Rahul Kumar Jha et. al. "sersc"	"Optimal Machine Learning Classifiers for Prediction of Heart Disease"	2020	SVM, KNN, DT, RF, DNN and NB using Rapid Minor Tool	Cleveland dataset	Comparison between severel classical algorithms and DNN. DNN outperformed.	93%
[7]	W. Mehrdad Aghamohammadi et al. "Computational Science"	"Predicting Heart Attack Through Explainable Artificial Intelligence"	2019	ANFIS, K-fold cross validation, GA	Cleveland dataset	ANFIS using K-fold cross validation and genetic algorithm gives better accuracy.	84.43%
[8]	Negar Ziasabounchi "International Journal of Electrical &	"ANFIS Based Classification Model for Heart Disease Prediction"	2014	ANFIS, GA	Cleveland dataset	GA if combined with ANFIS, can give better results	92.30%

	Computer Sciences"						
[9]	Oluwarotimi Williams Samuel et al. "Expert Systems with Applications"	"An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction"	2017	ANN, Fuzzy_AHP	Cleveland dataset	The hybrid integrated decision support system is capable of providing personalized HF risk predictions and treatment recommendations based on patient-specific data.	91.10%
[10]	A.V. Senthil Kumar [Journal of Artificial	"Diagnosis of Heart Disease using Fuzzy Resoluti	2012	ANFIS, MATLAB	Cleveland dataset	ANFIS performs well for HD prediction and could be better	91.83%

	Intelligence]	on Mechanism"				if included with other algorithms.	
[12]	Zeinab Arabasadi et al. [Computer Methods and Programs in Biomedicine]	"Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm"	2017	ANN, GA	Z-Alizadeh Sani	Genetic algorithm is a good combination and help in boosting performance of neural network for heart disease prediction.	93.85%
[89]	Diman Hassan et al. [Biomedical Signal Processing and Control]	"Heart disease prediction based on pre-trained deep neural networks	2022	DNN, PCA	Cleveland dataset	Utilizing a pre-trained DNN with PCA and LR can outperform in terms of	91%

		combined with principal component analysis"				performance.	
[90]	Mrs. K. Uma Maheswari and Ms. J. Jasmine "INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY"	"Neural Network based Heart Disease Prediction"	2017	ANN, LR	Cleveland dataset	Combining ANN for feature extraction with LR for prediction can give result for HD prediction.	84%
[91]	C.V. Aravinda et al "Deep Learning for Medical Applications with Unique Data"	"A deep learning approach for the prediction of heart attacks based on	2022	NN, RF, DT	Cleveland dataset	NN with Random Forest and Decision Tree can lead in good accuracy result for	90%

		data analysis				HD prediction .	
[13]	Kaan and Ahmet [Procedia Computer Science]	Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks ”	2017	ANN-Fuzzy_AHP, GARFNN	Cleveland dataset	Comparison done between ANN-Fuzzy_AHP and GARFNN that combines GA with Recurrent Fuzzy Neural Network. GARFNN show better results.	91.10%
[14]	G. S. G. Thippa Reddy et al.	“Hybrid genetic algorithm and a fuzzy logic classifier for heart	2020	GA, FL	Cleveland dataset	System uses a combination of fuzzy logic and genetic algorithm	90%

		disease diagnosis”				(AGAFL), rough set feature selection, and fuzzy rule-based classification to make predictions about heart disease.	
[15]	Akgul M et al. [INFUS]	“Diagnosis of Heart Disease Using an Intelligent Method: A Hybrid ANN – GA Approach”	2019	ANN, GANN	Cleveland dataset	comparison study between ANN and hybrid GANN and measured the performance over Cleveland dataset. GANN outperformed.	95.82%

[17]	Sneha Nikam [International Journal of Latest Trends in Engineering and Technology]	"Cardiovascular Disease Prediction using Genetic Algorithm and Neuro-fuzzy System"	2017	GANFIS	Cleveland dataset	Neural Fuzzy System and Generic Algorithm improve efficiency and reduce error rate.	
[30]	Santhanam et al	"Heart Disease Prediction Using Hybrid Genetic Fuzzy Model"	2015	GA, FS	Cleveland dataset	Hybrid Genetic-Fuzzy System gives good result	Satisfactory
[31]	MAbushariah et al. [Journal of Software Engineering and Applications]	"Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and	2014	MLP, ANN, ANFIS, MATLAB	Cleveland dataset	ANN show better result than ANFIS. Improvement requires in ANFIS to be able to	87%

		Adaptive Neuro-Fuzzy Inference Systems (ANFIS) Approaches"				perform well.	
[80]	Azam Davari Dolatabadi et. al [Computer Methods and Programs in Biomedicine]	"Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM"	2017	SVM and PCA		The combination of SVM and PCA were able to read HRV signal in ECG correctly and model performed well.	99%
[81]	Zerina Masetic and Abdulhamit Subasi [Computer Methods and	"Congestive heart failure detection using random forest	2016	KNN, SVM, DT, RF and ANN	Cleveland dataset	Classification models KNN, SVM, DT, RF and ANN were	100%

	Programs in Biomedicine]	classifier"				experiment to check best performing model. RF beat all due to its learning capacity.	
[82]	Moloud Abdar et. al [Computer Methods and Programs in Biomedicine]	"A new machine learning technique for an accurate diagnosis of coronary artery disease"	2019	GA, PSO algorithm	Z-Alizadeh Sani	The combination of GA and PSO algorithm give better results	93.08%
[83]	Shashikant et al.	"Predictive model of cardiac arrest in smokers using machine	2019	LR, DT and RF	Heart dataset, MITU Skillogies Pune, India	Classical models LR, DT and RF were compared to find best	93.61%

		learning techniqu e based on Heart Rate Variabili ty parameter"				optimal solution.	
[29]	Yaowei Li et al. [IEEE]	"Combin ing Convolu tional Neural Network and Distance Distribut ion Matrix for Identific ation of Congesti ve Heart Failure"	2018	CNN	MIT-BIH RR Interval Databas es	FuzzyGM En-generated DDM and Inception _v4 from CNN give good result.	81.85%

[96]	V. K. Sudha and D. Kumar "SN Computer Science"	"Hybrid CNN and LSTM Network For Heart Disease Prediction"	2023	CNN and LSTM	Cleveland	The hybrid CNN and LSTM method give good result	89%
[84]	Sultan Noman Qasem, Monirah Alsaidan [International Journal of Advanced Computer Science and Applications]	"A New Hybrid Intelligent System for Prediction of Medical Diseases"	2018	NN-GSO	Cleveland dataset	NN-GSO algorithm outperform for most of the disease but NN-GA give better results for heart	
[85]	Mohammad Shafenoor Amin et. al [Telematics and	"Identification of significant features and data mining	2019	DT, NB, KNN, SVM, LR, Vote and ANN	Cleveland dataset	Comparison done between classical models to find best model.	

	Informatics]	techniques in predicting heart disease"				Vote performed well.	
[87]	Rajesh Nichenametla et al. [International Journal of Engineering & Technology]	"Prediction of Heart Disease Using Machine Learning Algorithms"	2018	NB, DT	Cleveland dataset	Naibe Bayes demonstrated best result with large set of data and decision tree show better result for small dataset	

2.3 Comparison of Efficiency of ML based Classical Methodologies

Below, *table 2.2* lists the comparison analysis of various machine-learning classical methods used in HD prediction.

Table 2.2: Performance summary of classical ML techniques used in heart disease prediction

Reference	Method Used	Methodology	Dataset	Result
[86]	SVM, 10-Fold cross-validation	Cloud based prediction system using Machine learning technique SVM used. 10-fold CV method used for validation.	Cleveland heart disease dataset	Accuracy: 97.53% Sensitivity: 97.50% Specificity: 94.94%
[80]	SVM, PCA	ECG extracted HRV signals with reduced extracted features applying PCA experimented with SVM.	Long-Term ST database	Accuracy: 99.2%
[12]	Neural network, GA	Experiment with Feed-forward (1 hidden layer) structured hybrid Neural Network-Genetic Algorithm experimented with 54 features from dataset to compute efficiency of model.	Z-Alizadeh Sani	Accuracy: 93.85%

[3]	LR, SVM, ANN, K-NN, NB, DT, Random Forest, mRMR, K-fold cross-validation	Classifiers were experimented using 3 feature selection algorithm "Relief, mRMR, and LASSO" applying K-fold cross validation method.	Cleveland heart disease dataset	SVM with mRMR outperformed Accuracy: 88% ANN (MLP) with Relief outperformed
[13]	Fuzzy_AHP, ANN	Model trained by ANN, weights were computed by Fuzzy_AHP algorithm	Cleveland heart disease dataset	Accuracy: 91.10%
[84]	NN-GA, NN-GSO and NN-PSO	Algorithm were trained, performance was compared to reveal the best model	Cleveland heart disease dataset	NN-GA outperformed for heart disease
[85]	K-NN, DT, NB, Logistic Regression, Vote, SVM and Neural Network	Classifiers algorithms were experimented and results were compared.	Cleveland and UCI Statlog heart disease dataset	Vote gives better accuracy

[29]	CNN and DDM (distance distribution matrix)	Convolutional neural network (CNN) was used and experiment was done to calculate a distance-distribution matrix (DDM) in entropy	MIT-BIH RR Interval Databases	Accuracy: 81.85%
[83]	DT, Logistical Regression, and Random-Forest	DT, Logistical-regression and random-forest model trained to find the best performer among the models.	Data science research group, MITU Pune, India	Accuracy: 93.61% Sensitivity: 92.11% Specificity: 95.03%
[88]	Fuzzy Expert System	Fuzzy Expert System using MATLAB	Cleveland heart disease dataset	Accuracy: 94%
[96]	Hybrid CNN and LSTM	The hybrid CNN and LSTM method was employed to classify the heart disease	Cleveland	Accuracy: 89%
[109]	RF, SVM, KNN, LR, and AdaBoost	Performance was compared with several standard classifiers,	Heart failure dataset accessible	Accuracy: 76.25%

		including RF, SVM, KNN, LR, and AdaBoost	from UCI repository	
[115]	RF, SVM, KNN NB and XGBoost, ANN	Methods were experimented to compute the probability. XGBoost classifier, when paired with the wrapper technique gives best output.		Accuracy: 73.74%
[102]	ANN	The experiment was conducted using MATLAB, incorporating various ECG features	Physio Net ECG database	
[100]	NB, DT, RF, KNN	Experiment conducted using different ML method to predict heart disease	Cleveland	Accuracy: 90.8%
[107]	LR, NB, DT, RF, KNN, SVM	Experiment conducted using different ML method to predict heart disease	Cleveland	Accuracy: 82.92%

[101]	ANN, RF	Aims to enhance the accuracy of HD prediction. Included “Hybrid Random Forest” with a “Linear Model (HRFLM)” integrated ANN	Cleveland	Accuracy: 88.7%
[127]	RF, DT, Hybrid model	RF, DT and a hybrid model comprising RF and DT experimented. Hybrid model give best result	Cleveland	Accuracy: 88%
[105]	RF, DT, SVM, LR	Model trained to predict likelihood of stroke. RF outperformed		Accuracy: 98.94%
[103]	RF, SVM, LR, NB, Gradient Boosting	Model were trained using R software for prediction of heart disease. LR outperformed	Cleveland, Hungarian, Switzerland, and Long Beach VA datasets	Accuracy: 86.5%
[104]	RF, MLP, KNN, ET, XGB, SVC, SGD, ADB,	Experiment done using different	Cleveland	Accuracy: 92%

	CART, and GBM	method to predict heart disease		

2.4 Comparison of methodologies comprises of Neural Network and Genetic Algorithm and others

Below, table 2.3 shows the comparison of approaches using neural network and genetic algorithm with other techniques.

Table 2.3: Comparison of methodologies comprises of Neural Network and Genetic Algorithm and others

Author name	**Proposed approach**	**Technique**	**Dataset**	**Result**
Akgul, M [15]	Comparison of ANN with GANN	ANN, GANN	Cleveland	Accuracy (ANN): 85.02% Accuracy (GANN): 95.82%
Mrs. K. Uma Maheswari, Ms. J. Jasmine [90]	Combination of LR and ANN	ANN, LR	Cleveland	Accuracy (ANN): 84%
Diman Hassan et al. [89]	Experimented using pre-trained DNN with PCA and LR	DNN, PCA, LR	Cleveland	Accuracy: 91%
MAbushariah, M.A.M. et al. [31]	Comparison study with Multilayer Perceptron	ANN, ANFIS	Cleveland	Accuracy: 87%

	structure on the ANN, and ANFIS.			
C.V. Aravinda et al. [91]	Deep learning algorithms applied with NN, RF and DT classifiers to analyse patients' data for HD prediction	NN, RF, DT	Cleveland	Accuracy: 90%
Ramakrishnan, K et al. [11]	GA based RFNN with 13 input, 7 hidden neurons evaluated for HD prediction	GA, RFNN	Cleveland	Accuracy: 96.63%
Alizadehsani, R et al. [12]	Experiment done on hybrid GA-NN Algorithm with 54 features to find the prediction of heart disease	GA, NN	Z-Alizadeh Sani	Accuracy: 93.85%
Sultan Noman Qasem, Monirah Alsaidan [84]	Experiment carried out using NN with PSO, GSO and GA	NN-GA, NN-GSO and NN-PSO	Cleveland	NN-GA outperformed

Animesh Kumar Paul et al. [130]	Experimented using GA and a modified dynamic "multi-swarm particle swarm optimization (MDMS-PSO)" for the prediction of heart disease	GA, MDMS-PSO	Cleveland, Hungarian, 71 Switzerland, Long Beach, and Heart dataset from UCI repository	92.31% using Cleveland dataset, 95.56% for the Hungarian, 89.47% with Switzerland dataset, 91.80% for the Long Beach dataset, and 92.68% using Heart dataset

2.5 Comparison of Efficiency of Fuzzy Inference based Methodologies

Below, table 2.4 demonstrate the comparison of fuzzy system and neural networks. Different fuzzy based approaches were compared to record the performance.

Table 2.4: Comparison of fuzzy system based existing models

Author(s)	**Proposed approach**	**Technique used**	**Accuracy (%)**
Bhuvaneswari Amma N G [92]	Fuzzy-based clinical detection	PCA, ANFIS	93.2%
Oluwarotimi Williams Samuel et. al [9]	Hybrid system which combines Fuzzy	ANN, Fuzzy AHP	91.1%

	Analytic Hierarchy Process, ANN		
Mehrdad Aghamohammadi et. al [7]	GA, ANFIS	ANFIS	84.43%
Yaowei Li, Yao Zhang, Chengyu Liu, Li Zhang, Lina Zhao, Liuxin Zhang, Yang Zhang [29]	FuzzyGMEn, Inception_v4	CNN	81.85%
Negar Ziasabounchi, Iman Askerzade [8]	Adaptive neural fuzzy inference system based model	ANFIS	92.30%
Mohammad A. M. Abushariah et. al [31]	ANN and ANFIS	ANN, ANFIS	87.04%
Hameed A.Z. and Ramasamy B. et al. [124]	Weighted fuzzy rule approach in conjunction with GA	Fuzzy System, GA	68%, 48%, 55% with Cleveland, Hungarian, and Switzerland respectively
Purushottam Sharma and Kanak Saxena [125]	Comparison study between classic method and fuzzy system and GA	DT, NB, KNN, FL, GA	88.11%

Nagarajan R. and Thirunavukarasu R. [126]	Neural fuzzy based system trained to predict the disease	NFIS	87.7%
Gunasekaran Manogaran, R. Varatharajan, and M. K. Priyan [135]	Combination of multiple kernel learning and ANFIS for the purpose of prediction of heart disease	MKL, ANFIS	Sensitivity: 98%
M. Faris Al Hakim et al. [131]	Fuzzy inference system using the Tsukamoto method was utilized for prediction of risk of heart attack	FIS, Tsukamoto method	58%
P.K. Anooj [119]	Mamdani fuzzy inference system used for predicting heart disease	FIS, Mamdani	62.35%, 46.93%, and 51.21% with Cleveland, Hungarian, and Switzerland datasets respectively
Gabriella Casalino et al. [122]	A fuzzy rule-based framework designed to augment expert decision-making in the realm of cardiovascular diseases	FIS	91%

Osman Taylan et al. [132]	MATLAB used with Sugeno-type fuzzy rule-based ANFIS model, integrating Gaussian membership functions used to predict heart disease	ANFIS	96.5%

2.6 Benefits of ML based Intelligent Hybrid Systems

Based on above analysis, result clearly explains that hybrid system can produce better efficiency with comparison of traditional ones. Below is the gist of analysis done so far.

- GANN algorithm gives better result for datasets having small number of samples so is better for heart disease with small dataset like Cleveland.
- Hybrid systems gives better result than traditional classic models.
- Classifiers in combination of feature selection methods gives better result than alone.
- Reduced size of features improves performance in terms of accuracy. It can also help in reduction of execution time of the classification models.
- Techniques like ANN and SVM are widely used in health disease prediction due to their flexibility and requirement of smaller training datasets.
- Generation of more accurate results will be possible by further enhancement in existing hybrid systems converting it into intelligent hybrid systems.

2.7 Research Gap

Despite having multiple ML framework available for the prediction of the disease wherein, works had focus on some key points which was worked upon but it left with some gaps too. To propose this work, thorough study has been done and after analysis few gaps has been identified that has been taken as roadmap for proposed work discussed in Chapter 3. Deep analysis has been done after reviewing the existing studies and below research gaps has been identified:

- *Identification of suitable prediction model:* Many models have been proposed in past but there is a still need of identification of suitable predictive model for heart attack diagnosis.
- *Investigate the effectiveness of ML algorithms:* There is a need for studies that investigate the effectiveness of neural network-based ML algorithms for heart disease prediction compared to traditional statistical models.
- *Appropriate feature selection:* Appropriate feature selection improve classification accuracy. Irrelevant feature can affect the ML classifiers performance.
- *Handling data from diverse population:* It is very important that trained model should be generalized enough to cover groups such as based on age or gender.
- *Account for individualized risk factors:* There is lack of study discussing the individualized risk factors like family history of patient or any habit like addiction of cigarette or alcohol. There is scope to analyze patient record keeping these parameters in scope.
- *Improvement in performance of machine learning predictive models:* There is a still need of improvement in existing predictive model for heart attack.
- *Selecting appropriate classification model:* There is a lack of knowledge base for selecting appropriate classification model for finding heart attack probability.
- *Model execution time:* Though it has been found that model execution time is improving one after another experiment but there is still a scope of improvement.

- *Cover temporal changes in patient health:* There is scope to cover temporal changes in health parameter of patient and calculate the probability based on the range of the feature changes.
- *Notification to patient after predicting issue:* There is no IoT based mechanism for sending notification and prescription to the person regarding their some of the specified health parameter.
- *Solution to provide appropriate preventive advices to patient:* There is a scope to build an intelligent solution, which can provide appropriate preventive advices to patient.
- *Exposure of model for external uses:* Once model is trained and ready for test, it is important that it should get available for external uses like device integration or any third-party integration for medical use.

2.8 Summary

In this chapter several studies have been reviewed which are related to the chosen subject. During literature review we found many machine-learning technology like SVM, KNN, LR, DT, RF, NB, ANN, CNN, DNN or hybrid system are chosen for doing experiment and finding the optimal solution. We analysed results of many experiments having classical methods or neural networks or hybrid systems and found that with application of feature selection or adding any other data pre-processing techniques helps in improving the model result. Reduced number of features can also help in improving model efficiency as well as it saves cost as well. Hybrid system has better optimization capacity than the classical one and it can result better as compared to classic methods. Research gaps has been identified based on the literature reviews which will be used in define research objectives discussed in chapter 3.

Chapter 3: Research Methodology

3.1 Introduction

Each scientific investigation inherently encompasses a methodological framework, analogous to the designed blueprint. This methodology serves as the guiding structure being used in completing the research work, steering it towards the intended objectives. Through this meticulous journey, our aim is to provide not just an investigation, but a comprehensive map of the scientific voyage undertaken to unearth insights into the vital domain of heart disease prediction and machine learning. In the preceding sections of this thesis, we embarked through the foundational underpinnings of heart disease and its intersection with the intricate realm of machine learning, meticulously detailed in Chapter 1. Following this foundation, Chapter 2 dug into a comprehensive review of relevant literature, illuminating the landscape of prior research in this domain. In this chapter, methodology for all the objectives has been discussed in depth. Within its confines, we get on a detailed exploration of the methodologies meticulously crafted to propel our research forward. Herein, we meticulously outline the research objectives, unveil our proposed conceptual framework, and delineate the precise tools and software instrumental to our work. Furthermore, in this chapter, we venture into the depths of methodological details designed to address each of our research objectives. We also unveiled the methodological apparatus deployed in data collection, setting the stage in the subsequent chapters where we shall engage in experimental evaluations.

3.2 Research Objectives

This is the key stage in any research work life cycle and there must be some objective defined before starting a research work. Defining research objectives is a critical cornerstone in the research process for several compelling reasons. Firstly, it provides a clear direction, acting as a guiding beacon, offering a clear trajectory throughout the research. This ensures that the study remains focused, maintains its course, steering

clear of tangential or inconsequential pursuits and does not deviate into unrelated or inconsequential path. A well-defined objective establishes the criteria for evaluating the success or failure of the research, offering a measurable yardstick for progress, they also facilitate in resource allocation, allowing researchers to allocate time, manpower, and financial resources towards achieving the defined specified goals. Furthermore, clear objectives promote effective communication and collaboration within research teams, as everyone is aligned towards a common purpose, hence, fosters a more efficient and cohesive research environment. Ultimately, well-structured research objectives lay the foundation for producing meaningful and impactful findings, contributing substantively to the body of knowledge in a particular field. Prior to formulating these objectives, an extensive review of literature was conducted, drawing upon the insights of esteemed researchers whose work closely aligned with the proposed study. *Table 3.1* lists all the research objectives that have been framed based on the analysis done during the literature reviews.

Table 3.1: List of research objectives

ID	Description
RO-01	To study various ML approaches used in heart disease diagnosis
RO-02	To build a hybrid framework for the prediction of heart disease
RO-03	To develop a hybrid framework to find the probability of heart attack
RO-04	To test and evaluate the proposed machine learning based hybrid system algorithm

Using the aforementioned goals as a guide, a comprehensive research framework has been developed to comprehend the proposed system's workflow, elucidated in the next sections. Illustrated in Figure 3.1 below, this schematic representation elucidates the sequential progression towards achieving the proposed objectives. The initiation point lies in the formulation of the research objective RO-01 that predominantly centered on the extensive exploration of literature reviews on the subject area; the workflow

embarks on a diligent exploration of knowledge reservoirs. Popular databases have been searched with subject area keywords to filter research papers which has been further narrowed down to filter most relevant papers, thus constituting the foundation for the subsequent literature review. Subsequently, the research advances into the experimental domain, after the first objective is completed, next stage is to carry out the experiments that collectively encapsulates the RO-02 and RO-03 combinedly. This multifaceted undertaking entails the acquisition of the Cleveland dataset from the venerable UCI repository. The acquired dataset is subjected to a rigorous regimen of data pre-processing, meticulously honing it to encompass only the salient features essential for the research objectives; model has been trained and output is recorded. At this stage, an innovative hybrid neural network incorporating a “Genetic Algorithm Neural Network (GANN)” has been developed. GANN combines the traditional backpropagation algorithm with a genetic algorithm to optimize the weights of the neural network, able to find the optimal set of weights for NN, leading to improved accuracy and performance. The GANN model underwent multiple epochs, with a definite batch size, in order to optimize the network's weights. During this process, mutation was applied to generate new offspring and increase the diversity of the population, model trained to get best solution based on population generation using genetic algorithm. Various ML methods have been experimented for comparison with the proposed GANN method. Different criterion has been experimented like classical methods with different feature selection methods and cross validation method. Later proposed Performance from classical methods have been compared with GANN to see if proposed method is more capable than the classical methods and other methods proposed in the state of the art. In essence, the intricate architecture of this research framework not only underscores the meticulous delineation of research objectives but also showcases the ingenious interplay of advanced techniques such as the “Genetic Algorithm-Neural Network (GANN)” and “Fuzzy Inference Systems”. The subsequent stage involves the prognostication of the likelihood of a heart attack, necessitating the development of an intricate neural fuzzy inference system. This system emulates human-like decision-making paradigms, scrutinizing fluctuations in medical parameter readings to arrive at pertinent clinical determinations. This pivotal system operates at the crux of the diagnosis process, orchestrating a nuanced evaluation of potential

cardiac events. In essence, the proposed research framework intricately integrates cutting-edge methodologies, specifically the GANN and fuzzy inference system, to usher in a paradigm shift in the diagnostic landscape of cardiovascular health. Fuzzy classifiers, despite their unique characteristics, may not outperform other well-established approaches such as statistical methods, decision trees, or neural networks when solving classification problems, however, their distinctive strengths lie in their linguistic interpretability, intuitive handling, and overall simplicity, which play crucial roles in enhancing the acceptance and practicality of a solution. The ability of fuzzy classifiers to provide human-readable interpretations of results and facilitate user-friendly interactions can be particularly beneficial in real-world applications, where transparent and understandable models are desirable for decision-making processes. In addition to the mentioned advantages, fuzzy classifiers are capable of handling uncertain and imprecise data, making them suitable for scenarios where traditional crisp models might struggle due to data variability or noise. Furthermore, the incorporation of fuzzy logic allows the integration of expert knowledge or domain-specific rules, contributing to the development of more personalized and context-aware classification systems. However, it is essential to consider that the performance of fuzzy classifiers may heavily depend on the nature of the dataset, the complexity of the classification problem, and the tuning of fuzzy rules and membership functions. In certain situations, hybrid approaches that combine fuzzy classifiers with other machine learning techniques might prove to be the optimal solution, taking advantage of both the interpretability and robustness of fuzzy logic and the predictive power of other methods. Overall, while fuzzy classifiers may not always yield state-of-the-art accuracy, their unique benefits and adaptability to diverse scenarios make them valuable tools in the arsenal of classification algorithms, especially when human interpretability and ease of use are essential requirements for successful adoption in various domains. In this phase, the model undergoes the process of fuzzification and defuzzification, thereby imbuing it with the capacity to reason and draw inferences akin to human cognitive processes. This is achieved by incorporating a set of membership functions and fuzzy rules that collectively constitute the knowledge repository of the inference system. The system has the ability to make decisions that are similar to human decision-making and can predict the probability of heart attack based on medical parameters provided to the

model. To fortify precision, a meticulous crafting and rigorous testing of these membership functions and fuzzy rules is undertaken, ensuring the model's proficiency in rendering precise prognoses. Subsequently, the research endeavour advances to address research objective RO-04, pivoting towards the seamless integration of the model with an Application Programming Interface (API). This necessitates the exportation of the duly trained model into a cohesive bundle, which is then seamlessly amalgamated with the API infrastructure. This integration culminates in the exposure of endpoints, thus rendering the model accessible for public utilization. To expedite this process, the Python FASTAPI library emerges as the linchpin, enabling a streamlined and efficient integration process. Through this orchestrated integration, the model's capabilities are democratized, opening avenues for widespread access and utilization.

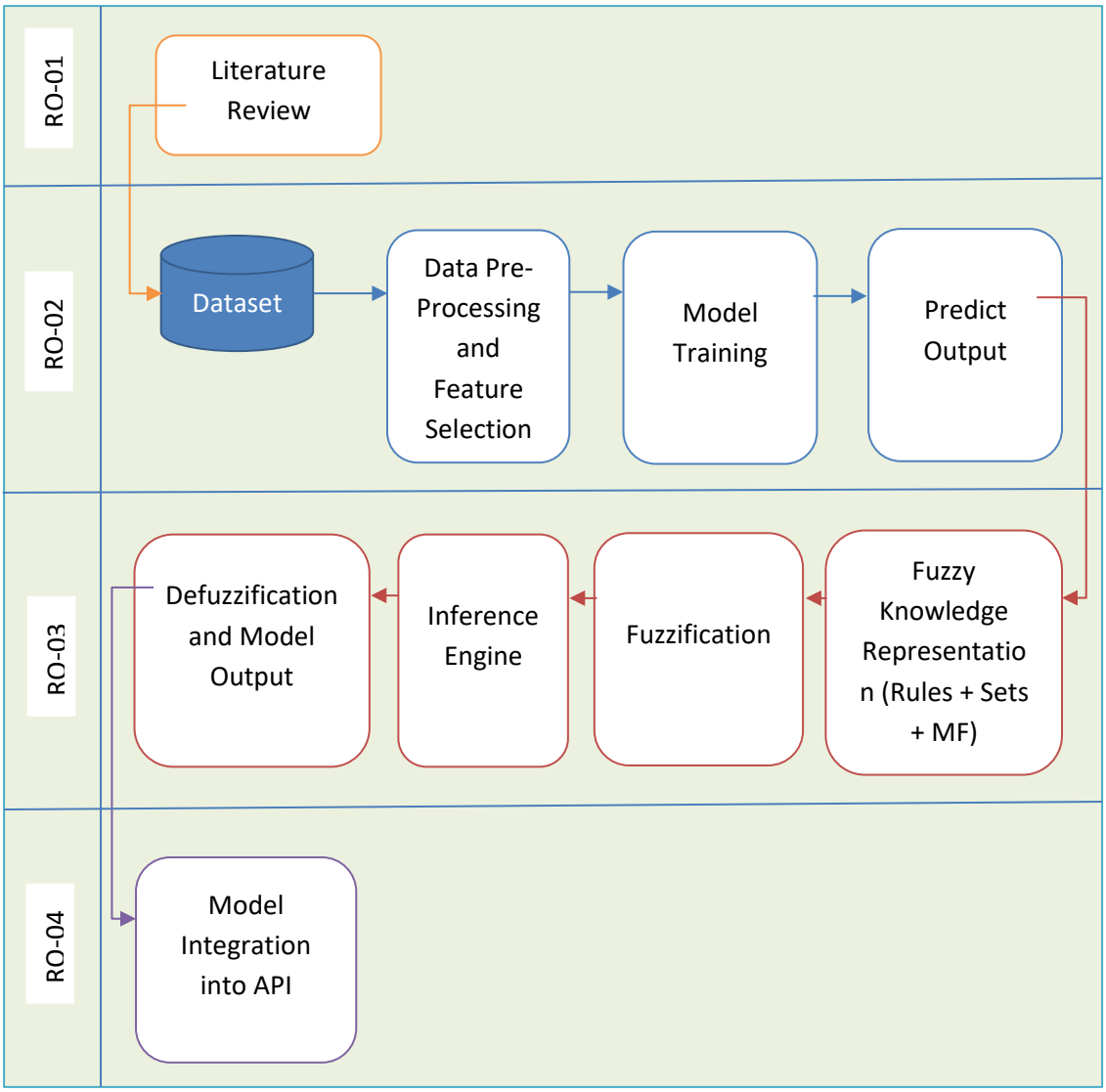


Figure 3.1: Proposed work flow depicting stages in completion of research objectives

Below subsections elaborate the path to complete all the objectives one by one starting with RO-1, 2, 3 and 4.

3.2.1 RO-01 - To study various ML approaches used in heart disease diagnosis

In pursuit of the stated objective, a comprehensive and systematic review of pertinent academic literature has been undertaken. This process commenced with the meticulous querying of prominent scholarly databases such as SOPUS and ScienceDirect, employing precisely calibrated search terms to yield a corpus of papers germane to the research endeavour. Further refinement was achieved through the application of

tailored filters encompassing subject area, publication year, proposed frameworks, and other pertinent parameters. This judicious filtration process culminated in the compilation of a discerningly curated collection of papers, poised for deeper inspection and reference in subsequent phases of the study. For an exhaustive exploration of this process, a comprehensive discussion is offered in Chapter 2. The perceptive perusal of this extensive body of literature yielded valuable insights. It emerged that while a plethora of research efforts have been directed towards the prediction of heart disease utilizing an array of machine learning methodologies, a noticeable scarcity exists in the specific domain of prognosticating the likelihood of a heart attack. This discernment illuminates a relevant research gap within the domain, underscoring the novelty and significance of the current study. The literature review, thus served as a crucible for refining the research focus and positioning it within the broader landscape of existing scholarship. Moreover, the comparative analysis undertaken during the literature review phase further enriched the understanding of the extant body of work. This analytical exercise, illustrated upon in detail in Chapter 2, stimulated a deep understanding of the relative strengths and limitations of diverse methodologies and approaches adopted in prior research endeavours. It not only facilitated the identification of methodological best practices but also offered valuable insights into areas warranting innovation and refinement. This comprehensive survey of the existing academic corpus thus provided the researchers with a sturdy vantage point, anchoring their subsequent investigations within the evolving boundary of knowledge.

3.2.2 RO-02 - To build a hybrid framework for the prediction of heart disease

To accomplish the overarching objective of enhancing the prediction accuracy of heart disease, we have undertaken a multifaceted approach in our research. This approach encompasses a thorough and systematic literature review, the development and utilization of a hybrid neural network known as GANN (Genetic Algorithm Neural Network), and the rigorous experimentation and analysis of results. The first step in our research journey involved an exhaustive exploration of relevant academic databases, including SOPUS and ScienceDirect. By deploying pertinent search terms and

employing filters such as subject area, publication year, and proposed frameworks, we meticulously curated a collection of research papers that were most germane to our investigation. This process of rigorous curation aimed to ensure that we obtained a refined list of scholarly works that could serve as invaluable references in our quest to enhance heart disease prediction (a comprehensive exposition of our literature review can be found in Chapter 2). As we delved into the existing body of knowledge, we discerned a distinctive pattern in the literature landscape. While numerous solutions had been proposed for the prediction of heart disease utilizing an array of machine learning methodologies, there was a notable dearth of methodologies specifically tailored to predict the probability of a heart attack. This insight highlighted a critical research gap that we were determined to bridge. To address this research void and advance the state-of-the-art in heart disease prediction, we embarked on the development of a hybrid neural network, GANN. The rationale behind employing a hybrid neural network, specifically GANN, stems from its unique amalgamation of neural network architecture with the evolutionary optimization strategies of genetic algorithms. This synergy engenders a dynamic computational paradigm capable of discerning intricate patterns and relationships within the dataset, ultimately leading to more accurate prognoses. The neural network component endows the model with the capacity for sophisticated learning and pattern recognition, while the genetic algorithm empowers it to iteratively refine its predictive capacity over successive generations. This symbiotic interplay between neural networks and genetic algorithms forms the crux of the GANN model's potency. The GANN model is a sophisticated fusion of neural networks and genetic algorithms. This amalgamation harnesses the pattern recognition capabilities of neural networks and couples them with the optimization prowess of genetic algorithms. To facilitate the training and evaluation of the GANN model, we leveraged the Cleveland dataset, which is widely recognized and utilized in the realm of heart disease research. Our overarching aim with the GANN model is to craft a predictive tool that not only improves the accuracy of heart disease prognosis but also expedites the diagnostic process. Timely and precise diagnoses are of paramount importance in medical practice, as they can significantly influence treatment strategies and ultimately patient outcomes. In the subsequent sections of this chapter and in Chapter 5, we have embarked on an expedition through the experiments

conducted with the GANN model. These experiments are designed to rigorously evaluate the model's performance, assess its predictive capabilities, and validate its potential to significantly enhance heart disease prediction. Through these empirical investigations, we aspire to not only contribute to the academic discourse surrounding heart disease prediction but also to offer practical insights that can have a tangible impact on clinical practice and patient care.

3.2.3 RO-03 - To develop a hybrid framework to find the probability of heart attack

The central goal of this specific objective is to construct a model capable of forecasting the likelihood of a heart attack. Such a predictive tool holds immense value, as it can serve as an early warning system for individuals, providing crucial insights into their current health status. Armed with this knowledge, individuals can take proactive measures to mitigate risks and potentially avert a heart-related crisis. To realize this objective, a fuzzy inference system has been meticulously crafted. This system, underpinned by the principles of fuzzy logic, functions as a virtual health advisor. It assesses an individual's health parameters and offers a comprehensible evaluation of their current health condition. By providing individuals with timely and accessible information, the fuzzy inference system empowers them to make informed decisions regarding their health and well-being. By providing early indicators, individuals can be better prepared for any potential health challenges that may arise. In essence, the system acts as a sentinel, diligently monitoring key health indicators and alerting individuals to any signs of potential cardiac distress. This proactive approach can be transformative, as it not only enhances individual health outcomes but also reduces the burden on healthcare systems by promoting preventative healthcare measures. The subsequent sections of this chapter and Chapter 5 delve into the intricate workings of this innovative fuzzy inference system. Detailed discussions encompass the system's design, its ability to process and interpret health data, and its capacity to provide actionable insights. Through a comprehensive exploration of its functionality and performance, the reader gains a profound understanding of how this system augments the landscape of cardiac health prediction. In sum, this objective represents a pivotal

juncture in the research journey, as it embodies the translation of cutting-edge technology into a practical and potentially life-saving application. As the subsequent chapters unfold, a deeper appreciation of the fuzzy inference system's role in revolutionizing cardiac health prediction will emerge, reaffirming its significance in the realm of healthcare innovation. This proactive approach to healthcare not only fosters a culture of informed decision-making but also holds significant potential for improving health outcomes and overall well-being. Through these advancements, this study seeks to make meaningful contributions to the field of cardiovascular health and enhance the quality of care provided to individuals at risk of heart disease.

3.2.4 RO-04 - To test and evaluate the proposed machine learning based hybrid system algorithm

This specific objective is primarily focused on the comprehensive testing of the established framework. In order to facilitate this, a crucial step involves the exposure of an Application Programming Interface (API). This interface acts as a bridge, allowing the trained model to interact with external systems. This integration is pivotal as it enables the system to be seamlessly incorporated into broader healthcare ecosystems. Moreover, the API functionality extends beyond mere integration. It lays the groundwork for future enhancements, envisaging features such as automated notifications about a patient's health status. These notifications could provide invaluable guidance, outlining specific precautions and recommended actions to stabilize the situation. This proactive approach to healthcare empowers individuals with actionable insights, fostering a culture of informed decision-making and self-care. In the subsequent Chapter 6, this objective will be thoroughly explored, providing an in-depth examination of the testing procedures. This will encompass an evaluation of the API's performance, ensuring that it effectively communicates with the trained model and delivers accurate outcomes. Furthermore, the integration process with external systems will be detailed, shedding light on how the framework seamlessly fits into the broader healthcare landscape. Additionally, the potential for further feature development will be discussed, emphasizing the system's adaptability and scalability. This includes considerations of how the API can serve as a foundation for future

enhancements, potentially revolutionizing the way healthcare information is disseminated and utilized. By rigorously addressing this objective, the research endeavours to establish a robust and reliable framework that not only meets current needs but also lays the foundation for future advancements in healthcare technology. Through effective testing and integration, this framework is poised to make significant strides in improving the accessibility and quality of healthcare services.

3.3 Research Framework

The central objective of this proposed research framework lies in the development of a solution geared towards predicting the likeliness of a heart attack based on specific medical parameters. The underlying aim is to furnish timely alerts to patients and offer them personalized preventive measures. This not only empowers individuals to take precautionary steps but also ensures timely medical intervention. In order to accomplish this, the framework leverages the intelligence of Artificial Intelligence (AI) in conjunction with Machine Learning (ML) algorithms. By implementing this research framework, the study endeavours to bridge the gap between medical data analysis and actionable insights. It not only showcases the potential of AI and ML in healthcare but also underscores their capacity to revolutionize patient care by providing personalized, proactive healthcare solutions. This powerful combination forms the cornerstone of the model, is subsequently structured into two distinct stages. Below stages represent crucial phases in the predictive process.

1 Framework to build a system for the prediction of heart disease.

2 Framework to find the probability of heart attack.

Proposed framework in *figure 3.2*, presents the high-level flow of model, this is further staged as stage 1 and stage 2:

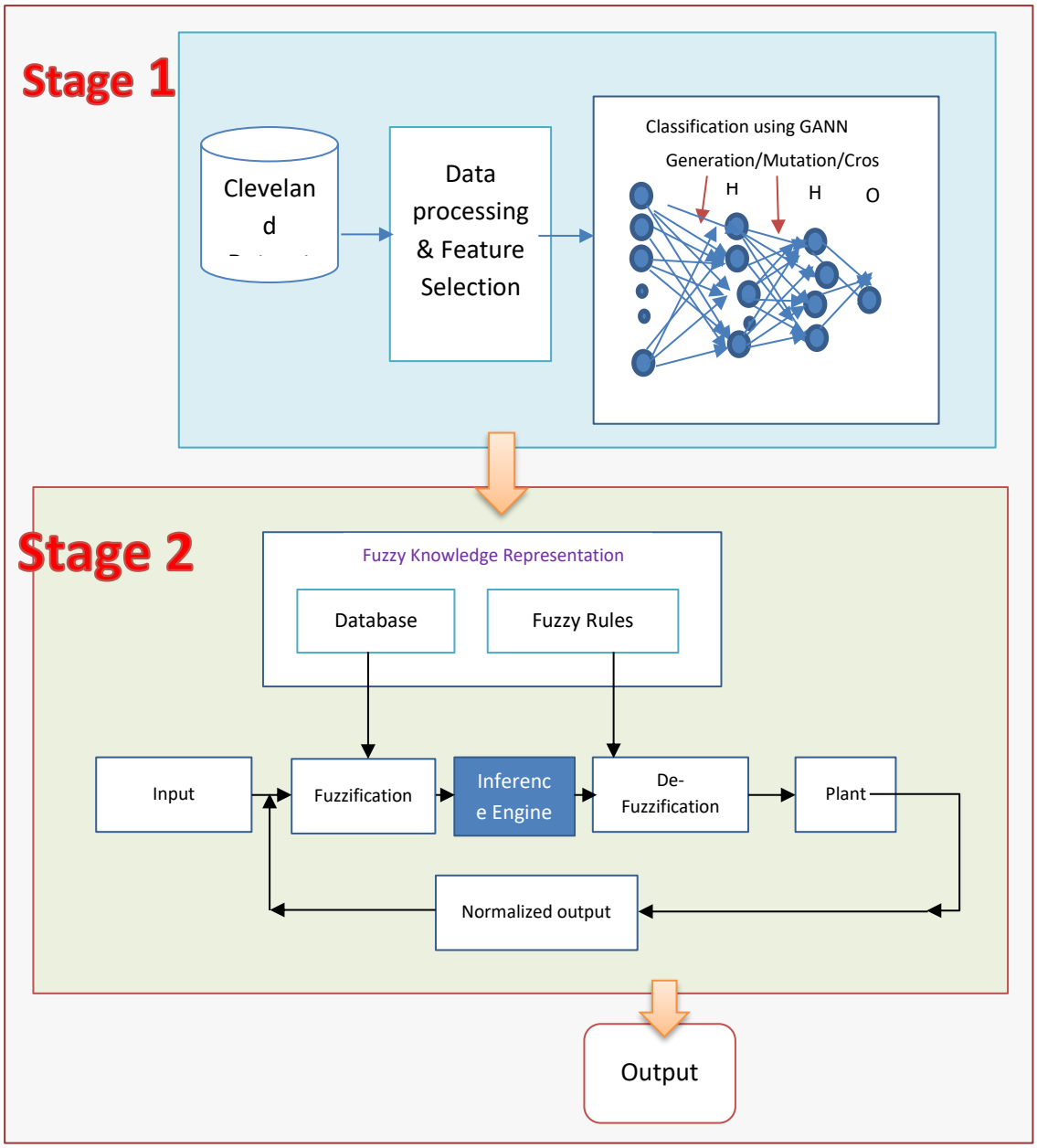


Figure 3.2: Sequential implementation of GANN based Fuzzy Inference System (GANFIS)

The schematic in Figure 3.2 exhales the structural framework of the proposed research work. This framework bifurcates into two apparent stages, each with a specific focus. The inaugural stage primarily revolves around the anticipation of heart disease, methodically engineered to harness the potential of machine learning and artificial intelligence. Subsequently, the second stage precisely describes the methodology for gauging the probability of a heart attack occurrence. This research employs an innovative hybrid system, seamlessly amalgamating the prowess of neural networks, genetic algorithms (GA), and a fuzzy inference system. This integration synergistically

empowers the model to accomplish the predefined objectives with a heightened level of accuracy and efficacy. The intricacies of this integrated system will be elucidated comprehensively in the ensuing subsections. The initial stage, dedicated to heart disease prediction, leverages the comprehensive Cleveland dataset from the UCI repository. This extensive dataset, encompassing a spectrum of physiological attributes, is subjected to a rigorous process of data preprocessing, effectively streamlining it for the subsequent stages of the model. The core of this stage lies in the development and training of a hybrid neural network, which uniquely incorporates the genetic algorithm (GA). This symbiotic amalgamation harnesses the strengths of both approaches, fostering a dynamic and adaptive learning process. Transitioning to the second stage, the focus pivots towards the assessment of heart attack probability. At this juncture, a pivotal component is the implementation of a fuzzy inference system. This sophisticated system emulates human-like decision-making, adeptly discerning patterns in the medical parameters provided. By intelligently processing these inputs through a series of membership functions and fuzzy rules, the system produces a nuanced evaluation of heart attack probability. The integration of these two stages epitomizes a novel approach towards comprehensive cardiac health assessment. The ensuing sections will delve deeper into the intricate mechanisms underpinning each stage, providing a thorough comprehension of the methodology's inner workings.

Stage 1 – Predicting heart disease: In the initial phase of the proposed research framework, the focus is directed towards the prediction of heart disease in patients, employing a meticulously structured multi-step process. This phase commences with a crucial data pre-processing step, wherein the input data undergoes a comprehensive cleansing and transformation process. This is a pivotal stage, as it lays the foundation for subsequent analyses. Following this, the mRMR algorithm, a feature selection technique, is deployed. This step entails the identification and extraction of the most pertinent features from the dataset. These features serve as the cornerstone for subsequent modelling. The subsequent step introduces the construction and training of a neural network, a formidable tool in machine learning. Here, a key innovation is integrated: the utilization of a Genetic Algorithm (GA). This optimization technique, inspired by the principles of natural selection, provides a sophisticated means to generate and refine the weights that govern the neural network's architecture. The

Genetic Algorithm Neural Network (GANN) solution thus endeavours to achieve an optimal neural network configuration, encompassing considerations such as the number of layers and nodes in each layer. The training phase of the neural network is paramount, as it strives to discern intricate patterns and relationships within the pre-processed data. This iterative process is further refined by the introduction of genetic operators, namely crossover and mutation techniques, which engender new offspring. These techniques are instrumental in enhancing the diversity and adaptability of the neural network population. As the training progresses, the fittest neural network, exemplifying superior predictive capabilities, is selected to form the nucleus of the subsequent generation. This cyclical process ensues until an optimal solution is attained, characterized by a neural network with a commendable proficiency in predicting heart disease. The conclusion of this phase concludes in the utilization of the best-performing neural network to predict the likelihood of heart disease occurrence. The output of the neural network serves as a foundational metric in the calculation of the probability of heart disease, predicated on the ensemble of medical parameters provided. This calculated probability serves as a pivotal indicator, furnishing valuable insights into the potential occurrence of heart disease in the individual under consideration. This first stage embodies a meticulous orchestration of advanced computational techniques and machine learning principles. It reflects a nuanced and systematic approach towards leveraging the potential of artificial intelligence in the domain of cardiac health assessment. The subsequent sections will delve deeper into the intricacies of each component, providing a comprehensive understanding of the methodology's inner workings.

Stage 2 – Finding heart attack probability: In the second phase of this research framework, the focus turns towards the critical task of calculating the likelihood of a heart attack. This is a pivotal stage as it involves the integration of a Fuzzy Inference System (FIS) into the existing system, enhancing its predictive capabilities. The input data, processed and refined in the preceding stages, serves as the foundation for this crucial computation. Within the Fuzzy Inference System, a comprehensive set of pre-generated fuzzy rules, numbering over 13,000, has been crafted to encompass a wide spectrum of scenarios, spanning from normal and risky cases situations. These rules serve as the cornerstone for making informed assessments based on the provided

medical parameters. For every medical parameter furnished to the model, a set of membership function is created. These functions play a pivotal role in characterizing the degree of association between a given parameter and the corresponding fuzzy set. This process, known as fuzzification, is a cornerstone of fuzzy logic systems, enabling them to process and interpret real-world data in a nuanced manner. Following the establishment of membership functions, the Mamdani and Sugeno systems are integrated into the computational framework. These systems represent distinct approaches to fuzzy logic-based decision-making, each with its unique strengths and applications. The choice between them is informed by the specific requirements of the computational task at hand. The input data, undergone the fuzzification process, is now transformed into fuzzy values, which encapsulate the relationship between the medical parameters and their corresponding fuzzy sets. This transformation process is fundamental in enabling the system to effectively process and analyse the input data, paving the way for accurate and insightful predictions. The output of the fuzzification process, now represented in fuzzy terms, undergoes a critical refinement process known as defuzzification. This vital step serves to sanitize the fuzzy output into a crisp, actionable result. Through a systematic evaluation of the fuzzy relationships and rules, a precise assessment emerges, providing a clear indication of the probability of a heart attack. This second stage embodies a sophisticated integration of fuzzy logic principles, leveraging their inherent capacity to handle uncertainty and imprecision in real-world data. The resulting system represents a powerful tool in assessing the likelihood of a heart attack, augmenting the diagnostic capabilities of the overall framework. This phase, characterized by its strategic application of fuzzy logic principles, forms an essential component of the research framework. Its successful execution further underscores the comprehensive and meticulous approach adopted in this research endeavour. The succeeding sections will delve deeper into the specific methodologies and techniques employed within this stage, offering a comprehensive understanding of its inner workings.

Below pseudo code explains the workflow for the proposed system.

Algorithm 3.1: Pseudo code for Hybrid neuro fuzzy inference system

---

Step 1: Input dataset for experiment

Step 2: Apply data pre-processing, design the model.

Step 3: Apply mRMR feature selection to generate feature indices.

Step 4: Load preliminary population with random uniform chromosomes.

Step 5: Apply GA and compute weight-matrix for population.

Step 6: Train the model

Step 7: Repeat the steps 4 to 6, record the scores.

Step 8: Perform model testing to predict HD.

Step 9: Generate inference rules using the priority and complexity of medical parameters from the dataset.

Step 10: Create MFs for each medical parameter based on the defined fuzzy set.

Step 11: Pass the input through the fuzzification process.

Step 12: Process the fuzzy output to defuzzification process to get crisp output.

Step 13: Test the model to measure efficiency.

Proposed framework has been discussed in details in next sub sections.

3.3.1 Framework to build a system for prediction of heart disease

To design the proposed framework, hybrid neural network has been used. The schematic diagram below, *figure 3.3* provides a simplified graphical illustration of the core components and stages that constitute the proposed system. The architecture is thoughtfully partitioned into five distinct stages, each playing a decisive part in the seamless operation of the framework: input dataset, data pre-processing, feature selection, model training, and decision-making. The expedition begins with the input dataset, which forms the foundation of the entire analytical process. This repository of information encompasses a comprehensive array of parameters and metrics, sourced from credible databases and repositories. These parameters encapsulate a diverse range of physiological and clinical attributes, each holding potential insights into the realm of

heart health. As the data progresses through the pipeline, it encounters the pivotal phase of data pre-processing. This phase is akin to a refining process, where the raw data undergoes a series of transformations to ensure it is in its most conducive form for subsequent analysis. Here, tasks such as normalization and standardization come into play, harmonizing the diverse units and scales of the input data. Outliers and erroneous entries are addressed, ensuring that the dataset is purged of any anomalies that may skew the ensuing analysis. Following this preparatory phase, the data advances to the feature selection stage. This is a crucial juncture, where the dataset is distilled to its most salient components. Advanced algorithms are deployed to discern the attributes that bear the highest predictive potential. By prioritizing these attributes, the framework optimizes computational resources, directing them towards the variables that wield the greatest influence in the analytical process. With the roster of selected features in hand, the focus shifts towards the training of the model. At the heart of this process lies the hybrid neural network, a sophisticated computational model that synergizes the power of NN with the optimization capabilities of GA. This union furnishes the model with an inherent adaptability, enabling it to autonomously refine its internal structure in response to the data it encounters. The outcome is a dynamic system, adept at extracting nuanced patterns and relationships from the input data. Within the crucible of the hybrid neural network, the selected features are subjected to a rigorous analytical process. Through a series of weightings and computations, the network distils the input data into a coherent prediction. This prediction, embodying the likelihood of heart disease, emerges as the culmination of a multi-layered analytical process, where each layer refines and refocuses the information gleaned from the data. The final output of this analytical journey stands as a testament to the potential of the proposed system. It constitutes a probabilistic assessment of the prediction of heart disease for the given set of input parameters. This output, a synthesis of advanced computational techniques and medical expertise, holds profound implications for augmenting diagnostic processes, enabling timelier interventions, and ultimately, safeguarding lives.

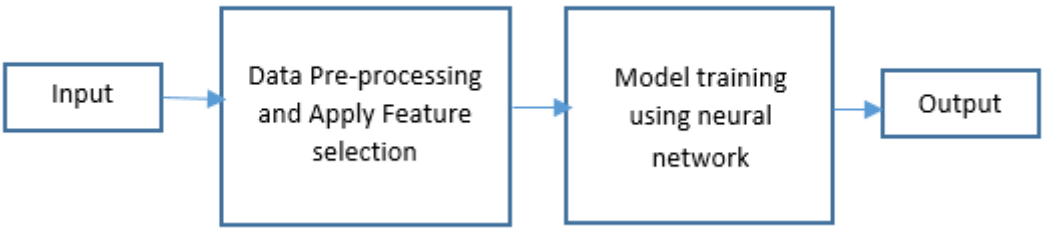


Figure 3.3: Block demonstrating building a system for HD prediction

The proposed framework stands poised to be a transformative tool in the realm of healthcare, specifically in the domain of heart disease prediction. At its core lies a meticulous process that commences with the acquisition of vital input from the Cleveland heart disease dataset, as expounded upon in the preceding Chapter 4. This dataset comprises a comprehensive set of 76 distinct parameters, each of which holds crucial medical significance. To unlock the full potential of this dataset, a systematic journey ensues. It embarks with a crucial phase known as data pre-processing. Here, the raw data is subjected to a series of preparatory steps aimed at refining its quality and ensuring that it is suitably formatted for subsequent analysis. These operations range from tasks as fundamental as data cleansing, where erroneous or extraneous entries are rectified or removed, to more intricate operations such as data normalization, which ensures that the data falls within standardized ranges, facilitating meaningful comparisons. Following this preparatory phase, the process advances to the domain of feature selection. This is a pivotal step, as it involves the identification of the most pertinent attributes from the dataset that will be instrumental in driving the classification process. Leveraging advanced algorithms, this stage acts as a discerning filter, sieving through the multitude of parameters to unearth those that exhibit the highest predictive potential. By doing so, it streamlines the subsequent stages of analysis, ensuring that computational resources are focused where they can yield the greatest impact. With the roster of relevant features now assembled, the framework shifts its focus towards the classification process. This pivotal step is entrusted to a hybrid neural network, a sophisticated model that marries the computational prowess of NN with the optimization capabilities of genetic algorithms. This union empowers the model to autonomously adapt and refine its internal structure in response to the data it encounters. The outcome is a highly adaptive system, capable of extracting nuanced

patterns and relationships from the input data. In the crucible of the hybrid neural network, the selected features are meticulously analysed and processed. The network employs a series of weightings and computations to transform the input data into a coherent prediction. This prediction, representing the likelihood of heart disease, emerges as the culmination of a multi-layered analytical process, where each layer refines and refocuses the information gathered from the data. The final output generated by the hybrid neural network serves as the definitive outcome of the proposed system. It encapsulates a probabilistic assessment of the likelihood of heart disease for the given set of input parameters. This output, poised at the node of advanced computational techniques and medical expertise, holds immense potential for enhancing diagnostic processes, enabling timelier interventions, and ultimately, saving lives. Below flow chart illustrates the detailed design for the prediction of HD using GANN.

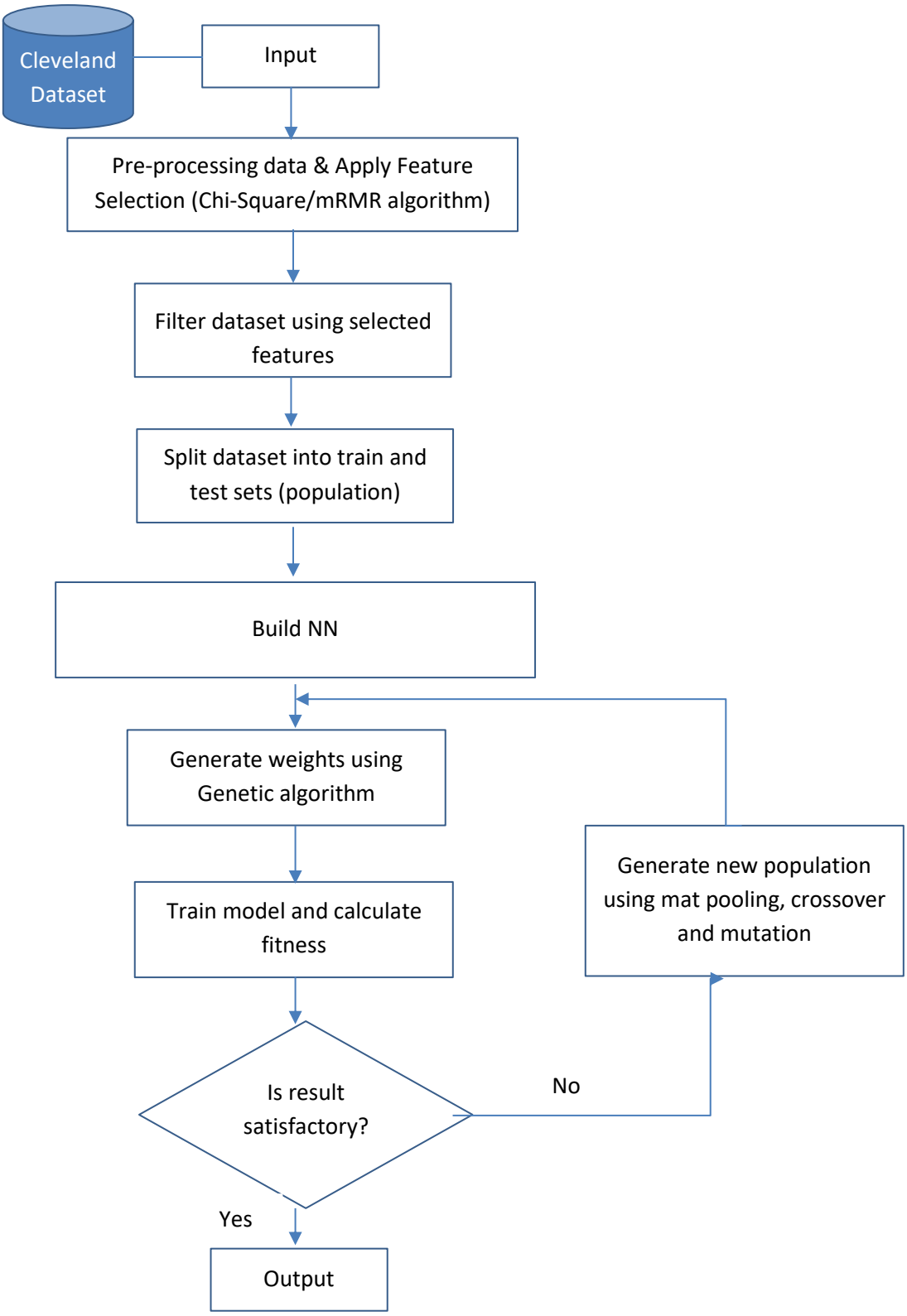


Figure 3.4: Flow chart illustrating detailed design for HD prediction

The process of predicting heart disease using Genetic Algorithm Neural Network (GANN) over Cleveland dataset is visualized in the steps outlined in *figure 3.4*. The

Genetic Algorithm works by evolving a population of potential neural networks through a process of selection, mutation, and crossover. The best performing neural network in the population is chosen for the subsequent generation, and the procedure is iterated until a satisfactory solution is obtained. Here are the steps:

- *Data collection:* Collect the "Cleveland" dataset, which contains 303 instances with 76 features [24].
- *Data pre-processing:* Pre-process the data and apply feature-selection to select the most relevant features for the task. This is done using the algorithms such as "chi-square" and "mRMR".
- *Train-Test data splitting:* After the pertinent features are chosen, the dataset is split into training and testing sets, following an 80:20 ratio. The training set is employed to train the GANN model, while the test set is reserved for evaluating the model's performance.
- *Build GANN Model:* GANN is built in stages starting with
    - *Initializing the population:* In this stage, the initial pool of population of potential neural networks is created. The pool size could be large or small and each member of the population is a potential NN architecture characterized by unique configurations of layers, nodes, and activation functions, for example, a potential neural network architecture consists of 2 hidden layers each containing eight nodes, and a sigmoid activation function.
    - *Evaluate the Population:* Once the initial population has been generated, the next stage is to train the neural networks on the dataset. In this stage, each neural network in the population is trained using backpropagation and a genetic algorithm to fine-tune the weights and biases. In the case of our specific dataset, it's divided into training and testing subsets, maintaining an 80:20 ratio. The training subset is employed to instruct each neural network within the population, whereas the testing subset is reserved for assessing the performance of each network. The fitness

function used to evaluate each neural network would be based on the accuracy of its predictions on the testing set.

- *Selection and Crossover:* the neural networks in the population are evolved using a combination of "selection", "crossover", and "mutation" operations. The selection operation selects the best-performing neural networks in the population to reproduce, while the crossover and mutation operations introduce new variations in the neural network architectures. This process is repeated until a satisfactory solution is found.
- *Mutation:* In the next stage, a small proportion of the newly produced solutions are randomly mutated to introduce new variations in the population.
- *Condition verification:* Once the genetic algorithm has converged and the best neural network architecture has been identified, the final stage is to test and validate the model on new, unseen data. The performance of the final model is evaluated based on its accuracy

o *Evaluate the Model:* In later part, the model's performance is evaluated on the testing data using appropriate metrics.

3.3.2 Framework to find the probability of heart attack

In the succeeding phase of the framework, the spotlight is directed towards determining the likelihood of a heart attack. This critical task is entrusted to a specially designed fuzzy inference system, a computational engine adept at processing imprecise data and making nuanced decisions based on a set of predefined rules. The handoff from the neural system to the FIS marks a pivotal transition in the analytical process. The input data, enriched with insights gleaned from the neural network, is now poised to navigate the working of the fuzzy logic domain. This domain thrives on the premise that not all information is categorical or absolute; rather, there exists a spectrum of uncertainty and ambiguity within medical parameters. The fuzzy inference system serves as a bridge between this nuanced reality and the realm of precise decision-making. At the heart of

this system are a curated set of fuzzy rules, meticulously crafted to encapsulate the myriad possible relationships between input parameters and the likelihood of a heart attack. These rules serve as the compass, guiding the system through the complex terrain of medical data. Each rule encapsulates a specific scenario, defined by a combination of input parameters and their respective memberships within defined fuzzy sets. This leads to a pivotal juncture in the process: the application of fuzzification. At the crux of the fuzzification process lies the fundamental tenet of linguistic variable representation, wherein crisp numerical inputs are seamlessly translated into linguistically interpretable fuzzy values. This paradigmatic shift from crisp to fuzzy representations endows our computational framework with the requisite flexibility and adaptability to navigate the intricate nuances and uncertainties inherent in real-world data. By encapsulating the inherent vagueness and imprecision pervading the realm of human cognition, fuzzification lays the groundwork for the subsequent inferential processes to unfold, thereby facilitating the seamless integration of human-like decision-making capabilities into our computational framework. Moreover, the fuzzification process engenders a granular delineation of the input space, enabling the delineation of subtle nuances and variations that elude conventional numerical representations. Through the adept manipulation of linguistic terms and membership functions, each input value is stratified across the spectrum of fuzzy sets, thereby furnishing a nuanced portrayal of its degree of association with each constituent set. This multifaceted characterization of input values not only enhances the interpretability and comprehensibility of our computational model but also augments its capacity to discern subtle patterns and correlations latent within the input data. Here, the crisp, concrete data from the neural system is converted into its fuzzy counterpart, a nuanced representation that accounts for the inherent uncertainty in medical parameters. This transformation is facilitated by the carefully calibrated MFs, which assign degrees of membership to each parameter within the defined fuzzy sets. This step is analogous to viewing medical data through a lens of gradations, acknowledging the inherent variability in real-world clinical metrics. The journey continues with the process of rule evaluation, where each rule's contribution to the decision-making process is assessed. This evaluation takes into account both the strength of the rule, determined by the degree of fulfilment of its antecedent, and the shape of the MF, reflecting the degree of

overlap between the input parameter and the fuzzy set. With these evaluations in hand, the system navigates towards defuzzification, the process by which the nuanced, fuzzy outputs are transformed back into a concrete, actionable decision. This stage calls upon well-defined methods, such as centroid-based defuzzification, to compute a precise output value that reflects the system's assessment of the probability of a heart attack. This meticulously orchestrated process culminates in the generation of a decisive output: the calculated likeliness of heart attack based on the amalgamation of medical parameters. This output, embodying the system's integration of fuzzy logic and medical expertise, holds profound implications for clinical decision-making. It represents a convergence of advanced computational techniques and domain-specific knowledge, poised to augment diagnostic precision, facilitate timely interventions, and ultimately, enhance patient outcomes.

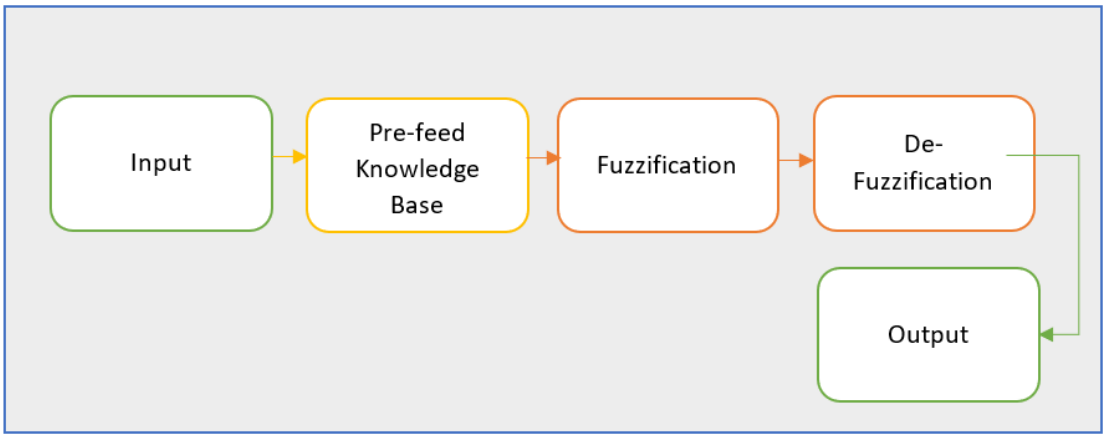


Figure 3.5: Block diagram illustrating alert system using neural and fuzzy inference system

Block diagram in *figure 3.5* illustrates the high-level plantation of proposed system. Framework is divided into four stages – passing input, pre-feed knowledge base, fuzzification and defuzzification.

Stage one (Input): In first stage, input is passed to the model for further processing. These inputs belong to the Cleveland dataset having various medical parameters that can create symptoms for heart attack.

Stage two (Knowledge base): At this stage, system is filled with huge collected knowledge base like rule base which create a base for fuzzy inference system and enable the system to pretend as having human like decision making capabilities. In the

research work, fuzzy rules will be generated based on the probability combination of priority medical parameters, their thresholds and their criticalities listed in table 3.2, 3.3 and 3.4.

Stage three (Fuzzification): In this stage, input is passed to the inference system, fuzzy sets are generated, crisp input $x \in U$ is mapped into fuzzy set $A \in U$ which then along with fuzzy rules are passed to membership functions via various available fuzzifiers like triangular, trapezoidal, gaussian or any other for further processing to convert crisp input into desired values. The main working of these fuzzifiers is to map input x into fuzzy set A with membership functions $\mu_A(x)$ [37]. Furthermore, the symbiotic fusion of fuzzification with the underlying principles of fuzzy logic affords our computational framework with the requisite robustness and resilience to navigate the inherent uncertainties and ambiguities pervading the real-world domain. By embracing the inherent vagaries and indeterminacies intrinsic to human cognition, fuzzification transcends the limitations of conventional deterministic approaches, thereby furnishing the framework with the requisite flexibility and adaptability to navigate the complex terrain of real-world data.

Stage four (Defuzzification): This is the last stage in the proposed system which is just the opposite of fuzzification, here crisp output *y is* produced from the aggregated output of fuzzy set. This transformative metamorphosis from fuzzy to crisp output is quintessentially orchestrated through a repertoire of defuzzification methods, each imbued with its unique set of principles and methodologies aimed at distilling the inherent fuzziness pervading the aggregated output into a tangible and interpretable form. At the vanguard of defuzzification lies a diverse array of methodologies, each meticulously tailored to suit the idiosyncratic requirements and constraints of the underlying computational framework. One such paradigmatic exemplar is the "maximum defuzzification" method, which operates on the principle of selecting the maximum membership value from the aggregated output of fuzzy sets, thereby affording primacy to the most salient and influential fuzzy set in delineating the crisp output. By prioritizing the fuzzy set with the highest degree of membership, the "maximum defuzzification" method epitomizes a pragmatic approach to distilling the aggregated output into a definitive and decisive crisp output, thereby streamlining the

decision-making process and enhancing interpretability. In addition to the maximum defuzzification method, an equally formidable contender in the pantheon of defuzzification methodologies is the centroid defuzzification method. Rooted in the principles of geometric centrality, this method seeks to compute the "centroid or center of mass" of the aggregated output fuzzy set, thereby furnishing a representative and emblematic crisp output that encapsulates the collective influence of all constituent fuzzy sets. Furthermore, the "means of maxima" defuzzification method represents another seminal contribution to the pantheon of defuzzification methodologies, operating on the principle of computing the arithmetic mean of the maximum membership values across all constituent fuzzy sets. This method endeavours to strike a delicate balance between the salient features of each constituent fuzzy set, thereby furnishing a holistic and comprehensive crisp output that encapsulates the collective influence of all constituent fuzzy sets.

Detailed framework has been illustrated in below figure:

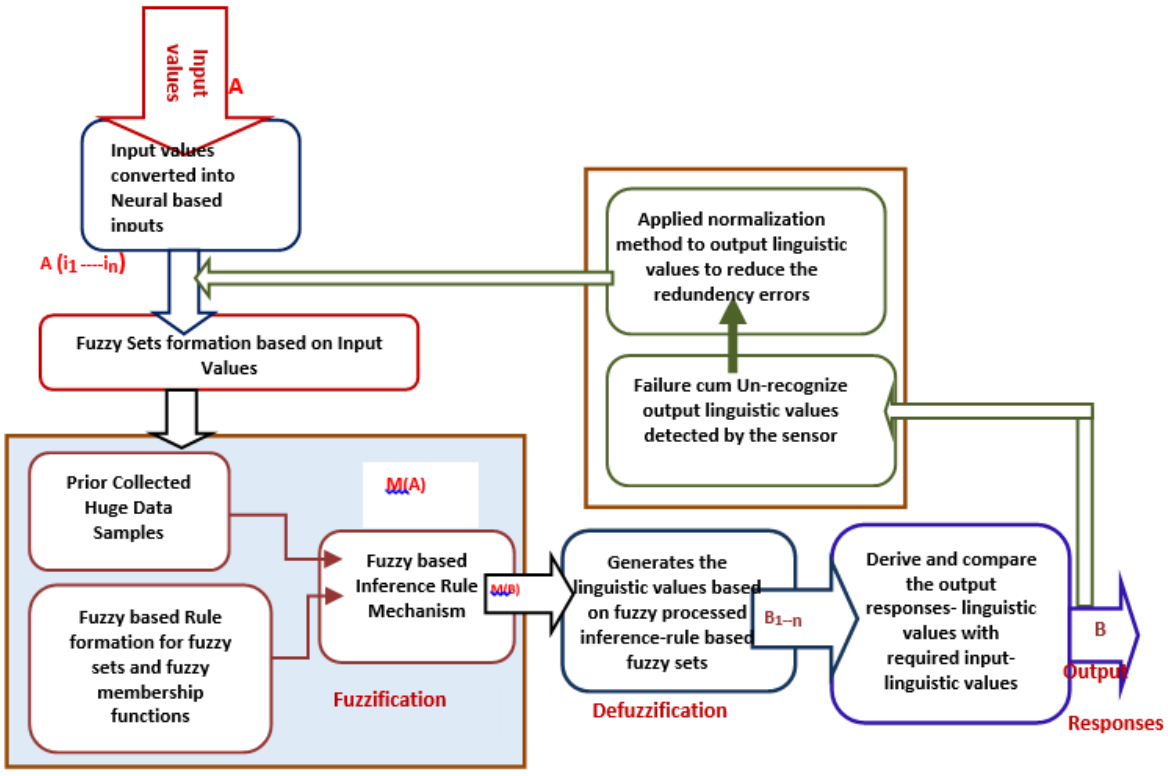


Figure 3.6: Diagram illustrating design for finding the probability of heart attack [24-25]

Figure 3.6 outlines a methodically composed process aimed at gauging the likelihood of a heart attack through the implementation of a NFIS. This intricate framework is underpinned by a stepwise progression, wherein each stage is imbued with a specific function and purpose, collectively contributing to the refinement and elucidation of the final output. By dissecting this process into discrete components, we can unravel the intricate tapestry of computational intelligence underpinning heart attack prediction.

The foundational stage of this paradigm centres on the pivotal process of data normalization, indicating a strategic endeavour poised to streamline subsequent analyses and engender robust insights into the underlying dataset. Positioned at the forefront of the analytical pipeline, data normalization emerges as an indispensable prerequisite, serving as the cornerstone upon which the veracity and interpretability of resulting analyses hinge. Data normalization represents a transformative process geared towards standardizing input values, thereby fostering uniformity and homogeneity across the multifarious dimensions of the dataset. This strategic makeover is underpinned by a singular objective i.e. to mitigate the harmful impact of assortment and disparate scales that often saturate raw datasets, thereby furnishing a conducive situation for the unconstrained application of subsequent analytical techniques. At its core, data normalization is predicated on the principled alignment of data points along a standardized scale, thereby engendering a harmonious convergence of disparate variables onto a uniform metric. This harmonization is achieved through a judicious calibration of input values, wherein each data point is systematically transformed to adhere to a predefined range or distribution. By effectuating this standardization, normalization engenders a sense of parity and coherence across the dataset, effectively levelling the proverbial playing field and obviating potential confounds arising from variegated units or scales. The strategic utility of data normalization transcends the confines of immediate analytical endeavours, permeating into the very fabric of computational frameworks and machine learning algorithms. In the context of neural network architectures, for instance, the efficacy of model training and predictive accuracy hinges critically upon the standardization of input values achieved through data normalization. By aligning input features along a uniform scale, normalization

augments the convergence properties of optimization algorithms, thereby expediting model convergence and enhancing predictive performance.

Following normalization, the input data traverses into the hallowed realms of the system's knowledge base, an expansive repository housing a compendium of pre-defined fuzzy rules. These rules, meticulously curated by domain experts and informed by empirical evidence, serve as the bedrock upon which informed decision-making is predicated. By drawing upon this rich reservoir of expert knowledge, the system gains invaluable insights into the intricate interplay of variables influencing the likelihood of a heart attack, thereby fostering nuanced and contextually relevant analyses.

Within the hallowed confines of the knowledge base, the establishment of fuzzy sets emerges as a linchpin in the computational framework, facilitating the seamless transition from raw input data to actionable insights. The inception of fuzzy sets assumes a pivotal role within the computational framework, serving as a fulcrum for the seamless translation of raw input data into actionable insights. Through judicious calibration, every parameter undergoes a meticulous mapping onto its corresponding membership function, thereby adeptly encapsulating the intricate nuances and subtleties that underlie the dataset. This process of fuzzy set establishment lays a robust foundation for subsequent stages of analysis, endowing the system with the acumen to discern and scrutinize input data through a multifaceted and nuanced lens. Such a nuanced approach facilitates a comprehensive exploration of the dataset, enabling the system to unravel latent patterns and discern subtle relationships that may elude traditional analytical techniques. Thus, the establishment of fuzzy sets couriers a transformative journey towards expounding the underlying complexities of the dataset, furnishing the analytical framework with the requisite tools to navigate the intricate terrain of data interpretation and analysis.

The subsequent stage of fuzzification represents a quintessential hallmark of fuzzy logic systems, wherein the input data undergoes a transformative metamorphosis into fuzzy values. These fuzzy values encapsulate the degree of association with the established membership functions, thereby imbuing the data with a nuanced and contextually relevant semantic framework. By virtue of this transformation, the system is empowered to navigate the complexities of heart attack prediction with heightened

acuity and precision, leveraging the rich tapestry of fuzzy relationships to glean actionable insights.

The denouement of this computational odyssey culminates in the critical phase of defuzzification, wherein the rich tapestry of fuzzy relationships and rules is distilled into a precise and actionable assessment of heart attack probability. Through a systematic evaluation of the aggregated fuzzy output, a definitive result emerges, offering invaluable insights into the likelihood of a heart attack and guiding subsequent clinical interventions. This transformative process of defuzzification represents the apotheosis of computational intelligence, seamlessly bridging the chasm between abstract fuzzy relationships and tangible clinical outcomes. In summation, Figure 3.6 represents the intricate interplay of computational intelligence and clinical acumen in the realm of heart attack prediction. By delineating a stepwise process encompassing data normalization, knowledge base interrogation, fuzzy set establishment, fuzzification, and defuzzification, this computational framework offers a nuanced and multifaceted approach to heart attack prediction, empowering clinicians with actionable insights and facilitating informed clinical decision-making.

3.3.3 Medical parameters and threshold definition

Various medical parameters exhibit differential impacts on disease pathology and can serve as pivotal determinants for disease prediction. Each medical parameter harbours its unique set or range of threshold values, dictating disease susceptibility and progression. The severity of a heart attack is intricately intertwined with diverse medical parameters, with potential impacts spanning from mild to critical. For instance, perturbations in blood pressure dynamics can profoundly impact cardiac function, precipitating dire consequences such as heart failure in vulnerable patients. Likewise, aberrations in blood cholesterol levels can exert a profound influence on the incidence and progression of heart failure. These parameters act as crucial determinants in the predictive models utilized to assess disease risk and progression. Each medical parameter possesses its own distinct impact on disease manifestation, acting as a key contributor to the predictive accuracy of diagnostic models. For instance, deviations from the optimal range in parameters such as blood pressure, blood cholesterol levels,

blood sugar levels, heart rate, cigarette consumption, and family history of heart disease can significantly influence the onset and severity of heart-related conditions. A comprehensive examination of existing literature in Chapter 2, Section 2.2 elucidates the pivotal role of medical parameters and their associated threshold values in delineating disease severity [136-142]. This empirical evidence underscores the intricate interplay between various physiological metrics and disease outcomes, thereby highlighting the imperative of discerning and leveraging these parameters in predictive modelling frameworks. Medical parameters encompass a broad spectrum of physiological indices, ranging from biochemical markers to hemodynamic variables, each bearing unique prognostic significance in the context of cardiovascular diseases and beyond. Notably, the intricate interactions between these parameters necessitate a nuanced understanding of their individual and collective impacts on disease progression and severity. Among the myriad medical parameters examined, blood pressure emerges as a crucial determinant of cardiovascular health, with deviations from normal ranges posing significant risks for adverse cardiovascular events. Fluctuations in blood pressure dynamics can precipitate deleterious effects on cardiac function, exacerbating the risk of myocardial infarction and heart failure in susceptible individuals. Research has shown that sustained hypertension, characterized by elevated blood pressure levels, is strongly associated with an increased risk of heart failure, myocardial infarction, and stroke. Furthermore, hypertension is often considered a silent killer, as it may remain asymptomatic for extended periods, underscoring the importance of routine blood pressure monitoring in clinical practice. Similarly, blood cholesterol levels play a crucial role in cardiovascular health. Elevated levels of low-density lipoprotein (LDL) cholesterol, often referred to as "bad" cholesterol, can lead to the accumulation of plaque within arterial walls, a process known as atherosclerosis. This buildup of plaque can impede blood flow to the heart, increasing the risk of coronary artery disease and heart attack. Conversely, high levels of high-density lipoprotein (HDL) cholesterol, or "good" cholesterol, are associated with a reduced risk of cardiovascular events, as HDL cholesterol aids in the removal of excess cholesterol from the bloodstream. Bllod sugar level is also another significant parameter in cardiovascular risk assessment, particularly in individuals with diabetes mellitus. Chronic hyperglycaemia, characteristic of uncontrolled diabetes, can contribute to endothelial dysfunction,

oxidative stress, and inflammation, all of which are implicated in the pathogenesis of cardiovascular disease. Individuals with diabetes are at an elevated risk of developing coronary artery disease, peripheral arterial disease, and diabetic cardiomyopathy, underscoring the importance of glycaemic control in the prevention of cardiovascular complications. Heart rate, the number of times the heart beats per minute, is another essential parameter in cardiovascular assessment. Resting heart rate serves as a marker of cardiovascular fitness, with lower resting heart rates generally associated with better cardiovascular health. Conversely, persistently elevated resting heart rates may indicate underlying cardiovascular dysfunction, such as autonomic nervous system imbalance or impaired cardiac function. Research has shown that elevated resting heart rates are independently associated with an increased risk of cardiovascular mortality, making heart rate monitoring a valuable tool in risk stratification. In the same way, cigarette consumption, a modifiable risk factor, exerts a profound influence on cardiovascular health. Smoking tobacco exposes individuals to a myriad of harmful chemicals, including nicotine and carbon monoxide, which can damage blood vessels, promote inflammation, and accelerate atherosclerosis. Smoking is a well-established risk factor for coronary artery disease, peripheral vascular disease, and stroke, with smokers exhibiting a significantly higher risk of cardiovascular events compared to non-smokers. Moreover, exposure to frequent smoke can also adversely affect cardiovascular health, highlighting the importance of tobacco control measures in cardiovascular disease prevention. Additionally, family history of heart disease serves as a valuable predictor of cardiovascular risk, reflecting both genetic predisposition and shared environmental factors. Individuals with a family history of premature coronary artery disease are at an increased risk of developing cardiovascular complications themselves, underscoring the importance of early risk assessment and preventive interventions in high-risk populations. Family-based screening strategies can help identify individuals at elevated risk, enabling targeted interventions aimed at mitigating cardiovascular risk factors and optimizing long-term health outcomes. In summary, an in-depth understanding of various medical parameters and their respective thresholds is essential for accurate disease prediction and risk assessment in cardiovascular medicine. By incorporating these parameters into predictive models and risk stratification algorithms, clinicians can optimize patient care and tailor interventions to

individual patient needs, ultimately improving cardiovascular outcomes and reducing disease burden. Table 3.2, provided below, furnishes a comprehensive list of the medical parameters along with their respective thresholds. These parameters, as outlined in Chapter 1, Subsection 1.2.2, represent crucial indicators that may significantly impact heart health and potentially lead to failure. They are sourced from the Cleveland dataset, a well-established repository of medical data, and serve as the foundational input for the proposed system. This detailed flow design not only underscores the systematic approach adopted in this research framework but also highlights the critical role played by each stage in determining the probability of a heart attack.

Table 3.2: Comprehensive Guide to Medical Parameters: Descriptions and Thresholds for Probability Calculation

S. No.	Feature Name	Description	Threshold
1	RestingBP_RBP	BP measured when admitted at the hospital	Systolic/diastolic (in mm Hg) Normal: < 120/80 Abnormal: 120-140/80-120 Critical: 140/120 > [33, 34]
2	SerumCholesterol_SCH	In mg/dl	Total Cholesterol: Normal: < 200, Moderate: 200 – 239, High: 239> HDL level: Normal: 40 – 60 Low: < 40, High: 60 > LDL level:

			Normal: <130, Moderate: 130 – 189, Critical: 189 >
3	FastingBloodSugar_FBS	Blood sugar level at fasting	0 for Normal,1 for High
4	RestingECGResult_RES	ECG result	0 for Normal,1 and 2 are other types
5	RestingHeartRate_RHR	Heart rate at resting	0-65 – Normal, 65-85 – Average, Above 85 high
6	MaxHeartRate_MHR	Max heart rate	Till 100 – Normal, 100-140 – Moderate, 100-180 – High, Above 180 – Critical
7	CigratePerDay_CPD	Cigrate intake per day	Higher is the intake, more is the failure risk
8	HeartDiseaseFamilyHistory	Heart disease history in family	0 for No,1 for Yes
9	IsHeartPatient	Is patient a heart patient	0 for No,1 for Yes

A comprehensive grasp of diverse medical parameters and their associated thresholds listed in above table holds paramount importance in precision disease prediction and risk evaluation within cardiovascular medicine. Through the integration of these parameters into predictive models and risk stratification algorithms, clinicians can refine patient care strategies and customize interventions to suit individual patient requirements, thereby enhancing cardiovascular prognosis and alleviating disease burden. This nuanced approach enables clinicians to harness the intricate interplay

between physiological metrics and disease pathology, fostering a holistic understanding of cardiovascular health and enabling proactive management strategies aimed at optimizing patient outcomes. Below, table 3.3 serves as a cornerstone within our research framework, delineating critical threshold values for essential medical parameters extracted from our comprehensive dataset. These thresholds are not merely numerical boundaries; rather, they represent pivotal benchmarks that underpin the foundation of our subsequent fuzzy rule generation process. This process, pivotal to endowing our system with human-like decision-making process, necessitates a profound comprehension of the intricate landscape of medical diagnostics. The paramount importance of these threshold values lies in their role as discriminative markers, delineating between normal physiological states and pathological deviations. Each threshold encapsulates a nuanced understanding of the physiological range within which a particular medical parameter operates optimally. Beyond mere numerical values, these thresholds embody a wealth of diagnostic information, serving as gatekeepers to the realm of pathological conditions. Moreover, the derivation of these threshold values entails a meticulous analysis of empirical data, leveraging sophisticated statistical methodologies to discern patterns and trends. This analytical rigor ensures that each threshold is grounded in empirical evidence, reflecting the diverse spectrum of physiological variability observed within our dataset. As such, these thresholds represent distilled insights gleaned from extensive data mining endeavors, crystallizing the collective wisdom of medical science into tangible numerical benchmarks. To illustrate, consider the impact of blood sugar levels on cardiovascular health. Extensive research has elucidated the intricate interplay between glycaemic control and cardiovascular outcomes, with elevated blood sugar levels posing a significant risk factor for heart disease. By establishing a threshold value indicative of optimal glycaemic control, our framework empowers clinicians to proactively monitor and intervene, mitigating the risk of adverse cardiovascular events. Similarly, the influence of heart rate variability on cardiovascular health underscores the dynamic nature of physiological parameters. With aberrations in heart rate variability serving as harbingers of autonomic dysfunction and cardiovascular morbidity, delineating an optimal threshold for heart rate variability becomes paramount. Through meticulous analysis and data-driven insights, our research

endeavors to unravel the complex tapestry of physiological variability, shedding light on the intricate nexus between medical parameters and disease pathology. Each parameter, be it blood pressure, cholesterol levels, or any other relevant metric, possesses a range of values that can be considered normal, elevated, or indicative of potential health concerns. These thresholds essentially acting as signposts for clinicians, indicating when a parameter turns into a critical territory. In our research framework, these thresholds serve as the cornerstone for the development of fuzzy rules. These rules are akin to decision guidelines for the system, mirroring the nuanced assessments a human expert might make based on these critical thresholds. They encapsulate the logic governing how the system interprets and responds to variations in these medical parameters. For instance, consider a scenario where blood pressure exceeds a certain threshold. This could trigger a fuzzy rule that flags this elevation as a potential concern. The system, empowered by these rules, can then apply this logic across the spectrum of parameters, mimicking the kind of cognitive process a clinician might undertake. The role of these thresholds extends beyond mere numerical values. They encapsulate the collective wisdom of the medical community, distilled into precise cutoff points that signal shifts from normalcy to potential health risks. In essence, they imbue our system with a level of clinical acumen, enabling it to process and interpret medical parameters in a manner that aligns with established medical knowledge. This integration of thresholds into our research framework represents a convergence of clinical expertise and computational precision. It enables our system to not only process raw data but to do so with a level of discernment that mirrors the decision-making process of a trained medical professional. It ensures that our framework doesn't merely generate outputs, but rather, it delivers insights that are contextually rich and clinically relevant.

Table 3.3: Threshold listing for medical parameters

	MP= Medical Parameter		**MPA1= Medical Parameter Analyzer1**		**MPA2= Medical Parameter Analyzer2**		**MPA3= Medical Parameter Analyzer3**	
Feature Name	**MP Min**	**MP Max**	**MPA1 Min**	**MPA1 Max**	**MPA2 Min**	**MPA2 Max**	**MPA3 Min**	**MPA3 Max**
Patient age								
Gender of patient (Sex)								
Heart beat (Thalach)	70	150	80	160	90	170	100	180
Cholesterol (Chol)	199	209	214	224	229	239	239	249
Resting BP (RBP)	80	110	90	120	100	130	110	140
Resting ECG result (RestECG)	120	200	130	210	140	220	150	230
Heart status (Thal)								
Blood Sugar at Fasting (FBS)	100	125	110	135	120	145	125	155
Pain due to Exercise (Exang)								
Slope for Peak-Exercise (Slope)								
ST-Depression due to Exercise (OldPeak)	0.5		1		2		3	

These thresholds are the key player that empowers our system to bridge the gap between raw data and meaningful insights embedded within our system represent a pivotal component, serving as the linchpin in translating raw data into actionable insights. These thresholds function as the guiding beacons navigating the system's decision-making processes, endowing it with the acumen to distinguish between benign fluctuations and potential health hazards. Indeed, they serve as the bedrock upon which our research framework is erected, seamlessly melding the rigor of clinical benchmarks with the finesse of computational analysis. At its core, our system relies on these thresholds to parse through the intricate tapestry of medical data, extracting pertinent information that underpins accurate disease prediction and risk assessment. Through meticulous calibration, these thresholds are meticulously calibrated to encapsulate the nuanced interplay between various physiological parameters and their corresponding health implications. This calibration process involves a meticulous examination of empirical evidence and clinical guidelines, ensuring that each threshold is finely tuned to capture subtle deviations indicative of pathological processes. Moreover, these thresholds are not static entities but rather dynamic constructs that adapt to the evolving landscape of medical knowledge and diagnostic paradigms. As our understanding of disease mechanisms advances and new insights emerge, these thresholds are recalibrated to reflect the latest evidence-based practices and clinical guidelines. This iterative refinement process underscores our commitment to maintaining the highest standards of accuracy and relevance in our predictive models. In essence, these thresholds serve as the linchpin in our quest to unlock the predictive potential of medical data, providing the essential scaffolding upon which our research endeavors rest. They are the foundation upon which our research framework stands, fusing the rigor of clinical thresholds with the precision of computational analysis. Through their judicious application, we aim to harness the full spectrum of information encoded within medical datasets, leveraging it to drive actionable insights and transformative advancements in cardiovascular medicine. By seamlessly integrating the precision of computational analysis with the depth of clinical expertise, we endeavor to forge new frontiers in disease prediction and personalized patient care, ultimately enhancing outcomes and alleviating the burden of cardiovascular disease. Utilizing the threshold values delineated in table 3.3 as a foundation, the subsequent table 3.4 enumerates a

comprehensive listing of features alongside their assigned priorities. Each medical feature is meticulously categorized based on its criticality, a crucial step in orchestrating the synthesis of fuzzy rules essential for the development of our system aimed at fulfilling the overarching research objective. Through this meticulous feature listing process, we discern the intricate hierarchy of medical parameters, each imbued with its unique significance in shaping the predictive prowess of our system. These priorities serve as guiding principles, steering the formulation of fuzzy rules that encapsulate the complex interplay between various physiological variables and their corresponding health implications. Furthermore, the prioritization of features underscores the iterative refinement process inherent in our research methodology, whereby empirical evidence and clinical expertise converge to inform the selection and weighting of relevant parameters. This iterative approach ensures that the resultant fuzzy rules are imbued with the highest degree of accuracy and clinical relevance, paving the way for robust predictive models capable of discerning subtle nuances indicative of pathological processes.

Table 3.4: Priority listing for medical parameters

Features with priority	**Description**
RBP(0,1,2,3) CHOL(0,1,2,3) FBS(0,1,2) RES(0,1,2) RHR(0,1,2) MHR(0,1,2,3) CPD(0,1,2) HD_FH(0,1) HHD(0,2)	0=> Normal 1=> Moderate, 2=> High, 3=> Critical

The features enumerated in the preceding table are meticulously prioritized on a scale ranging from 0 to 3, wherein 0 signifies a state of normalcy, while 3 denotes the highest level of criticality indicative of an elevated risk of experiencing a heart attack. This comprehensive feature set comprises a myriad of vital physiological parameters, each bearing profound implications for cardiovascular health and disease susceptibility. Examples of such pivotal features include blood pressure, heart rate, cholesterol levels, blood sugar levels, daily cigarette consumption, a surrogate measure of smoking addiction, familial predisposition to heart disease, and the presence of symptomatic manifestations indicative of an underlying cardiac pathology. The prioritization of these features is informed by an extensive review of the existing literature, as expounded upon in the preceding chapters. For instance, the inclusion of blood pressure as a high-priority feature underscores its pivotal role in cardiovascular health, with fluctuations in blood pressure levels serving as a harbinger of imminent cardiovascular events, including myocardial infarction. Similarly, parameters such as cholesterol levels and heart rate are accorded high priority due to their established associations with cardiovascular morbidity and mortality, with elevated levels of cholesterol and aberrant heart rates posing a significant risk for adverse cardiovascular outcomes. Conversely, features such as blood sugar levels and smoking habits are assigned relatively lower priorities, reflecting their more gradual impact on cardiovascular health. While fluctuations in blood sugar levels and chronic smoking habits can contribute to the development of cardiovascular disease over time, their immediate effect on disease susceptibility is less pronounced compared to parameters such as blood pressure and cholesterol levels. Furthermore, the inclusion of family history as a feature underscores the intricate interplay between genetic predisposition and disease susceptibility. Indeed, individuals with a familial history of cardiovascular disease are inherently predisposed to a heightened risk of experiencing similar cardiovascular events, underscoring the importance of genetic predisposition in disease pathogenesis. Analogously, the presence of symptomatic manifestations indicative of an underlying cardiac pathology serves as a red flag, warranting heightened vigilance and diagnostic scrutiny to ascertain the underlying etiology and guide appropriate therapeutic interventions. To illustrate, consider a hypothetical scenario wherein two individuals—A and B—present with similar demographic characteristics and medical histories, yet diverge in terms of their

familial predisposition to cardiovascular disease. Individual A possesses a familial history of cardiovascular disease, whereas individual B lacks any such genetic predisposition. Despite their comparable baseline characteristics, individual A is inherently at a heightened risk of experiencing cardiovascular events due to their genetic predisposition, underscoring the role of family history as a potent determinant of disease susceptibility. In conclusion, the prioritization of features within the proposed framework reflects a nuanced understanding of the intricate interplay between physiological parameters, genetic predisposition, and disease susceptibility in cardiovascular medicine. By leveraging this comprehensive feature set, clinicians and researchers can develop robust predictive models capable of discerning subtle nuances indicative of pathological processes, thereby facilitating early risk stratification and targeted therapeutic interventions aimed at mitigating the burden of heart attack. In essence, the feature listing process serves as a crucial prelude to the development of our predictive framework, laying the groundwork for the subsequent generation of fuzzy rules that will drive the system's decision-making capabilities. By anchoring our methodology in the rigorous delineation of feature priorities, we strive to imbue our system with the discriminative acumen necessary to navigate the complex landscape of medical diagnostics and prognostication.

3.4 Research flow

Complete research work has been divided into several stage which starts with literature review, then defining research framework, selecting and working with dataset, performing experiments and analysing results and test implementation.

Literature Review: The research journey boarded upon a meticulous exploration through the vast expanse of existing literature to lay a sturdy foundation for the subsequent stages. The literature review served as a compass, guiding the research towards understanding the intricacies of predicting heart disease and unravelling the complex interplay of medical parameters in determining the probability of a heart attack. Insights gleaned from a numerous of scholarly articles, research papers, and clinical studies provided valuable context and perspective, shaping the research framework and delineating the path forward. The foundation of the research endeavour

lies in an exhaustive review of existing literature. The quest to predict heart attacks has been a long-standing challenge in the field of cardiovascular medicine. Extensive research has been conducted over the years to identify the key factors and parameters that contribute to the onset of heart disease and increase the risk of heart attacks. Studies have consistently highlighted the importance of various medical parameters such as blood pressure, heart rate, blood sugar levels, cholesterol levels, smoking habits, and family history of disease in determining an individual's susceptibility to heart attacks. Through a comprehensive review of published studies, researchers gain insights into the state-of-the-art methodologies, key findings, and gaps in knowledge that inform the design and implementation of their own research framework. The exploration of predicting heart attacks and assessing cardiovascular health status begins with an extensive review of existing literature in Chapter 2. This critical examination delves into past research endeavours, highlighting key findings, methodologies, and insights gleaned from studies focused on cardiovascular health and disease prediction. Drawing upon a wealth of scholarly articles, research papers, and clinical studies, this chapter provides a comprehensive overview of the current state of knowledge in the field, setting the stage for subsequent research endeavours.

Defining Research Framework: With insights gleaned from the literature review, we proceed to define the research framework that will guide the investigation. The research framework outlines the objectives, methodologies, and parameters that will be employed to achieve the research goals. In the context of predicting heart attacks, the framework delineates the medical parameters to be considered, the predictive models to be utilized, and the evaluation metrics to assess model performance. This stage involves careful consideration of various factors, including the selection of appropriate datasets, the choice of machine learning algorithms, and the formulation of research hypotheses. The research framework comprises two primary objectives: predicting the likelihood of heart disease and determining the probability of a heart attack in at-risk individuals. These objectives are driven by a comprehensive analysis of medical parameters and risk factors associated with cardiovascular health. The framework integrates machine learning algorithms, fuzzy logic techniques, and data preprocessing methodologies to develop robust predictive models capable of accurately assessing an individual's cardiac health status. Chapter 3 serves as the cornerstone of the research

framework, where the foundational elements of the study are defined and articulated. Here, the overarching objectives of predicting heart disease and determining the probability of heart attacks are delineated, providing a clear roadmap for the ensuing research activities. The chapter elucidates the methodologies, algorithms, and techniques that will be employed to achieve these objectives, laying the groundwork for the subsequent stages of the research process.

Selecting and Working with Dataset: One of the seminal datasets used in cardiovascular research is the Cleveland dataset, which contains a wealth of information on patients diagnosed with heart disease. Researchers have leveraged this dataset to develop predictive models and algorithms aimed at accurately identifying individuals at high risk of heart attacks. The dataset provides valuable insights into the relationship between different medical parameters and the likelihood of heart disease, serving as a foundation for subsequent research endeavours. The Cleveland dataset serves as the cornerstone of our research efforts, providing a rich source of data for training and testing predictive models. This dataset contains a diverse array of medical parameters, including demographic information, clinical measurements, and patient history, making it well-suited for cardiovascular research. Prior to analysis, the dataset undergoes rigorous preprocessing to address missing values, normalize data, and ensure consistency across variables. In Chapter 4, meticulous attention is devoted to detailing the dataset that forms the backbone of the research endeavour. The Cleveland dataset, renowned for its rich trove of medical parameters and patient information, takes centre stage as researchers delve into its intricacies and nuances. This chapter provides a comprehensive overview of the dataset's composition, structure, and characteristics, offering valuable insights into the variables and attributes that will inform subsequent analyses and experiments.

Performing Experiments: With the dataset in hand, the research transition into the experimental phase, where a series of rigorous analyses and predictive modelling techniques is employed to discern patterns and extract meaningful insights. Utilizing advanced machine learning algorithms and statistical methods, the research uncovers the intricate relationships between medical parameters and the likelihood of heart

disease and heart attack. Through meticulous experimentation and iterative refinement, predictive models are trained and evaluated, with a keen focus on accuracy, sensitivity, and specificity. Experimental procedures involve training machine learning models and fuzzy inference systems using the pre-processed dataset. Various algorithms, including classical methods, ANNs, GA and fuzzy inference systems, are employed to develop predictive models for heart disease and heart attack risk assessment. The experiments involve partitioning the dataset into training and testing sets, training the models on the training data, and evaluating their performance on the testing data using appropriate evaluation metrics. Researchers iterate on the experimental process, fine-tuning model parameters and exploring different feature combinations to optimize predictive performance. Chapter 5 represents the heart of the research endeavour, where experiments are meticulously designed, executed, and analysed to uncover insights into cardiovascular health and disease prediction. Leveraging advanced machine learning algorithms, fuzzy logic techniques, and rigorous statistical analyses, researchers embark on a journey to develop predictive models capable of accurately assessing an individual's risk of heart disease and heart attacks. This chapter provides a detailed account of the experimental methodologies, results, and insights garnered from these endeavours.

Result analysis: The results of the experiments are analysed in chapter 5 to assess the efficacy and performance of the predictive models in accurately predicting heart disease and estimating the probability of heart attacks. Comparative analyses are conducted to evaluate the strengths and limitations of different algorithms and techniques. Insights gained from the analysis provide valuable feedback for refining and optimizing the predictive models, enhancing their predictive accuracy and clinical utility.

Test Implementation: In parallel with the analytical endeavours, the research ventures into the realm of practical implementation, where the predictive models and methodologies are put to the test in real-world scenarios. Through test implementations and validation exercises, the feasibility and applicability of the research findings are assessed, providing valuable feedback and insights for future refinement and optimization. This iterative process of experimentation, analysis, and validation forms the crux of the research methodology, fostering a robust and iterative approach to

knowledge generation and scientific inquiry. This stage involves implementing the developed models in real-world scenarios to assess their practical applicability and effectiveness. Test implementations may involve deploying the predictive models in clinical settings, healthcare systems, or mobile applications designed to assist healthcare providers and patients in assessing cardiac health status and mitigating the risk of heart attacks. Feedback from users and stakeholders is solicited to further refine and improve the predictive models, ensuring their relevance and usability in clinical practice. The culmination of the research journey unfolds in Chapter 6, where the developed predictive models are put to the test in real-world scenarios. We exposed the model with the help of API using “FastAPI” python library which is easy to integrate in existing applications. Test implementations involve deploying the models in clinical settings, healthcare systems, and mobile applications, where they are evaluated for their efficacy, usability, and practical applicability. This chapter provides a comprehensive overview of the test implementation process, offering insights into the challenges, successes, and lessons learned from deploying predictive models in real-world settings.

3.5 Summary

In this chapter, we embark on a comprehensive discussion concerning the proposed research framework. This involves a meticulous exploration of the framework's sequential stages, where we delve into a detailed examination of various machine learning algorithms. Additionally, we expound upon the intricacies of the data employed in our framework, shedding light on the specific features that were meticulously selected for inclusion. The subsequent chapter will pivot towards an in-depth examination of the data source and the terminologies associated with data collection. Here, we aim to provide a clear elucidation of the origins of our dataset, as well as the methodologies employed in its systematic accumulation. Furthermore, we will devote considerable attention to a comprehensive discussion on the medical features that play a pivotal role within our proposed framework. This will include an exploration of their individual functions, priorities, and significance in the larger context of our research endeavour. To offer a preview of what lies ahead, we begin by dissecting the framework into its constituent stages. Each stage, carefully designed and

integrated, contributes uniquely to the overarching objective of predicting heart diseases. By comprehensively examining these stages, we establish a solid foundation for understanding the framework's inner workings. Furthermore, our discussion extends to a detailed evaluation of various ML algorithms that have been strategically employed in the framework. Each algorithm brings its own set of strengths and nuances to the predictive process. Through a clear and accessible explanation, we aim to provide a comprehensive overview of their functionalities and how they synergistically contribute to the predictive accuracy of our model. Equally significant is our exploration of the data that underpins the entire framework. We delve into the details of data selection, emphasizing the specific features that have been curated for inclusion. This process is underpinned by a rigorous evaluation of each feature's relevance and contribution to the predictive capabilities of the framework. As we transition into the subsequent chapter, our focus will shift towards the foundational elements of data sourcing and collection. This critical aspect of our research methodology warrants meticulous attention. We will expound upon the sources from which our dataset was derived and elucidate the methodologies employed to ensure its accuracy, reliability, and comprehensiveness. Moreover, we will dedicate considerable discourse to the medical features themselves. These features, carefully selected based on their clinical significance, play a pivotal role in the accuracy and efficacy of our predictive framework. Through a systematic examination, we will explain the individual functions and priorities of these features, offering a comprehensive understanding of their collective impact on our research work.

Chapter 4: Methodology for Data Acquisition and Preparation

4.1 Introduction

Data collection and processing serve as foundational cornerstones in the successful execution of any research endeavour. This critical phase underpins the integrity and accuracy of the entire research framework, possessing the potential to significantly influence the outcome percentages and subsequently, the efficacy of the model. Through literature reviews, we've observed a noteworthy phenomenon that the same algorithm can yield better results depending upon the nature of the collected data and the approach employed for data processing. This underscores the paramount importance of prioritizing these tasks as preliminary steps before get on to any research experiment. This chapter stands as a comprehensive exploration of these pivotal aspects. It encompasses a thorough discussion of the dataset, its salient features, and the methodologies meticulously employed for data acquisition within the scope of our research. Furthermore, it digs into the complex techniques harnessed for data pre-processing, which encompass a spectrum of operations ranging from data cleansing to the augmentation or elimination of specific columns, all with the overarching aim of refining the dataset for subsequent analysis. The dataset itself represents the bedrock upon which our research endeavours rest. It represents a curated collection of data points, each bearing unique attributes and characteristics pertinent to our research domain. These attributes span a diverse array of parameters, from physiological indicators to demographic details, collectively contributing to a holistic representation of the subject matter under investigation. The careful selection of these attributes is guided by a confluence of factors, including their documented relevance to the research domain, their capacity to yield actionable insights, and their availability within credible sources. Data pre-processing emerges as an indispensable phase in the journey from raw data to meaningful analysis. It encompasses a suite of operations, each tailored to rectify specific facets of the dataset. Impurities and irregularities within the data are

meticulously addressed through processes of cleansing and filtering. This serves to enrich the overall quality of the dataset, expunging any artifacts or outliers that might otherwise distort subsequent analyses. Additionally, the augmentation or pruning of specific columns is executed judiciously, driven by a discerning consideration of their contribution to the overarching research objectives. Moreover, data preprocessing encapsulates techniques to handle missing or incomplete data, a common quandary in real-world datasets. Imputation methods are strategically employed to infer and populate these gaps, ensuring a comprehensive dataset for subsequent analyses. The judicious application of these techniques not only safeguards against potential biases but also fortifies the dataset's capacity to yield robust, reliable results. In the ensuing subsections, we embark on an exhaustive exploration of the specific dataset that underpins our research pursuits. This includes a detailed examination of its origin, structure, and the rationale behind its selection. Simultaneously, we expound upon the intricacies of the data pre-processing methodologies leveraged, elucidating the rationale and techniques deployed at each juncture. This detailed exposition serves to equip the reader with a profound understanding of the foundational steps undertaken in preparation for the subsequent phases of our research framework.

4.2 Overview of Dataset employed in the Research Work

The research work utilizes the "*Cleveland Heart Disease*" dataset, which is a widely accessible medical dataset that can be obtained from the UCI ML repository. The dataset contains information on 303 patients and includes 76 medical parameters [24] out of which, after excluding incomplete patient records, there exists 282 good patient data excluding few erroneous, incomplete and missing data divided into genders i.e., male and female covering different age groups, thereby facilitating a gender-specific analysis, which may yield insights that are vital for tailored medical interventions. In addition to the Cleveland dataset, the research has been further enriched through the inclusion of two additional datasets procured from the UCI repository. These supplementary datasets, sourced from Hungary and VA Long Beach, represent invaluable complements to the primary dataset from Cleveland [24]. Despite the geographic variance in their origin, the datasets share uniformity in their feature set,

affording a consistent framework for comparative analyses. It is imperative to underscore that the uniformity in dataset features across these diverse sources underpins the coherence and comparability of our analyses. This strategic choice allows us to draw meaningful parallels and distinctions between patient cohorts from different demographics, enabling a nuanced understanding of the factors that may influence the likelihood of heart disease across varied populations. Below, *table 4.1* list down all the features in Cleveland dataset with description and datatypes.

Table 4.1: Cleveland Dataset Features: Description and Data Type [24]

S. No.	Feature Name	Description	Data Type
1	id	Identification number of patient	integer
2	ccf	SSN number	integer
3	age	Patient age	integer
4	sex	Gender of patient	binary
5	painloc	Chest pain location	no data
6	painexer	Provoked by exertion	no data
7	relrest	Relieved after rest	no data
8	pncaden	Total of painloc, painexer, relrest	no data
9	cp	Type of pain in chest	ordinal
10	trestbps	Resting BP	integer
11	htn		?
12	chol	Serum cholesterol	integer
13	smoke	Is smoker	binary
14	cigs	Number of cigarettes in a day	integer
15	years	Smoking since number of years	integer
16	fbs	Sugar level in blood at fasting > 120 then 1 else 0	binary

17	dm	Diabetes history	binary
18	famhist	Family history of heart disease	binary
19	restecg	Resting ECG results	ordinal
20	ekgmo	ECG reading - month of exercise	integer
21	ekgday	ECG reading - day of exercise	integer
22	ekgyr	ECG reading - year of exercise	integer
23	dig	Digitalis used furring exercise	binary
24	prop	Beta blocker used during exercise	binary
25	nitr	Nitrates used during exercise	binary
26	pro	Calcium channel blocker used during exercise	binary
27	diuretic	Diuretic used during exercise	binary
28	proto	Exercise protocol	integer
29	thaldur	Test duration of exercise	integer
30	thaltime	ST-depression time	timestamp
31	met	Mets achieved	binary
32	thalach	Max heart rate	integer
33	thalrest	Heart rate at rest	integer
34	tpeakbps	BP at peak of exercise (systolic)	integer
35	tpeakbpd	BP at peak of exercise (diastolic)	integer
36	dummy		?
37	trestbpd	BP at rest	integer
38	exang	Exercise induced angina	binary
39	xhypo		binary

40	oldpeak	S-T depression in ECG	float
41	slope	S-T segment slope in ECG	ordinal
42	rldv5	Height at rest	integer
43	rldv5e	Height at peak exercise	integer
44	ca	Number of major vessels	integer
45	restckm	Irrelevan	?
46	exerckm	Irrelevan	?
47	restef	Irrelevan	?
48	restwm	Rest wall (sp?) motion abnormality	integer
49	exeref	Irrelevan	?
50	exerwm	Irrelevan	?
51	thal	Thallium scan	ordinal
52	thalsev	unused	?
53	thalpul	unused	?
54	earlobe	unused	?
55	cmo	Month of cardiac cath	integer
56	cday	Day of cardiac cath	integer
57	cyr	Year of cardiac cath	integer
58	num	Heart disease diagnosis	ordinal
59	lmt	irrelevant	?
60	ladprox	irrelevant	?
61	laddist	irrelevant	?
62	diag	irrelevant	?
63	cxmain	irrelevant	?

64	ramus	irrelevant	?
65	om1	irrelevant	?
66	om2	irrelevant	?
67	rcaprox	irrelevant	?
68	rcadist	irrelevant	?
69	lvx1	not used	?
70	lvx2	not used	?
71	lvx3	unused	?
72	lvx4	unused	?
73	lvf	unused	?
74	cathef	unused	?
75	junk	unused	?
76	name	patient name	string

Above is the list of all the features available in Cleveland dataset where there are many unused, irrelevant or contains incomplete bad record. These features were filtered out before proceeding the experiments. We also compared this dataset with others which is discussed in next subsection.

4.2.1 Dataset feature analysis

In the quest of the research objectives, we downloaded the dataset from the esteemed UCI Machine Learning repository, specifically, we gathered three distinct datasets: the Cleveland, Hungarian, and VA Long Beach, all of which are fundamentally linked to the domain of heart health. Notably, each of these datasets is characterized by a commonality of 76 attributes, thereby establishing a foundational consistency across our data sources. Among this triad of datasets, the Cleveland dataset emerges as a real cornerstone in heart disease research because of the number of good data. With a

substantial repository of 303 individual patient records, it stands as a base for experimentation with diverse machine learning techniques [24]. In alignment with the specific focus of our study, a judicious selection process was undertaken to sanitize the subset of attributes from this extensive pool. Ultimately, 16 attributes were deemed relevant for inclusion as input features, alongside a pivotal target feature denoted as "num." The rationale behind this selection process was twofold: first, to ensure relevance to our research objectives, and second, to mitigate computational overhead by excluding inessential attributes. These selected attributes, detailed in table 4.3, collectively constitute the foundational layer upon which our analytical endeavours are constructed. Within this paradigm, the "num" feature assumes supreme importance. This feature is characterized as a categorical attribute of ordinal datatype, manifesting values within the inclusive range of 0 to 4. This range encapsulates a nuanced spectrum, wherein a value of 0 signifies the absence of heart disease, while values ranging from 1 to 4 progressively indicate escalating degrees of heart disease severity. This ordinal scale thus serves as a critical key player in our analytical framework, providing a graded categorization of patients based on the presence and intensity of heart disease. In order to enhance the interpretability and efficacy of our analytical framework, a strategic transformation of the "num" feature was undertaken. This transformation yielded a novel binary feature denoted as "HasHeartDisease," thereby effecting a crucial shift in the representational paradigm. Within this binary scheme, a value of 0 designates the absence of heart disease, a value of 1 denotes the categorical values 1 to 4 from the original "num" feature, signifying the presence of heart disease in varying degrees of severity. This binary representation distils the diagnostic insight into a streamlined and actionable format, thereby expediting the decision-making process. Below, *table 4.2* lists the Cleveland dataset features classification with counts to keep the relevance features and remove the odd ones.

Table 4.2: Cleveland dataset feature relevance categorization

Features Categories	**Total Count**	**Features List**
Features with irrelevant/erroneous/empty data	32	"painloc, painexer, relrest, pncaden, htn, dummy, xhypo, restckm, exerckm, restef, exeref, exerwm, thalsev, thalpul, earlobe, lmt, ladprox, laddist, diag, cxmain, ramus, om1, om2, rcaprox, rcadist, lvx1, lvx2, lvx3, lvx4, lvf, cathef, junk"
Features less/no relevant to subject	26	"id, ccf, name, years, ekgmo, ekgday, ekgyr, dig, prop, nitr, pro, diuretic, proto, thaldur, thaltime, met, rldv5, rldv5e, tpeakbps, tpeakbpd, restwm, cmo, cday, cyr, dm, smoke"
Duplicate Features	1	"trestbpd" (Resting BP) [feature 10, 37]
Features relevant to subject	17	"age, sex, cp, trestbps, chol, cigs, fbs, famhist, restecg, thalach, thalrest, exang, oldpeak, slope, ca, thal, num"

In above table, out of 76 available features, 32 features have erroneous or bad data, same has been highlighted at UCI repository [24], 26 features are of less or no relevant

to the subject. Features like id, identity number, name are personal details, similarly features ekgmo, ekgday,ekgyr, thaldur, thaltime, cmo, cday, cyr only talks about the duration so have less impact on subject. Features like dm and smoke talks about if person has any diabatic history or he is a smoker. Regarding smock habit, other feature like cigs is included. Feature *trestbpd* records resting BP of patient which is also covered by feature *trestbps*. all these features have been removed from dataset and finally we left with 17 features where 16 have been taken as input feature and 1 (num) as output; out of it "cigs and famhist" features are behavioural and family history so they are removed while predicting heart disease; feature resting heart rate is also excluded as we already have max heart rate which shows the heart health when we do any activity; but all these three features are included in feature set while predicting the likeliness of heart attack because at that time they are important to be known; below, *table 4.3* list the 17 parameters (input and target).

Table 4.3: List of selected attributes from dataset

Age of patient in year (Age)	Gender of patient (Sex)	Maximum heart rate (Thalach)
Cholesterol level (CHOL)	Type of pain in chest (CP)	Resting Blood Pressure (TrestBPS)
Resting ECG result (RestECG)	HeartStatus (Thal)	Major vessels count coloured by Fluoroscopy (CA)
Fasting blood sugar level (FBS)	Exercise induced Angina (Exang)	Slope for Peak-Exercise (Slope)
S-T Depression (OldPeak)	Diagnosed heart disease (Num)	Cigarette per day (Cigs)
Resting Heart Rate (Thalrest)	Patient Family Heart Disease History (Famhist)	

Below *table 4.4*, lists the features with description and their corresponding values. features.

Table 4.4: Dataset features details

Feature	**Description**	**Values**
Age	Patient's age	
Sex	Patient's gender	Male=1 Female = 0
Cp	Chest pain type	1 = atypical angina 2 = typical angina 3 = asymptomatic 4 = nonanginal pain
Trestbps	BP measured when admitted at the hospital	Systolic/diastolic (in mm Hg) Normal: < 120/80 Abnormal: 120-140/80-120 Critical: 140/120 > [33, 34]
Chol	Cholesterol level	Total Cholesterol: Normal: < 200, Moderate: 200 – 239, High: 239> HDL level: Normal: 40 – 60 Low: < 40, High: 60 > LDL level: Normal: <130, Moderate: 130 – 189, Critical: 189 >

FBS	Blood sugar level at fasting	0 for Normal,1 for High
Restecg	ECG result	0 for Normal,1 and 2 are other types
Thalrest	Heart rate at resting	0-65 – Normal, 65-85 – Average, Above 85 high
Thalach	Max heart rate	Till 100 – Normal, 100-140 – Moderate, 100-180 – High, Above 180 – Critical
Exang	Pain due to physical activity	1, 0 (yes, no)
Oldpeak	Old peak	
Slope	Measurement of the upslope or downslope of the ST segment	1 = up 2 = flat 3 = down
Ca	Number of major vessels colored by fluoroscopy	(0-3)
Thal	Thallium scan	3 = normal 6 = fixed defect 7 = reversible defect
Cigs	Cigarette intake per day	
Famhist	Heart disease history in family	1, 0 (yes, no)

To check if data is good enough for experiment and has no biasness, we compared few parameters and computed the statistics. Below, *table 4.5* represents a statistic showing

comparison between distribution of data among healthy and risky patient based on feature “num” in Cleveland dataset.

Table 4.5: Patient health record distribution as healthy and risky

No HD	53.87%
Patient having HD	46.13%

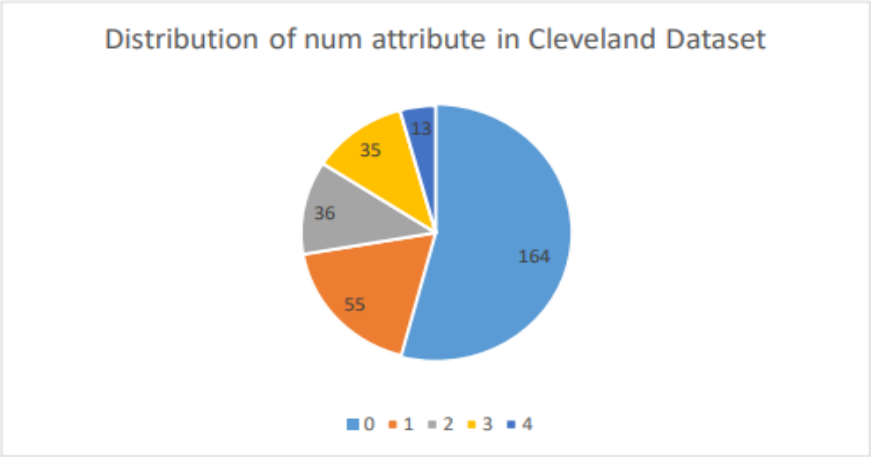


Figure 4.1: Data distribution representation among healthy and risky patient [5]

The patient statistics shown in table 4.5 are presented in the above graphical representation. The dataset has been classified into a numeric status range of 0-4, where 0 represents a healthy individual, and 1-4 indicate an individual's level of risk, with 1 being the lowest risk and 4 being the highest. The graph below illustrates the proportion of healthy individuals to those at risk. The statistical analysis affirms that the dataset has been properly normalized, maintaining a balanced ratio between the two categories, hence, indicates that the dataset is well-suited for experimental purposes.

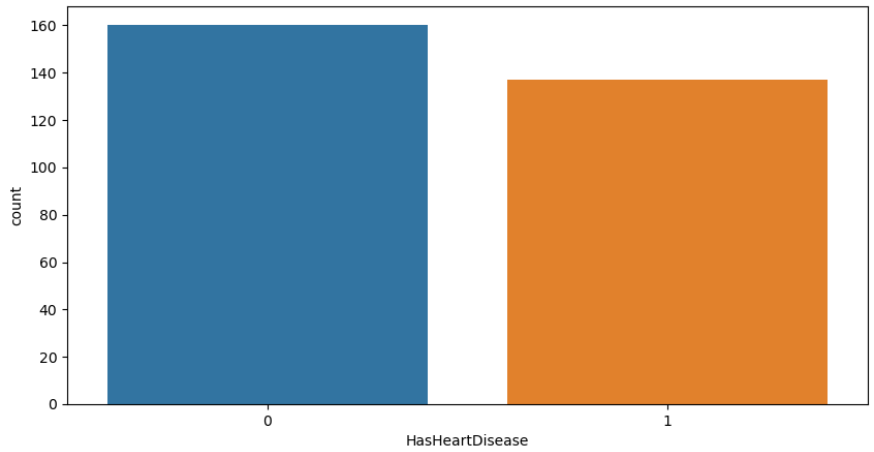


Figure 4.2: A visual analysis showing statistical comparison of healthy and non-healthy patients

Like in Cleveland dataset, Hungarian and VA Long Beach are not very appropriate for experiments due to missing data in few of the features like CA, OldPeak, THAL and others but they have been combined to make a suitable dataset with lesser features and are used in heart attack prediction (more details has been covered in chapter 5).

4.2.2 Significance of clinical parameters

In the preceding sections, we outlined the parameters instrumental in our experiments. Now, let's delve into the significance of these parameters. We'll explore why they hold crucial importance in predicting heart disease and determining the probability of a heart attack.

- *Age:* This health parameter holds significant import, especially in light of empirical evidence which underscores its pivotal role. Studies have consistently indicated a notable escalation in the susceptibility to heart attacks and the development of CVD amongst individuals in the advanced stages of adulthood, particularly those aged above 65 or adults surpassing the age of 40. This demographic cohort emerges as a focal point of concern due to its heightened predisposition to cardiac ailments. The intricate interplay of physiological and environmental factors in this age bracket accentuates the gravity of this health parameter. It serves as a critical juncture for intervention strategies, necessitating vigilant monitoring and tailored healthcare measures.

Moreover, the discernible contrast in risk profiles across age groups underscores the imperative for targeted preventative measures and early detection protocols. Consequently, the age factor assumes a paramount significance within the broader spectrum of heart health assessment, guiding the formulation of proactive healthcare strategies tailored to the unique needs of distinct age demographics. In essence, this parameter serves as an essential in the orchestration of comprehensive cardiac care, warranting thorough attention in the design and implementation of healthcare interventions [66].

- *Gender:* Analysing the available data, a notable trend emerges in the realm of cardiovascular health. It is noteworthy that the incidence of CVD manifests at a relatively lower frequency among women when compared with their male counterparts. This observed gender disparity in CVD frequency prompts a closer examination of underlying physiological and demographic factors. Interestingly, while women may exhibit a lower incidence rate, it is crucial to acknowledge the nuanced dimension of cardiovascular health outcomes in this demographic. Contrary to the relatively lower occurrence, women tend to experience a typical course in the aftermath of acute cardiovascular events. Mortality rates among women afflicted by such events tend to be comparatively higher, a phenomenon that warrants careful consideration and targeted intervention. The observed trends not only shed light on the physiological intricacies at play but also accentuate the critical need for tailored preventative measures and responsive post-event care for women in the context of cardiovascular health [67].
- *Heart rate:* This medical parameter is most associated with heart disease and is very useful in diagnosing the symptom of imbalance related to heart. Within the domain of cardiovascular health, heart rate emerges as a crucial physiological parameter, bearing a deep association with the onset and progression of heart disease. It serves as a reliable indicator, offering valuable insights into the equilibrium of

cardiac function. Any abnormal fluctuation in heart rate should be alarmed and preventive measure should be taken. Monitoring fluctuations in heart rate assumes paramount importance, as deviations from the established norms can signify underlying cardiac irregularities. The assessment of heart rate encompasses two distinct contexts: resting heart rate, recorded when an individual is in a state of rest, and maximum heart rate, obtained during periods of rigid physical activity. These two states are characterized by varying threshold values, each holding distinctive clinical significance. The resting heart rate serves as a baseline metric, offering a glimpse into the heart's performance at a state of relative rest. Deviations from the established resting heart rate range can serve as an early indication of potential cardiac issues. Conversely, the assessment of maximum heart rate is conducted during moments of physical exertion, providing valuable data regarding the heart's capacity to respond to increased physiological demands. This parameter, too, boasts its own set of threshold values, demarcating the bounds of normalcy. Heart rate can be measured as resting when a person is not doing any physical activity or during physical activity which measures as maximum heart rate. It is imperative to underscore the criticality of periodic heart rate monitoring, for any significant changes from the defined parameters alert prompt attention and necessitates the implementation of preventive measures. The nuanced evaluation of heart rate within the clinical context, accounting for both resting and maximum states, thus emerges as a vital component in the comprehensive assessment of cardiovascular health. This proactive approach not only enhances the prospects of effective cardiac management but also contributes to the overall wellbeing and quality of life of individuals [68-70].

- *Blood pressure:* Like heart rate, this is also an important medical parameter and reflects problem in heart and event it can harm the heart if the count is higher for longer [71,72, 110]. A high blood pressure

should always critically alarmed and immediate precaution is needed. Same way low blood pressure is also very important.

- *Blood sugar:* Elevated blood sugar levels, a hallmark of diabetes, constitute a significant risk factor for heart attacks and cardiovascular complications. Individuals grappling with diabetes exhibit an augmented susceptibility to heart-related ailments, rendering the management of blood sugar levels with high priority in the context of heart health. Persistently high blood sugar levels exert deleterious effects on the cardiovascular system. The intricate network of blood vessels in heart muscle, known as coronary arteries, is particularly vulnerable to the detrimental impact of elevated sugar levels. Prolonged exposure to heightened blood sugar can instigate a cascade of pathological processes, culminating in the narrowing and stiffening of these crucial vessels. This adverse transformation, termed atherosclerosis, impedes the seamless flow of blood to the heart muscle, thereby elevating the risk of heart attacks. Furthermore, diabetes can instigate a state of chronic inflammation within the body, amplifying the propensity for atherosclerotic plaque formation. These plaques, composed of cholesterol deposits and inflammatory cells, can obstruct coronary arteries, culminating in a potentially catastrophic reduction of blood flow to the heart. In severe cases, this compromised blood supply can precipitate “myocardial infarctions”, commonly referred to as heart attacks. Data from the NHA from 2012 shows that approx. 65 out of 100 people having diabetes may face serious problems related to HD or stroke [51][93]. This statistical revelation underscores the imperative for meticulous blood sugar management, positioning it as a linchpin in averting the perilous ramifications of heart-related afflictions among individuals with diabetes. Hence, it is advisable to individuals with diabetes, in collaboration with their healthcare providers, to implement periodic measures for blood sugar regulation. This endeavour not only mitigates the risk of heart attacks but also safeguards overall

cardiovascular well-being. By prioritizing glycaemic control, people having diabetes can fortify their defence against the risk of cardiovascular disease, promoting a foundation for sustained health and vitality.

- *ECG result:* The ECG feature emerges as a pivotal tool in assessing cardiac well-being. The electrical spike encapsulates crucial insights into an individual's heart health status, studying the patterns and spikes in the ECG tracing, clinicians can glean invaluable information regarding the heart's rhythm, potential irregularities, and overall functioning. It serves as a non-invasive yet insight informative window into the workings of the heart. The ECG feature assumes dominant significance in the comprehensive health evaluation of the cardiovascular system.
- *Chest pain type (CP):* Chest pain type also indicates abnormalities to heart but always it is not the sole reason for heart disease or attack, rather it mixes with other medical parameter and can help in indicating the disease. If a patient experience chest pain, particularly if it is accompanied by shortness of breath and feel fatigue, it is important to seek medical attention promptly.
- *Exercise induced Angina (Exang):* This is a type of chest pain or discomfort which causes during physical activity and indicates risk to heart patient who feels angina during exercise.
- *Slope for Peak-Exercise (Slope):* This is the measure of the rate at which heart rate increases during exercise. Though heart rate (HR) usually increases during exercise to meet the oxygen requirement but the rate at which HR increases, depends on various factors like age, fitness level or any medical condition like existence of heart disease or sometime failures.
- *Major vessels count coloured by Fluoroscopy:* Fluoroscopy, a medical imaging technique used to obtain images from inside of body using X-

Rays, can be used to visualize the major vessels of the heart, including the coronary arteries. The number of vessel count is vital to identify the risk e.g, lesser number of visible vessels indicate a blockage or narrowing of the arteries, decrease in count indicates the progression of risk to heart.

- *Cholesterol:* This parameter digs into a crucial symptom of heart disease. Elevated cholesterol levels in an individual can lead to the accumulation of fatty deposits within the veins, can prompt a heart attack by impeding the seamless circulation of blood and oxygen to the heart and other vital organs. This process occurs as the excess cholesterol in the bloodstream adheres to the arterial walls, gradually forming plaques. Over time, these plaques can restrict blood flow, thereby diminishing the supply of oxygen to the heart muscle, which then, can conclude to myocardial infarction or a heart attack. Understanding the intricate dynamics between cholesterol levels, monitoring and managing it emerges as a critical component in the broader strategy of cardiovascular health and risk mitigation.
- *S-T Depression (OldPeak):* S-T depression, is a measure of the ST segment of the ECG represents the period of time when the ventricles of the heart are contracting. Depression occurs when the ST segment is below the baseline, indicating that the heart is not receiving enough oxygen during exercise or activity.
- *Cigarette consumption:* This should not be considered as medical parameter but rather it is a habit which can cause illness. Now a days it has been observed that people started consuming cigarette a lot which can damage their heart and cause heart disease or heart failure as well if they are chain smoker [110]. According to the study, individuals who smoke experience a twofold increase in cardiovascular complications, their blood pressure remains at higher rate as compared to the one who are less addictive of smoke.

- Family history: The family medical background holds significant relevance in the assessment of an individual's susceptibility to any ailments. If any family member has the history of disease such as diabetes or heart disease, the likelihood of developing similar health concerns escalates. For instance, if there's a family predisposition towards diabetes, it heightens the probability of the individual also grappling with this metabolic disorder. Diabetes, in turn, can exert a profound influence on cardiac health owing to its impact on blood sugar levels. Elevated or poorly managed blood sugar levels can precipitate a range of cardiovascular complications over time. Recognizing and understanding these familial health markers is the key in pre-emptively managing and mitigating potential cardiovascular risks [94].

4.2.3 Comparison of Cleveland dataset with another available dataset in UCI repository

UCI repository has three more datasets along with Cleveland dataset that is “Hungarian dataset, Switzerland dataset and VA Long Beach dataset”. All four dataset are similar to each other and are publicly available for research but when compared Cleveland dataset with other three, we found this dataset having lesser number of bad and missing data so we preferred this dataset for continuing our research work for example as an illustration, the Hungarian dataset encompasses 294 instances and encompasses 76 attributes. Within this dataset, there are a total of 491 instances with missing values having columns having eight categorical columns and five numeric columns. Below, *table 4.6* shows the list of features with categorical and numerical data.

Table 4.6: Comparison of heart disease datasets

Dataset	Source	Instances	Attributes	Target Variable
Cleveland	Cleveland Clinic Foundation	303	76	Presence or absence of heart disease (ordinal scale 0-4)
Hungarian	Hungarian Institute of Cardiology	294	76	Presence or absence of heart disease (binary)
Switzerland	University Hospital, Zurich	123	14	Severity of coronary artery disease (ordinal scale 0-4)
VA Long Beach	Department of Veterans Affairs	200	76	Presence or absence of heart disease (binary)

The above comparison in *table 4.6*, clarify that Cleveland and Hungarian datasets are the largest in terms of instances and attributes, and have the same target variable of heart disease presence. While the VA Long Beach dataset is similar to the above two in size and target variable, the Switzerland dataset has fewer instances and attributes but measures the severity of coronary artery disease on an ordinal scale. Overall, the Cleveland and Hungarian datasets are the most commonly used for heart disease research. The number of records in each dataset consists of both good and bad data, shown in figure 4.3.

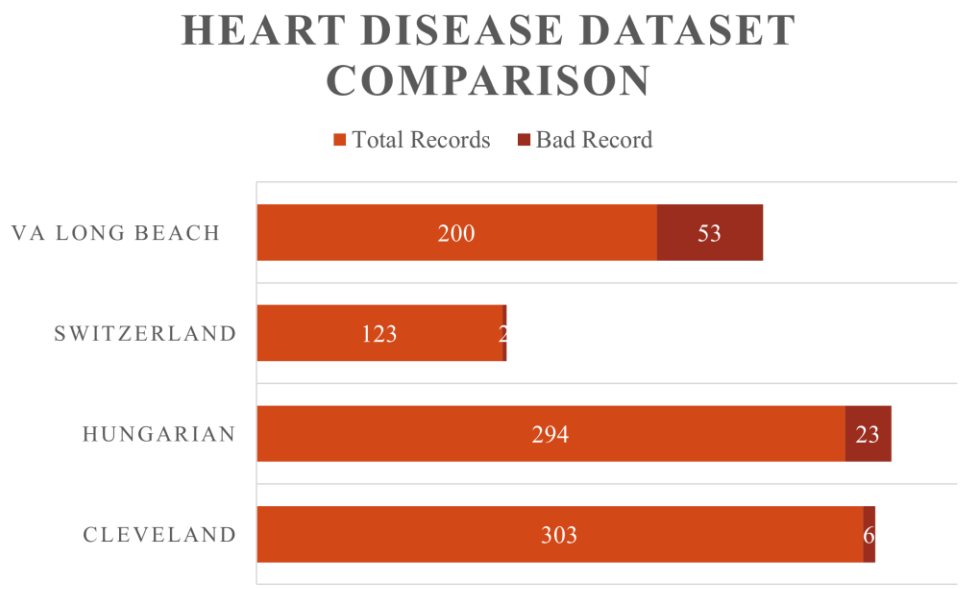


Figure 4.3: Dataset record comparison

In Figure 4.3, the statistics regarding the Heart Disease Dataset, particularly the Cleveland dataset, are depicted. This dataset comprises a total of 303 records, with a minor proportion of 6 records classified as bad data. Similarly, the Hungarian dataset encompasses 294 records, with a slightly higher count of 23 bad data instances. In contrast, the Switzerland dataset contains 123 records, with only 2 instances categorized as bad data. Lastly, the VA Long Beach dataset comprises 200 records, with a more significant portion of 53 records identified as bad data. These statistics provide insights into the quality and integrity of the data across different subsets of the Heart Disease Dataset, shedding light on potential discrepancies or anomalies that may impact subsequent analyses and interpretations, it was found that Cleveland dataset has lesser number impurities as compared to other datasets.

4.2.4 Limitation of Cleveland dataset

Though at one side we found many positive points in Cleveland dataset like it is very popular, balanced, have sufficient number of features to train models but during data preparation before using in the experiments, we found few limitations as well in this dataset, few of them are listed below:

- *Number of features:*

- Out of 76 features, there are some features which related to patient personal details and were left like id, name, ccf (social security number).
- There are many features which are not in use and their details are not provided in dataset so that were left during experiments for example htn, dummy, lvx1, lvx2, lvx3, lvx4, lvf, cathef, junk, thalsev, thalpul, earlobe, cmo, cday, cyr, and etc. These columns are not included in the experiments.
- There are few which are irrelevant like rldv5, rldv5e, exerckm, restckm
- There are some duplicate features as well like trestbps (feature number 10 and 37).
- Out of 76 features, 36 features were removed due to erroneous/empty data, 26 were irrelevant and 1 duplicate. After removing all such features, we left with 17 good features.

- *Number of good records:*
  - From a set of 303 records only 282 are deemed accurate, others contain incomplete data for many features included in the research work, for example for features *"resting heart rate", "is smoker", "cigarette per day", and "heart disease family history"* there are 14 missing records, along with some incomplete or erroneous data (*figure 4.4*).
- *Data division:*
  - The data is imbalanced with unequal distribution among different age groups. The majority of records with positive heart disease are in the age range of 40-75.
  - The data is also imbalanced between male and female patients, with a significantly higher number of male records compared to female records.
- *Healthy vs Risky records:*

- Out of 282 accurate data 127 were identified for patients as risks while others were of normal patients.
- Only six of the 127 at-risk patients were in the Adult-Group1 age range (18-40), and only one was in the Old-Group2 range. This indicates that the data is not correctly classified according to age groups.

Below, figure 4.4 shows the missing data for features “resting heart rate”, “is smoker”, “cigarette per day” and “heart disease family history” which is highlighted in orange. This also shows wrong data (highlighted in red) for resting heart rate for example “655.7, 1110, 9 and 1” which is usually not possible in real scenario.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Age	Sex	ChestPain	RestingBP	SerumCh	FastingBl	RestingEC	RestingHe	MaxHeart	ExerciseIn	OldPeak_	PeakExer	VCA	ThalliumS	HeartDise	HasHeartI	IsSmoker	CigratePe	HeartDisease
58	1	4	114	318	0	1	65	140	0	4.4	3	3	6	4	1	0	0	1
58	0	4	170	225	1	2	1	146	1	2.8	2	2	6	2	1	1	44	0
56	1	2	130	221	0	2	19	163	0	0	1	0	7	0	0			
56	1	2	120	240	0	0		169	0	0	3	0	3	0	0			
67	1	3	152	212	0	2		150	0	0.8	2	0	7	1	1			
55	0	2	132	342	0	0		166	0	1.2	1	0	3	0	0			
44	1	4	120	169	0	0		144	1	2.8	3	0	6	2	1			
63	1	4	140	187	0	2		144	1	4	1	2	7	2	1			
63	0	4	124	197	0	0	9	136	1	0	2	0	3	1	1			
41	1	2	120	157	0	0	655.7	182	0	0	1	0	3	0	0			
59	1	4	164	176	1	2	1110	90	0	1	2	2	6	3	1			
57	0	4	140	241	0	0		123	1	0.2	2	0	7	1	1			
45	1	1	110	264	0	0		132	0	1.2	2	0	7	1	1			
68	1	4	144	193	1	0		141	0	3.4	2	2	7	2	1			
57	1	4	130	131	0	0		115	1	1.2	2	1	7	3	1			
57	0	2	130	236	0	2		174	0	0	2	1	3	1	1			

Figure 4.4: Cleveland dataset showing incomplete records for few of the features

Below figure 4.5 shows the correlation between dataset features using heat map. A heat map is a visual representation of data where values are depicted using a spectrum of colours. It is a way to visualize the relationship between two or more variables. The values in a dataset are represented as colours, where higher values are typically shown in warmer colours like red or orange, and lower values are shown in cooler colour like blue or green.

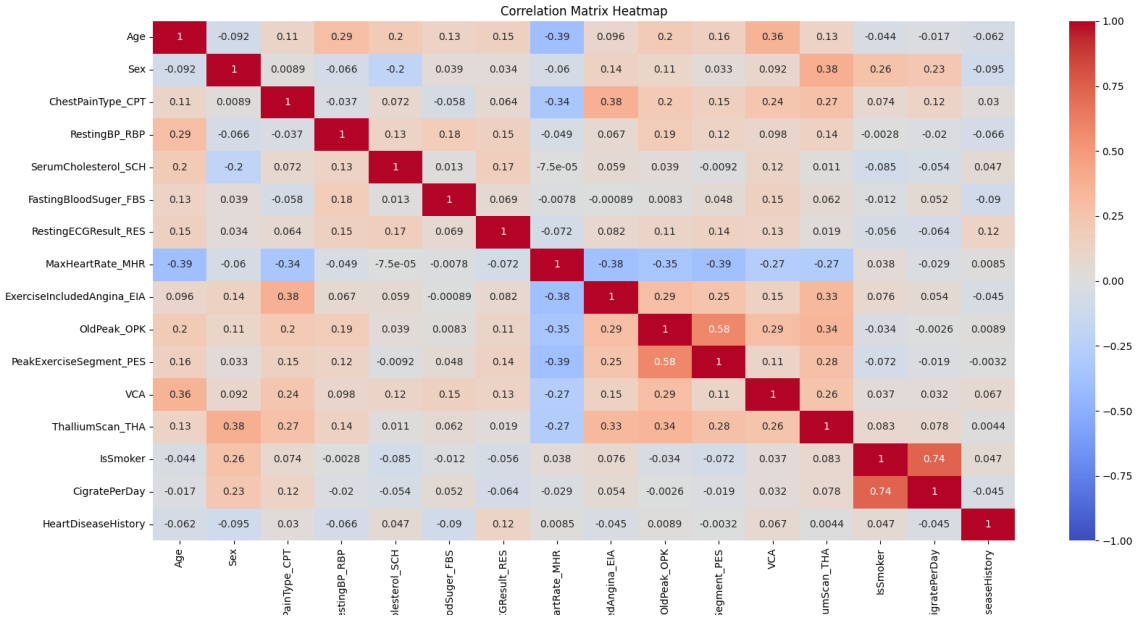


Figure 4.5:Heatmap showing correlation between dataset features

4.3 Stages in data pre-processing

Data pre-processing is an essential stage in data analysis and there are several stages in data pre-processing. After completing all of these, a data can be treated as ready to go for model training.

4.3.1 Feature Encoding

We applied the technique of "label-encoding" to transform the ordinal data into numeric form. For this we used the most famous data science library scikit-learn. In the dataset there is a feature named *"num"* which has details related to patient being diagnosed of heart disease. This feature contains value from 0 to 4, with 0 indicating a healthy patient and 1-4 signifies the risk level of HD patient. Since we are using binary classification in our experiments, hence converted these values into binary where 0 stands for healthy and 1 means patient at risk. For this, label encoding has been combined with mapping function, a data mapper added which has details for all the categories, added one more feature 'HasHeartDisease' and stored converted values there. *Figure 4.6* shows the schematic representation of feature encoding process.

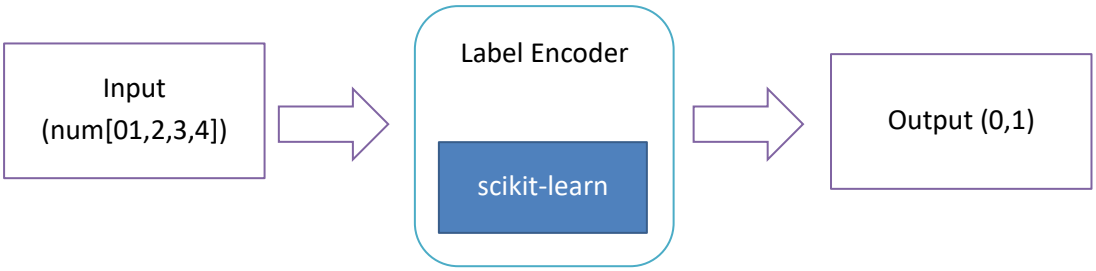


Figure 4.6:Block diagram showing feature encoding

4.3.2 Feature Selection

Cleveland dataset has 76 features including irrelevant and unused features discussed in section 4.2; this can cause lowering down the model accuracy. So, to optimize the model result, it is ideal to choose only the appropriate features that provide better efficiency of the model. Along with improving model efficiency, feature Selection also help in reducing over-fitting, complexity and cost of the model, reduces useless features from the dataset which irrelevant, improves accuracy, and reduces the training time. In the research work two feature selection technique i.e., Chi-Square and mRMR methods has been used. Below, *figure 4.7* shows schematic representation of feature selection flow.

Figure 4.7: Application of feature selection to get relevant features

4.3.2.1 Benefits of feature selection

There are several benefits of feature selection which are listed below:

- *Reduce Over-fitting/ Improve model accuracy:* Including irrelevant or redundant features in a model can result in overfitting and cause lower model accuracy. By selecting the most relevant features can improve model accuracy.

- *Reduce training cost:* Unnecessary features can be eliminated using feature selection that leads to faster computation times and reduced memory uses, hence reducing training cost as well. In our work, the dataset has 76 attributes which need to overview and elimination for faster processing.
- *Improve model interpretability:* A model, if has relevant features is easier to interpret and understand in comparison to the one that includes many irrelevant or redundant features.
- *Increase generalization:* By choosing only the appropriate features, the technique can help to increase the generalization of the model, making it more effective at predicting outcomes on new or unseen data.

4.3.3 Feature Scaling

Feature scaling is used in standardizing the dataset features in a fixed range and handles highly varying values. Without application of this method, system is not able to differentiate weighted values and consider both as one regardless of their unit associated like 1kg will tend to be smaller than 100 gram or vice versa [73]. Figure 4.8 shows the feature scaling flow in data pre-processing.

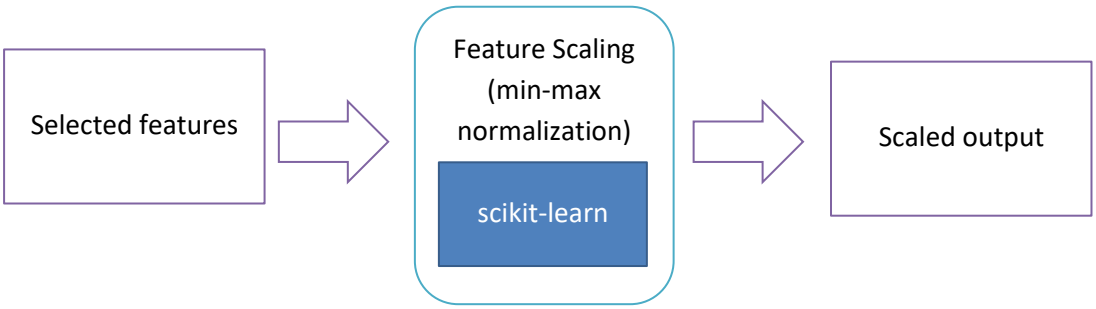


Figure 4.8: Data scaling flow

In above *figure 4.8,* schematic representation of feature scaling is shown using min-max normalization method.

4.3.3.1 Way to feature scaling

There are two ways to scale a feature:

- *Normalization or Min-Max Normalization:* It involves a process of subtracting the minimum value of each feature from their respective values, followed by

division by the range (i.e., the difference of min, max of feature values). This procedure scales the values within a standardized range of 0 to 1. This is shown in equation 4.1.

$$Xnorm = \frac{X - min(X)}{max(X) - \min(X)} \quad (4.1)$$

In this context, the set of data values designated as X undergoes normalization. Here, "min(X)" represents the minimum value, while "max(X)" stands for the maximum value. The transformed dataset, post-normalization, is denoted as "*Xnorm"*.

- *Standardization:* It subtracts value of each feature by the mean of all the values of features and divide by standard deviation, the square root of variants. Standardization can be calculated as in equation 4.2.

$$Xstand = X - \frac{mean(X)}{StandardDeviation(X)} \quad (4.2)$$

Here, *X* is the set of data values to standardize, "*mean(X)" and "StandardDeviation(X)"* are the arithmetic mean (average) and the standard deviation of the data values in *X* respectively. *Xstand* is the resulting standardized set of data values.

4.3.4 Data Splitting

Upon completing the data pre-processing, the final step involves splitting the data into two distinct sets - training and test. This division facilitates both the training and evaluation of the model. Typically, a widely accepted practice is to allocate 80:20 ratio of the data for training/test purposes. For this purpose, we used "Scikit-learn" python library.

4.4 Summary

In this chapter, we discussed about the process of data pre-processing, the stages and terminologies. We discussed that how different concepts like data encoding or data scaling can help in improving model output, how they can help in fasten the training and so on. Data pre-processing stage is very important in model optimization and should be used before proceeding with any experiments.

Chapter 5: Experiment Scenarios

5.1 Introduction

This chapter constitutes a key stage in the research work, building upon the foundation laid in preceding sections. In previous chapters we discussed about the research objectives, proposed framework, methodologies and dataset details. This chapter mainly covers the experimental scenarios covered to complete the proposed research work. In this chapter, experimental scenarios related to proposed framework has been discussed in details. The experimentation unfolds across different phases, each meticulously calibrated to align with the research objectives. Commencing with the introduction of input parameters into the system, the process starts through the realms of data pre-processing, concluding in the curation of pertinent features. The ensuing phase is marked by the establishment of a neural network, an essential component of our predictive model. Here, the neural network is methodically trained, absorbing the Cleveland dataset [24], a base of our experimental domain. This dataset, renowned for its comprehensive attributes, forms the crucible upon which our model refines its predictive insight. A crucial component in the research work is the creation of a rule base, a repository of decision making that offers outcome to our fuzzy inference system. The succeeding step involves the actualization of the inference system itself, an intricate collection of logic gates and membership functions, standardized to navigate the complexities of medical parameters. Our experimental panorama encompasses a spectrum of technical specifications, each tailored to strengthen the efficacy of our research framework. The dataset, selected with exacting precision, derives from the renowned Cleveland repository, spanning a rich set of attributes, is instrumental in furnishing our model with the requisite depth of insight along with the feature selection approach, an essential key player in the experimental designs, serves a crucial role wherein extraneous attributes are filtered, leaving behind a nucleus of indispensable parameters. The process of cross validation, akin to a crucible's tempering, ensures that our model's predictive ability is refined to a razor's edge. As the model training phase

unfurls, it adapts the details of the Cleveland dataset, forging an interdependent relationship between data and algorithm. The rule creation process meticulously shapes the heuristics that guide our fuzzy inference system. The membership functions, akin to a reference of medical semantics, confer our fuzzy system with the linguistic conditions requisite for decoding relevant medical parameters. The subsequent phase entails the experimental scenarios necessary for the robustness of our research framework, checks that our model's efficacy is truly revealed, as it navigates an array of scenarios, each designed to challenge its predictive efficacy. Ultimately, the chapter concludes in the unveiling of test results that underscores the efficacy and relevance of our proposed research framework. In below sections we will discuss all the experiment experimental scenarios in details including the system specifications used for experiments, dataset detailing, feature selection, cross validation, model training, rule creation, membership functions to be used in FIS, test cases, test results and etc.

5.2 System specification

To conduct experiment and to train the model smoothly a good quality of hardware having good configuration and software which is capable of performing the desired action is needed. All experiments have been conducted on machine with configuration: “64-bit windows 11 operating system, 11th Gen Intel(R) Core (TM) i7-1195G7 @ 2.90GHz 1.80 GHz and 16 GB RAM. Python libraries like NumPy, pandas, sklearn, Keras, and TensorFlow using PyCharm IDE and C# using Visual Studio 2019 Community version”. Along with this, Azure Machine Learning Designer and Google Colab have also been used as a part of research.

5.3 Dataset

Dataset has already been discussed in details in chapter 4 but to re-iterate the things, in this research work, Cleveland dataset along with Hungarian and VA Long Beach for HD has been downloaded from UCI for the experiment, data were analysed from all the datasets and post analysis the Cleveland dataset has been concluded as best in terms of data and features and also from the popularity among other researchers.

5.4 Evaluation Metrics

The experiments have been performed using ML techniques with many classification methods as well as hybrid system of neural network family. Several classifiers have been used to identify the model's effectiveness that fits for predicting heart disease and to ensure that model is giving expected result, model evaluation metrics comes into the picture. There are several evaluation metrics that can be used to calculate these metrics. In the research work we mainly computed model efficiency based on parameters accuracy, sensitivity, specificity.

5.4.1 Confusion Matrix

The analysis of results has been performed based on efficiency parameters, such as accuracy. The accuracy has been computed by using a 2X2 confusion matrix that has true and false combinations. Table 5.1 provides a list of these combinations. Parameters computed on the basis of type of heart disease prediction probability which decides if a patient is on risk or not. Along with this, it also calculates if prediction is done correctly by model or not.

Table 5.1: Confusion matrix

Actuals	**Positive Prediction (1)**	**Negative Prediction (0)**
Has HD (1)	TP	FN
No HD (0)	FP	TN

Above *table 5.1* shows the evaluation parameters of model which has been used in calculating the model score using confusion matrix by applying below formulas on the matrix listed in the table.

TP “True Positive” –Spotted as HD and predicted as HD +

TN “True Negative” –Not spotted as HD and predicted as HD negative

FP “False Positive” –Not identified HD but predicted as HD + (Called as Type1 error)

FN “False Negative” –Has HD but predicted as HD negative (this is also called Type2 error)

Based on above matrix, below equations shows the calculation for accuracy, specificity, sensitivity, Classification error and Precision:

$$Accuracy\ (ACC)\ = \frac{TP+TN}{TP+TN+FP+FN} X\ 100\% \qquad (5.1)$$

$$Sensitivity\ (SENS)\ = \frac{TP}{TP+FN} X\ 100\% \qquad (5.2)$$

$$Specificity\ (SPEC)\ (No\ HD) = \frac{TN}{TN+FP} X\ 100\% \qquad (5.3)$$

$$Classification\_Error\ \frac{FP+FN}{TP+TN+FP+FN} X\ 100\% \qquad (5.4)$$

5.5 Material and Methods

This segment starts the discussion on the techniques and the scenarios used in experiments. This section mainly covers the experimental scenarios which have been experimented to complete the proposed framework. Research work starts with heart disease prediction and later stage it predicts the heart attack probability which then exposed to public via API endpoint. Detailed discussion on its framework has been accomplished in chapter 3. In next few subsections, experiment scenarios have been discussed.

5.5.1 Experimental scenarios for prediction of heart disease

To cover up the first stage, various criterion was experimented so that best features could be selected and model efficiency could be enhanced. Experiment has performed

for the proposed framework where neural network has been clubbed with genetic algorithm to build GANN network. To compare the efficiency and justify the selection of proposed hybrid model, experiments has been executed on six classical ML methods "SVM", "KNN", "LR", "DT", "RF", "NB" and "NN" and results from GANN model have been compared with classical and neural network to get the best model. Algorithm to get the best relevant features has been applied and cross validation has been used for model validation (elaborated below) and to select the optimal feature set. Below, *table 5.2* lists the parameter setup for the experiments and *figure 5.1* shows the flow diagram of experiment flows for ML algorithms.

Table 5.2:Parameter setup for experiments

Description	**Value**
Dataset	Cleveland from UCI repository
Total number of features	76
Included in experiment	14 (13 input, 1 target)
Number of samples	303
Samples excluding bad data	282
Cross Validation	5-Fold
Data split ratio [Train: Test]	80:20

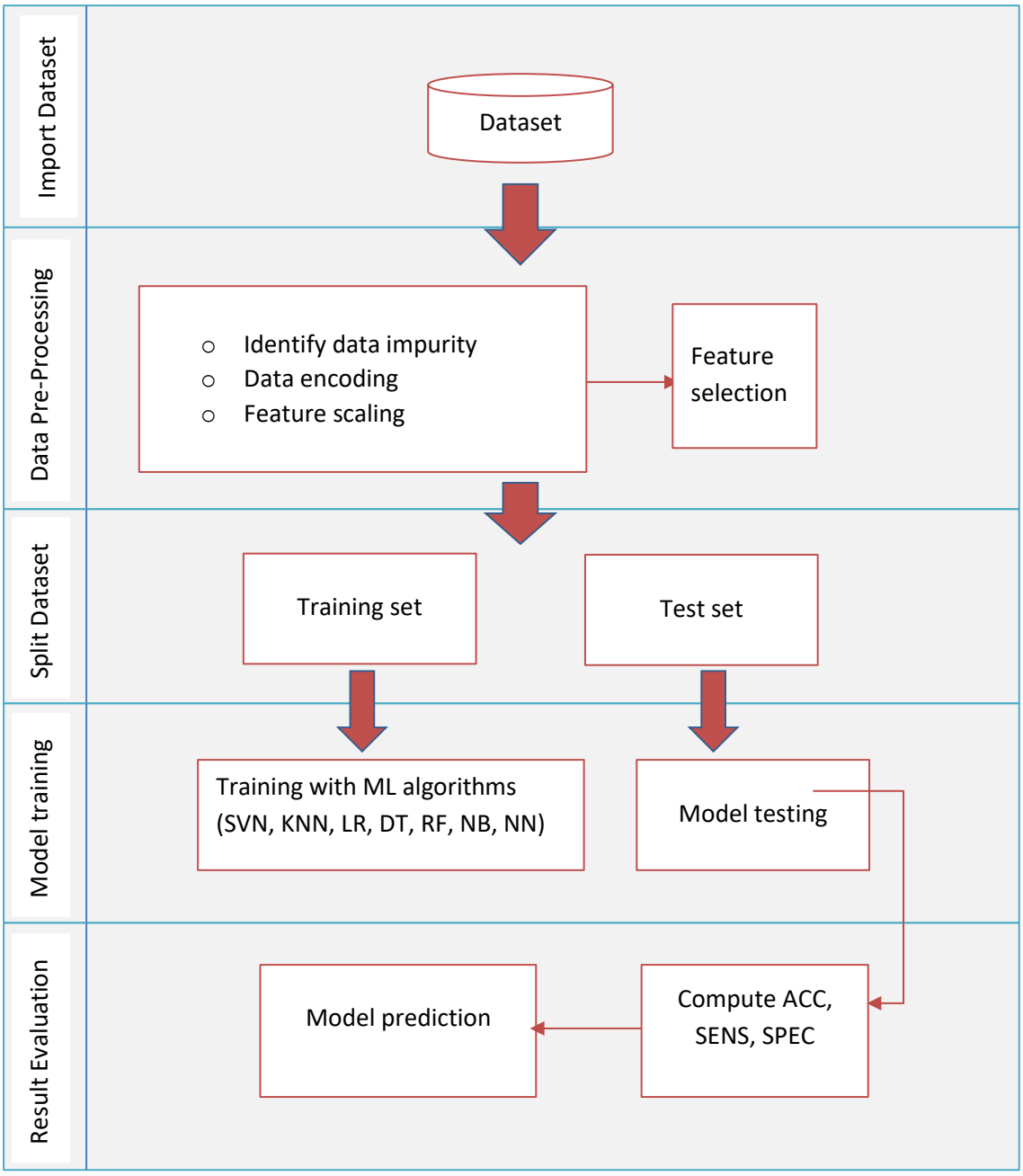


Figure 5.1: Schematic flow diagram of experiment flow for classical methods

Above *figure 5.1* demonstrates the experiment divided into five stages which are briefly explained below:

- *Stage 1 [Import dataset]:* This is the first stage in experiment. In this stage dataset is being imported in the system for further use. Here to import dataset python library has been used which first import already available

Cleveland dataset in csv format into the system and then convert data into data frame for further use.

- *Stage 2 [Data optimization]:* Pre-processing data is an important stage in any experiment; it impacts the model efficiency as well. In the second stage of experiment imported dataset undergoes pre-processing stage wherein data is cleaned according, missing data is filled if required, noisy data is removed, after that it goes through the data transformation where data is encoded and scaled (normalized or standardized) to make it fit according to the model, after that it goes through data reduction stage where extra data is reduced in terms of size and dimension and finally appropriate features are selected for next stage [74]. Data transformation is done using "sklearn" library and mRMR feature selection.
- *Stage 3 [Split dataset]:* The next stage is splitting dataset into train and test set. This is usually done in a particular ratio and most common is 80:20 which is practiced in the experiments as well. To accomplish this task, the widely utilized "sklearn" python library was employed. Following the dataset split, it was segregated into two segments - one for training purposes and the other for testing. Data splitting of dataset has been approached using "sklearn python library".
- *Stage 4 [Model building]:* In this stage, model is built using pre-existing information like available pre-processed dataset which is split into parts. Model has been trained using each ML algorithm i.e., KNN, one by one like with different settings to optimize the model efficiency.
- *Stage 5 [Result evaluation]:* This is the final stage wherein result is evaluated from the built model. This is done on the basis of confusion metrics wherein we weigh efficiency of model based on parameters like "accuracy", "sensitivity" and "specificity".

In below sections we have discussed the experimental scenarios covered in the research work.

5.5.1.1 Experiment applying feature selection

Experiment started with adding feature selection as a part of experiment along with the rest which has been used in the experiment discussed in above section. To apply feature selection there are several methods available but we choose "Chi-Square and mRMR" feature selection method; they are discussed below.

- *Feature selection using Chi-Square:* The technique has been applied while performing the experiment and used to select the most appropriate features for a classification problem. The utilization of the chi-square test in our experimentation process represents a critical facet of our research methodology. This statistical tool has been deployed to determine and designate the most pertinent features in the context of a classification problem. By quantifying the interdependence between the target variable and the feature variable, the chi-square test furnishes us with a metric that guides our feature selection process. The crux of this operation involves the computation of statistics for each feature, from which emerges a selection of features distinguished by their elevated chi-square statistic values, thereby attesting to their paramount significance in our analytical pursuit. Herein follows a succinct pseudocode that meticulously delineates the steps entailed in the application of the Chi2 method.

Algorithm 1: Chi-Square feature-selection

Step 1: Define the target variable (Y) and the feature variables (X1, X2, ..., Xn).

Step 2: Initialize an empty list to store selected features "featureList".

Step 3: For each feature variable Xi:

a) Create a contingency table between Xi and Y.

b) Calculate the expected and observed frequencies for each cell in the table.

c) Compute the chi-square statistic using the formula.

d) Compare the computed chi-square value with a predefined threshold.

e) If the chi-square value exceeds the threshold, add Xi to “featureList”.

Step 4: End loop.

Step 5: Return the list of selected features “featureList”.

This pseudocode encapsulates a systematic workflow, commencing with the definition of the target variable (Y) and the feature variables (X1, X2, ..., Xn). Subsequently, it establishes a container, “featureList” to host the chosen features. Iteratively, for each feature (Xi), the algorithm conducts a comprehensive analysis. It entails the construction of a contingency table contrasting Xi and Y, followed by the computation of expected and observed frequencies within the table. The chi-square statistic is then calculated, serving as a key measure in this process. This computed value is subjected to a thresholding mechanism, scrutinizing whether it surpasses a predefined “Chi2Threshold”. Should a feature's chi-square value exceed the stipulated threshold, it is deemed as sufficiently informative and is consequently appended to the roster of “featureList”. This iterative process ensures that only features with a demonstrable influence on the classification problem are retained. Ultimately, the algorithm concludes in the delivery of the final set of selected features, a witness to the diligence and rigor that underpins our research methodology. This facet, grounded in statistical insight, augments the robustness and efficacy of our classification framework, charting a course towards precise and insightful prognostications in cardiovascular health assessment.

- *Feature selection using mRMR (“Minimum-Redundancy-Maximum-Relevance”) method:* This works on the feature’s relevance and mutual information. It picks features that have high relevance with output class, means high mutual information with target, this is termed as maximum relevance. But with each feature they check for lower redundancy, means share low mutual information too. This is termed as minimum redundancy. It undertakes the computation of both relevance and redundancy scores for

every feature in contention. These scores serve as quantitative measures, delineating the feature's degree of relevance to the output class and its potential for redundancy, respectively. The pivotal aspect of this process is the discernment of an optimal set of features, one that maximizes relevance while concurrently minimizing redundancy. This nuanced selection process ensures that the chosen features not only possess a significant bearing on the output class but also do so without introducing superfluous or overlapping information. Furthermore, it's worth noting that the feature indices, which are essentially numerical identifiers, undergo a transformation into tangible, real-world features. In essence, the mRMR method represents a discerning approach to feature selection, underpinned by a sophisticated understanding of relevance and redundancy. By adhering to the principles of maximum relevance and minimum redundancy, this methodology stands as a linchpin in our pursuit of a refined and discriminating feature set for our classification framework. Through this, we aim to distil the most pertinent aspects of the data, furnishing our model with the capacity to deliver precise and insightful projections in the realm of cardiovascular health. Flow diagram for feature selection has been demonstrated in f*igure 5.2*.

Algorithm 2: mRMR feature-selection

Step 1: Compute the mutual information for each feature in relation to the target variable.

Step 2: Compute the redundancy between each pair of features using a “distance metric” i.e., “Euclidean distance”.

Step 3: Subtract the redundancy score from the relevance score to get the mRMR score.

Step 3: Determine the priority of dataset attributes by evaluating the mutual information they share with the output class and other attributes.

Step 4: Output: Attribute indices with higher mRMR score are picked as final feature set.

Below is the pseudo code explanation for the mRMR method:

Pseudo code for mRMR

Input:

- “data”: The dataset containing features and the target variable.

- “num_features”: The desired features to select.

Output:

- “selected_features”: The set of selected features.

Step 1. Initialization:

- Initialize an empty set “selected_features”.

- Initialize an empty set “remaining_features”.

Step 2. Initialize Remaining Features:

- For each feature in `data`, add it to the set “remaining_features”.

Step 3. Feature Selection Loop:

- For i from 1 to “num_features”:

a. Initialize “best_feature” to None.

b. Initialize “best_relevance_score” to infinity.

c. Initialize “best_redundancy_score” to infinity.

d. Calculate Scores:

- For each feature in “remaining_features”:

i. Calculate the relevance score between the feature and the target variable.

ii. Calculate the redundancy score between the feature and features in “selected_features”.

- If the feature's relevance score is greater than “best_relevance_score” and its redundancy score is less than “best_redundancy_score”:

- Set “best_feature” to the current feature.

- Set “best_relevance_score” to the relevance score.

- Set “best_redundancy_score” to the redundancy score.

e. Feature Selection:

- Add “best_feature” to “selected_features”.

- Remove “best_feature” from “remaining_features”.

Step 4. Return Selected Features:

- Return the set of “selected_features”.

The mRMR feature selection algorithm iteratively selects features based on their relevance to the target variable and their redundancy with respect to the already selected features. It maintains two sets: “selected_features” for the chosen features and “remaining_features” for the features yet to be considered. In each iteration, the algorithm identifies the feature with the highest relevance and lowest redundancy scores, ensuring a balance between informativeness and redundancy. The process is repeated until the desired number of features (“num_features”) is selected.

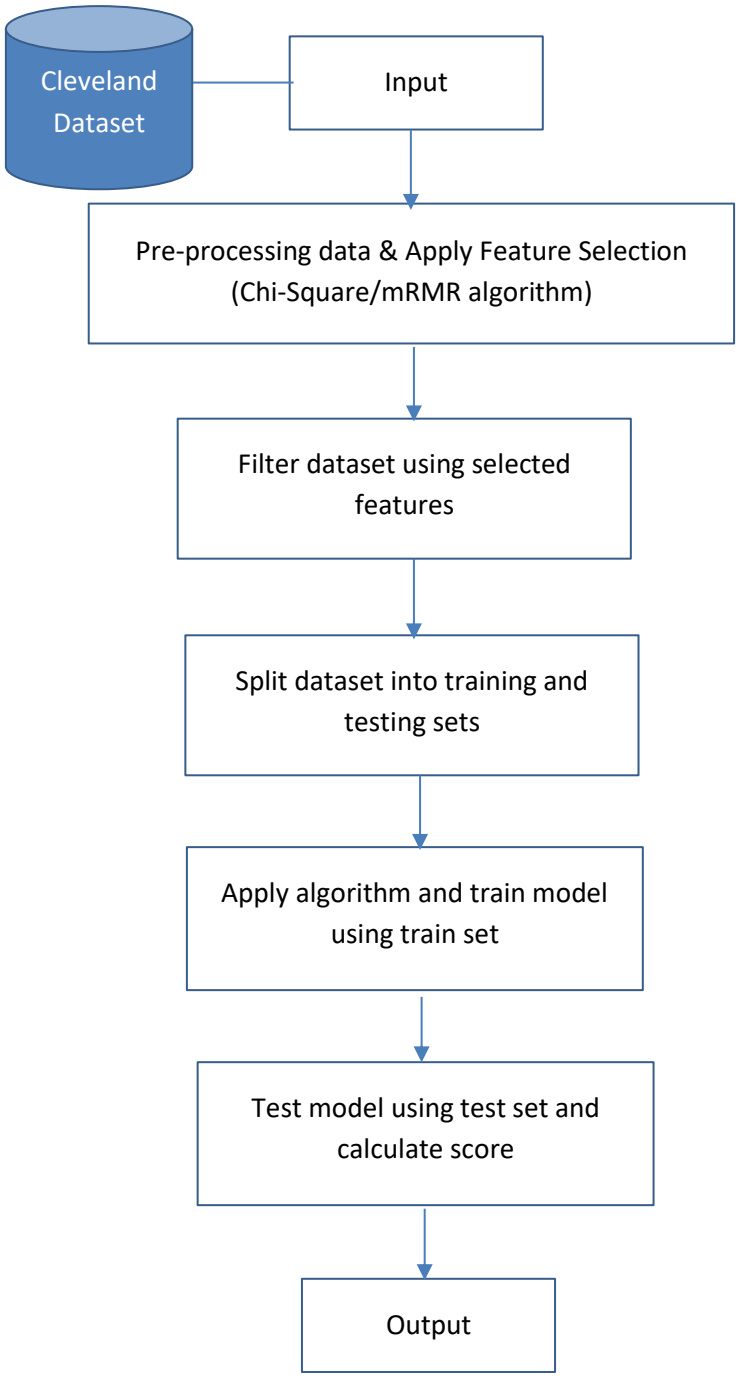


Figure 5.2: Flow chart showing model training applying feature selection

Figure 5.2 shows the flow chart showing model training incorporating feature selection. Experiment includes following stages - importing data, data pre-processing including cleaning, transformation and reduction of data, dataset splitting, model training and testing and result evaluation. Cleveland dataset is used in the experiment having 76 attributes which undergone through the feature selection and 16 attributes were taken

out (listed in section 4.2, chapter 4). Above flow has been repeated for all the ML algorithms used in the research work. Adding feature selection methods in model training improves the efficiency and also helps in reducing training cost. Experiment has been performed and results were noted. Results has been noted for all algorithms and analysed for finding the efficiency of model, we will discuss the result in later part of this chapter.

5.5.1.2 Experiment applying cross validation

After completing the experiment discussed in previous sub section, we came up with new scenarios wherein we added cross validation to the experiments. Important part of using model validation technique is that it gives the confidence that the trained model is of high efficiency and model will work well with unseen data [75-77]. Though there are many types of model validation like Hold-out, K-fold, Stratified K-Fold and Leave-One-Out (LOOCV) but in this experiment, LOOCV and K-fold CV has been used. Below we have briefly discussed both the approaches one by one.

- *Leave One Out (LOO) Cross-Validation:* To compute the viability of model, we started with "*LOO cross-validation*" technique for training model. In LOO technique n models are created using n sample where each model uses n-1 data size for training purpose. This way it consumes lot of cost to train model i.e., for just a small 300 sample size dataset, there will be 300 different temporary dataset created consisting of 299 (n-1) samples that will be used to train 300 models [77]. Below algorithm shows the implementation of LOO CV but due to high cost, this technique has been revoked and not used anywhere in the experiment.

Algorithm 3: Pseudo code for applying LOO CV technique for training model with classical methods

---

Step 1: Select priority features using feature selection technique.

Step 2: Apply LOO cross validation technique and create n model and n dataset removing current patient data from dataset and create test sample with removed record.

Step 3: Model training: Apply classification algorithm and train/test the model using training/testing datasets respectively.

Step 4: Result: Compute performance parameters i.e., accuracy, sensitivity, specificity

Step 5: Repeat step 3 to 6 with different feature set and compare the result till the best result show up.

Step 6: Output: Test model and output the result.

- *K-Fold Cross-Validation:* This is another CV technique used to train and validate model. It serves as a robustness check for the model. The process involves partitioning the dataset into k subsets of equal size. The model is then trained on "k-1" of these subsets and validated on the remaining one. This procedure is iterated k times, ensuring that each subset serves as both training and validation data at some point. Performance is calculated and features are selected based on n highest average score. We used CV in the experiment to validate the biasness of data. In algorithm 4, steps have been recorded for application of K-fold CV.

Algorithm 4: Pseudo code for applying K-Fold Cross-Validation technique for training model with classical methods

---

Step 1: Input data [24].

Step 2: Apply data pre-processing and refine the data.

Step 3: Apply feature selection using Chi-Square technique.

Step 4: Apply “K-Fold cross-validation” technique and split dataset into k-folds.

Step 5: Model training: Apply classification algorithm and train the model using selected features on each fold. For each fold in the dataset:

a) Designate the current fold as the validation set.

b) Utilize the remaining 'k-1' folds for training the model.

c) Evaluate the model's performance on the validation set.

Step 6: Record the performance metric for each iteration.

Step 7: Calculate the average performance score based on the recorded metrics.

Step 8: Identify the top 'n' features based on their average performance scores across all folds.

Step 9: Output: Test model and output the result.

Below *figure 5.3* shows the schematic representation of experiment work flow incorporating cross validation and feature selection as a part of experiment.

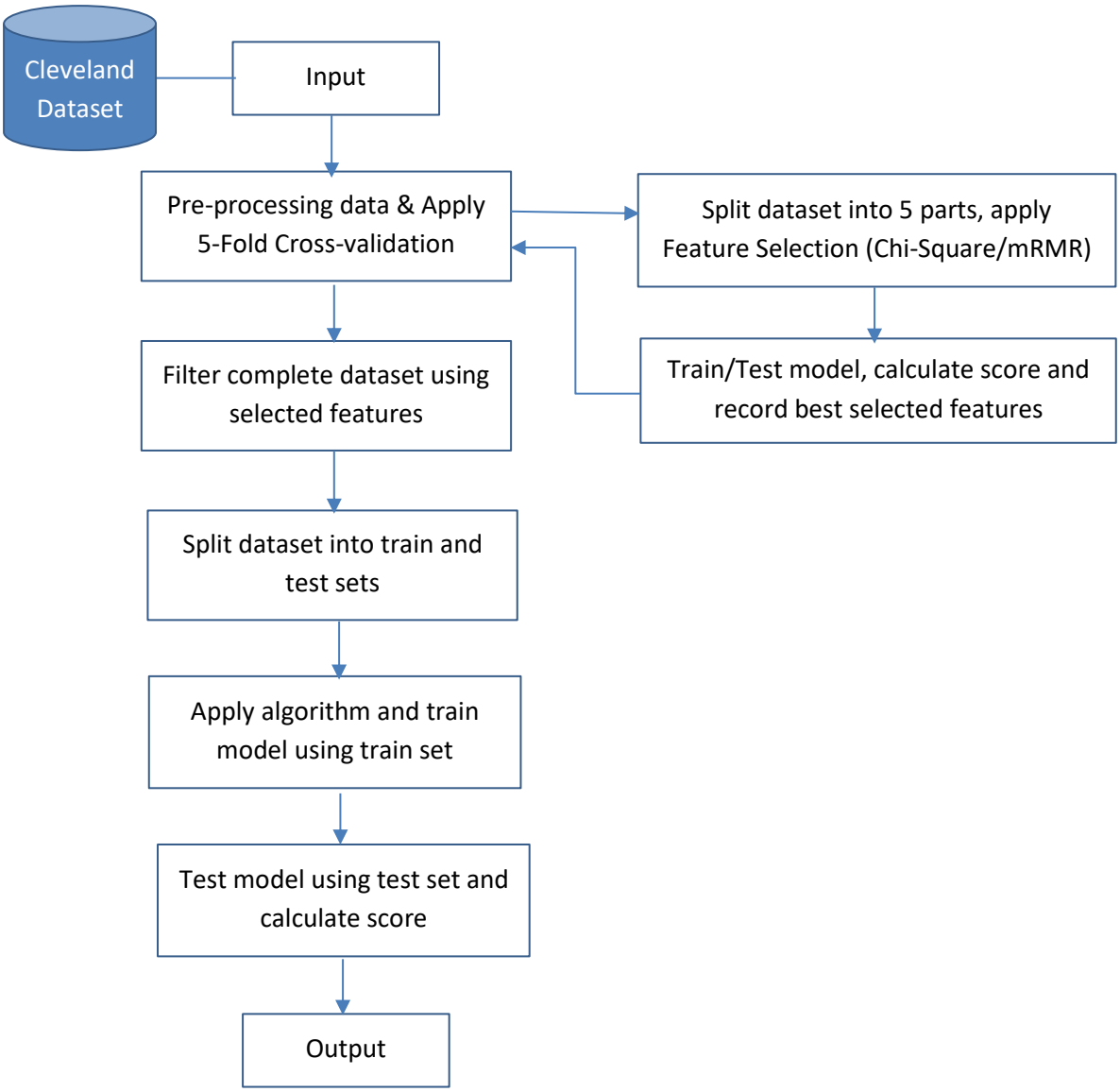


Figure 5.3: Flow chart showing model training using cross validation

5.5.1.3 Experiment with GANN

A Hybrid framework has been created which comprises of "genetic algorithm and neural network or GANN" in that GA is used to evolve the NN structure and parameters by selecting the best performing individuals from a population of randomly initialized networks and applying genetic operations "mutation and crossover" to produce new offspring. The fitness of each individual is determined by evaluating its performance on a given task or dataset. To crystallize the optimal set of features, we've enlisted the aid of the minimum Redundancy Maximum Relevance (mRMR) algorithm. It plays a pivotal role in enhancing the efficacy of GANN by selecting the most discriminative

features. Subsequently, the model undergoes an intensive training phase. Below is the pseudo code which shows the steps involved in the model training with “GANN and mRMR”.

Algorithm 5: GANN and mRMR feature-selection

Step 1: Input dataset for classification [24].

Step 2: Process input data and prepare for use in GANN. Include data-cleaning, apply feature selection.

Step 3: Create model using inputs, hidden and output neurons.

Step 4: Initialize population, fill chromosomes with randomly generated uniform values.

Step 5: Use activation methods and compute weight matrix for population.

Step 6: Apply fitness function and identity the higher chance of chromosome to being selected as parents for the next generation.

Step 7: Choose n most fit chromosomes. Apply crossover to generate new population.

Step 8: The newly created chromosomes may also undergo mutation, introducing additional genetic diversity into the population.

Step 9: Train the mutated chromosome and evaluate fitness on testing dataset. Get the best model and return. Continue from step 5 if generation is left for processing.

Step 10: The genetic algorithm stops when a termination criterion is reached, either maximum number of generations completes, or a satisfactory level of performance is achieved.

Step 11: Select the best neural network architecture as a solution.

GANN consists of several phases starting with initial population generation, evaluation and finding fitness score, selecting best population and applying mutation to generate

new population and finally getting the optimised solution for the problem. These stages are discussed below:

- *Initial phase:* First generation initialized with randomly generated population with set of parameters. The network architecture is defined with 416 neurons, comprising of 2 hidden layers and 1 output layer. Activation methods such as ReLU and Sigmoid are applied, and the optimization is performed using the gradient descent method.
- *Evaluation phase:* In this phase, every chromosome produced in the initial stage undergoes training, and its accuracy is assessed to determine its fitness score.
- *Selection phase:* The best neural network architecture found during the genetic algorithm process is selected as the final solution for classification. This selection is based on the fitness scores, with the network demonstrating the highest fitness being chosen while others are disregarded.
- *Cross-over phase:* The chosen parent undergoes mating process or crossover to generate new offspring.
- *Mutation phase:* Following the crossover, a few genes are randomly mutated to ensure diversity within the population. Below flow diagram shows the working of GANN.

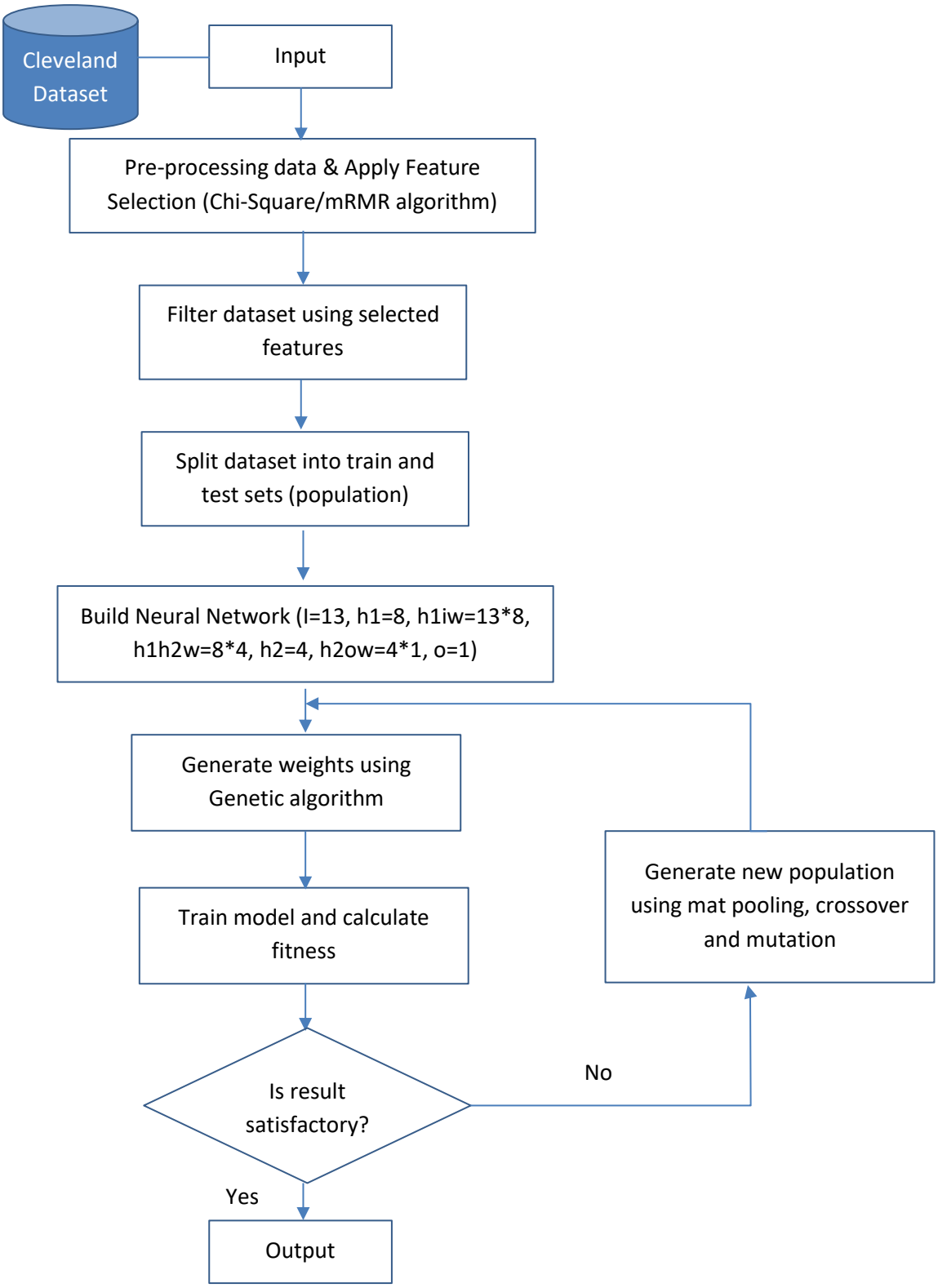


Figure 5.4: Experiment flow for GANN

5.5.2 Experimental scenarios for finding the probability of heart attack

After the neural network is prepared, the second stage of the experiment involves using a fuzzy inference system. In this stage, the input from the neural system undergoes fuzzification and defuzzification processes through the inference engine. Experiments has been performed with different set of fuzzy rules and set which derived different membership functions for the dataset. In below sub section this has been discussed in detail.

5.5.2.1 Medical parameters used in experiment

Detailed discussion on the dataset has already been gone through in chapter 4 and all the features have been explained in depth that are used in the research work. Based on the domain knowledge, certain medical parameters have been identified as high priority for predicting the probability of heart attack, *table 5.3* presents these priority medical parameters from the dataset.

Table 5.3: Priority medical parameters used in the experiment

Resting Blood Pressure	Resting ECG Result	Cigarette Per Day
Serum Cholesterol	Resting Heart Rate	Heart Disease Family History
Fasting Blood Sugar	Max Heart Rate	Is Heart Patient

The medical parameters enlisted in *table 5.3* are of utmost importance for accurately predicting the probability of a heart attack. Resting Blood Pressure and Resting ECG Result are essential indicators of cardiac health, while Cigarette Per Day serves as a significant risk factor that directly impacts the overall cardiovascular well-being. Serum Cholesterol and Fasting Blood Sugar levels are also key factors to consider, as they put direct impact on heart's health. The resting and maximum heart rate, being vital indicators of heart function, aid in identifying any abnormalities in heart rhythm. Family history of heart disease is an important parameter to consider as it indicates a potential genetic predisposition to heart conditions. Lastly, the parameter “IsHeartPatient” can help identify if the patient is already a heart patient or not. All

these parameters are instrumental in predicting the probability of a heart attack and can facilitate early detection and prevention of heart-related ailments, thus emphasizing the need to consider them during diagnosis and treatment.

5.5.2.2 Rule Creation

In the initial phase of the experiment, after input injection, the crucial task is to establish rules that will govern the subsequent stages. This rule-setting process involves a meticulous analysis of all the incorporated features. Each feature is scrutinized to determine its priority based on its potential impact on heart health. Noteworthy features include factors like "resting heart rate, daily cigarette consumption, blood sugar levels, and family history of heart disease." Once a comprehensive understanding of these variables is obtained, a variety of approaches are employed to formulate rules. The primary objective is to ensure that these rules possess the requisite potency to support the system effectively. Furthermore, they must empower the system to determine the probability of a heart attack occurring in an individual. It's important to note that while the number of antecedents can be more than one, consequents remain singular. Rule creation for this experiment is achieved through the integration of two distinct approaches.

The experiment's rule creation process was undertaken through two distinct approaches –

- The foundation of this study lies in the Cleveland dataset, which serves as the bedrock for our analysis. After undergoing a rigorous process of data processing, we extracted 282 high-quality data points, while 21 were deemed unfit for inclusion out of the initial pool of 303 data entries. In the pursuit of understanding and predicting risk levels, a comprehensive set of 126 rules was meticulously generated. These rules are rooted in the "num" feature, which delineates the risk categories into four distinct levels, denoted by values 1, 2, 3, and 4. Each of these rules embodies a unique combination of features, enabling the system to effectively assess and categorize the risk levels associated with each patient.

- The initial phase of our study has been marked by an in-depth exploration of various features, each offering a unique insight into the potential risks associated with heart health, for example, the maximum heart rate emerges as a critical indicator, its significance lying in the fact that an elevated heart rate can substantially impact cardiac well-being. Consequently, we have stratified the contribution of maximum heart rate into four distinct tiers: normal, medium, high, and critical. This stratification system empowers our model to finely discern and categorize the varying degrees of risk associated with different heart rates. Similarly, the blood sugar levels do play a role in cardiac health, our analysis has revealed that its impact is not as immediate as that of heart rate. We have established two thresholds for blood sugar - normal and medium. This nuanced approach allows us to accurately assess its influence on an individual's risk profile. In tandem, we have evaluated the contributions of other features as well, each offering a unique vantage point in the assessment of heart attack risk. Through this comprehensive analysis, we aim to provide a holistic understanding of a person's susceptibility to a heart attack.

Our work has culminated in the generation of an extensive set of 13,824 rules, carefully tailored to account for both normal and high-risk scenarios. These rules serve as the cornerstone of our predictive model, providing a robust framework through which we can assess the likelihood of a heart attack. As an illustrative reference, a selection of these inference rules has been outlined in Table 5.4, offering a glimpse into the complexity and precision of our risk assessment system.

Table 5.4: Fuzzy rules showing antecedents and consequent

Condition	**Attack Probability**
"BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Good AND Max Heart Rate IS Normal AND Cigarette Per Day IS Normal AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No"	Normal
"BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Good AND Max Heart Rate IS Normal AND Cigarette Per Day IS Moderate AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS No"	Normal
"BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Good AND Max Heart Rate IS Moderate AND Cigarette Per Day IS Moderate AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No"	Normal
"BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Good AND Max Heart Rate IS High AND Cigarette Per Day IS Moderate AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS No"	Normal

“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Good AND Max Heart Rate IS High AND Cigarette Per Day IS Moderate AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS Yes”	Moderate-Risk
“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Good AND Max Heart Rate IS High AND Cigarette Per Day IS High AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS Yes”	Moderate-Risk
“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Good AND Max Heart Rate IS Critical AND Cigarette Per Day IS Normal AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No”	Normal
“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Average AND Max Heart Rate IS High AND Cigarette Per Day IS High AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS Yes”	Moderate-Risk
“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS Normal AND ECG IS Normal AND Resting Heart Rate IS Average AND Max Heart Rate IS High AND Cigarette Per Day IS High AND	Moderate-Risk

HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS No”	
“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS High AND ECG IS Risk1 AND Resting Heart Rate IS High AND Max Heart Rate IS High AND Cigarette Per Day IS High AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No”	Moderate-Risk
“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS High AND ECG IS Risk1 AND Resting Heart Rate IS High AND Max Heart Rate IS Critical AND Cigarette Per Day IS Moderate AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS Yes”	High-Risk
“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS High AND ECG IS Risk2 AND Resting Heart Rate IS Good AND Max Heart Rate IS Normal AND Cigarette Per Day IS Normal AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS Yes”	Moderate-Risk
“BP IS Ideal AND Cholesterol IS Normal AND Blood Sugar IS High AND ECG IS Risk2 AND Resting Heart Rate IS Good AND Max Heart Rate IS Moderate AND Cigarette Per Day IS High AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS Yes”	Moderate-Risk

“BP IS Moderate AND Cholesterol IS Ideal AND Blood Sugar IS High AND ECG IS Risk2 AND Resting Heart Rate IS Average AND Max Heart Rate IS Critical AND Cigarette Per Day IS Ideal AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No”	Moderate-Risk
“BP IS Moderate AND Cholesterol IS Ideal AND Blood Sugar IS High AND ECG IS Risk2 AND Resting Heart Rate IS Average AND Max Heart Rate IS Critical AND Cigarette Per Day IS Ideal AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS Yes”	High-Risk
“BP IS Moderate AND Cholesterol IS Moderate AND Blood Sugar IS Ideal AND ECG IS Ideal AND Resting Heart Rate IS Good AND Max Heart Rate IS High AND Cigarette Per Day IS Ideal AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS Yes”	Moderate-Risk
“BP IS Moderate AND Cholesterol IS Moderate AND Blood Sugar IS Ideal AND ECG IS Ideal AND Resting Heart Rate IS Good AND Max Heart Rate IS High AND Cigarette Per Day IS Moderate AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No”	Normal
“BP IS Moderate AND Cholesterol IS Moderate AND Blood Sugar IS Ideal AND ECG IS Ideal AND Resting Heart Rate IS Good AND Max Heart Rate IS High AND Cigarette Per Day IS Moderate AND	Moderate-Risk

HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS Yes”	
“BP IS High AND Cholesterol IS Moderate AND Blood Sugar IS Ideal AND ECG IS Ideal AND Resting Heart Rate IS Good AND Max Heart Rate IS Ideal AND Cigarette Per Day IS High AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS Yes”	Moderate-Risk
“BP IS High AND Cholesterol IS Moderate AND Blood Sugar IS Ideal AND ECG IS Ideal AND Resting Heart Rate IS Good AND Max Heart Rate IS Moderate AND Cigarette Per Day IS Ideal AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No”	Normal
“BP IS High AND Cholesterol IS Moderate AND Blood Sugar IS Ideal AND ECG IS Ideal AND Resting Heart Rate IS Good AND Max Heart Rate IS Critical AND Cigarette Per Day IS High AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS Yes”	High-Risk
“BP IS High AND Cholesterol IS Moderate AND Blood Sugar IS Ideal AND ECG IS Ideal AND Resting Heart Rate IS High AND Max Heart Rate IS Ideal AND Cigarette Per Day IS Ideal AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No”	Normal

“BP IS Critical AND Cholesterol IS Critical AND Blood Sugar IS High AND ECG IS Ideal AND Resting Heart Rate IS High AND Max Heart Rate IS High AND Cigarette Per Day IS High AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS No”	High-Risk
“BP IS Critical AND Cholesterol IS Critical AND Blood Sugar IS High AND ECG IS Ideal AND Resting Heart Rate IS High AND Max Heart Rate IS High AND Cigarette Per Day IS High AND HeartDiseaseFamilyHistory IS Yes AND IsHeartPatient IS Yes”	Critical-Risk
“BP IS Critical AND Cholesterol IS Critical AND Blood Sugar IS High AND ECG IS Risk1 AND Resting Heart Rate IS Good AND Max Heart Rate IS Ideal AND Cigarette Per Day IS Moderate AND HeartDiseaseFamilyHistory IS No AND IsHeartPatient IS No”	Moderate-Risk

5.5.2.3 Fuzzy sets

This can be treated as major pillar for any fuzzy inference system and for a single feature in dataset there could be multiple fuzzy sets defined. In the experiment, we defined more than one fuzzy set for each dataset feature.

5.5.2.4 Membership functions

A membership function $\mu_A(x)$ enable us to present a fuzzy set graphically. Different types of MFs such as triangular, trapezoidal and gaussian have been used in the experiments. For all medical parameters employed in the system, MFs have been developed to create fuzzy sets. The graphical representation of the fuzzy set has been

achieved through the application of triangular, trapezoidal and gaussian functions in this study.

- *Triangular membership function:* The function of the curve is determined by the vector x and is influenced by the scalar parameters p, q, and r. where p ≤ q ≤ r. The triangle has a base of length (r-p) and a height of 1. Equation 5.6 shows the same. The triangular membership function is frequently utilized in scenarios where the limits of a fuzzy set are vague, and the level of membership slowly rises and decreases across the set range. It is also simple to execute and interpret, making it a preferred option in numerous applications.

$$f(x;p,q,r) = \begin{Bmatrix} 0, & x \le 0 \\ \frac{x-p}{q-p}, & p \le x \le q \\ \frac{r-x}{r-b}, & q \le x \le r \\ 0, & r \le x \end{Bmatrix} \quad (5.6)$$

- *Trapezoidal function:* The function is often used when the boundaries of a fuzzy set are not well-defined and the degree of membership gradually increases and decreases over the range of the set. This function has four parameters that determine the shape of the trapezoid: p, q, r and s wherein the parameter p represents the lower boundary of the fuzzy set, q represents the point where the degree of membership reaches the maximum value of 1, r represents the point where the degree of membership starts to decrease, and s represents the upper boundary of the fuzzy set. The vector function depends on the four scalar parameters p, q, r and s where p and s illustrate the lower bottom area of the curve and q and r symbols top one, equation 5.7 represent the same.

$$f(x;p,q,r,s) = \begin{Bmatrix} 0, & x \le 0 \\ \frac{x-p}{q-p}, & p \le x \le q \\ 1, & q \le x \le r \\ \frac{r-x}{r-q}, & r \le x \le s \\ 0, & s \le x \end{Bmatrix} \quad (5.7)$$

- *Gaussian function:* This is a bell-shaped curve that is commonly used in fuzzy logic systems. It is a symmetric function that is defined by a mean value and a standard deviation. The function is continuous and non-negative for all real numbers. The Gaussian function is a popular choice for fuzzy systems because it allows for a smooth transition between membership values. As the input value moves away from the mean value, the membership value decreases in a smooth and continuous manner. MF parameters, specified as the vector [σ c], equation 5.8 represents the gaussian function where σ is the standard deviation and c is the mean.

$$f(x;\sigma,C) = e^{-\frac{(x-C)^2}{2\sigma^2}} \qquad (5.8)$$

Below, graph for membership functions for all the selected medical parameters has been plotted wherein fuzzy sets along with the type of membership functions can be overviewed.

- *Resting BP:* In the realm of fuzzy logic, the input parameters are stratified into four distinct fuzzy sets "normal, moderate, high, and critical". Each of these sets is defined by specific ranges, elucidated in detail in Figures 5.5. For the medium and high sets, their membership functions are graphically depicted as trapezoidal shapes. These contours encapsulate the range of values associated with these categories. Conversely, the normal and critical sets are characterized by triangular shapes, again visually capturing the span of values attributed to each of these fuzzy categories. *Figure 5.6* shows the gaussian representation of the same.

Figure 5.5: MF for RBP

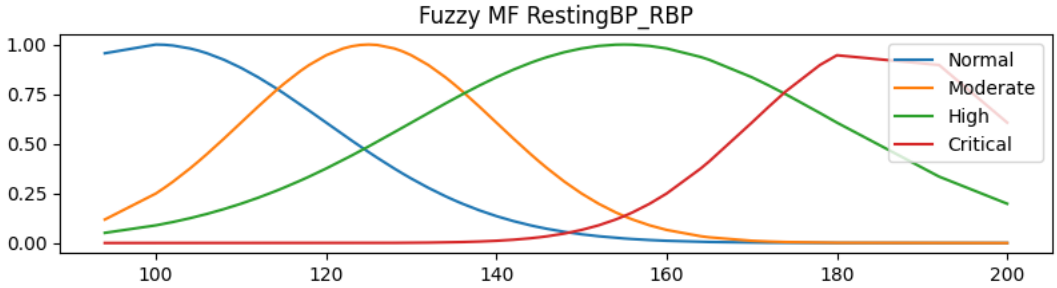


Figure 5.6: Gaussian MF showing input feature RBP

- *Cholesterol:* The input field has been partitioned into four fuzzy sets, namely "normal, moderate, high, and critical", with their corresponding ranges displayed in Figure 5.7. The MFs have been defined as a combination of "trapezoidal" and "triangular" functions. *Figure 5.8* represent the "gaussian" MF.

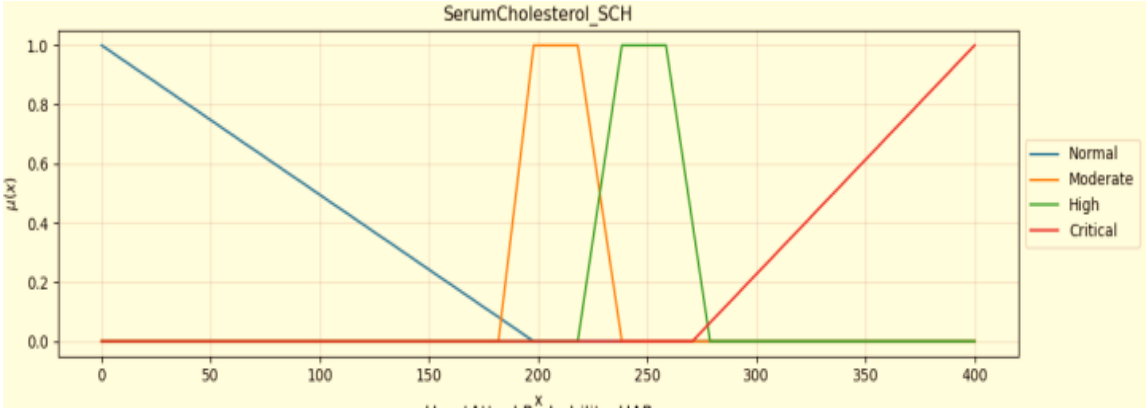


Figure 5.7: MF for SCH

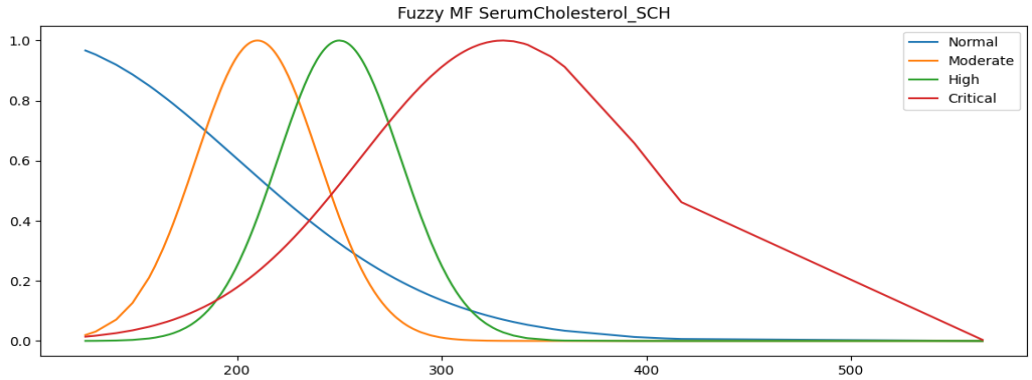


Figure 5.8: Gaussian representation for feature SCH

- *Sugar level (fasting):* Input is divides into 2 fuzzy sets "normal and high" with "triangular" MF, overviewed in *figure 5.9*.

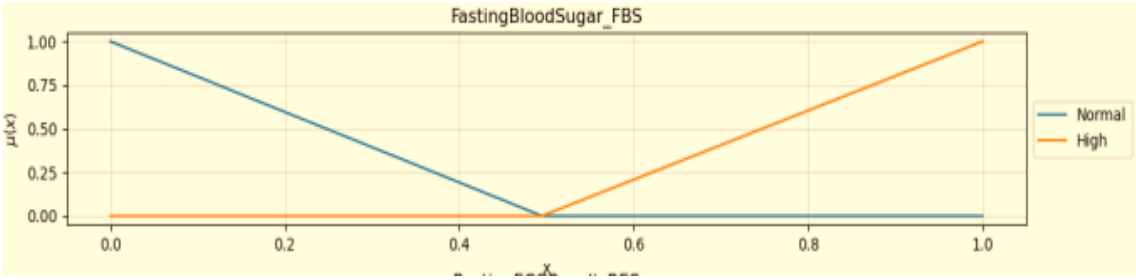


Figure 5.9: Membership function graph for input feature FBS

- *Resting ECG Result:* Input has fuzzy sets "normal, risk1 and risk2" having "triangular" MF, has been plotted in *figure 5.10*.

Figure 5.10: MF for RES

- *Resting HR:* Input covers in fuzzy sets "good, average and high" having "triangular" MF, is shown in *figure 5.11*.

Figure 5.11: MFs for input feature RHR

- *Max HR:* Input is covered in 4 fuzzy sets "normal, moderate, high and critical", their range are presented below. MF is defined as "triangular", can be seen in *figure 5.12* along with the "gaussian" representation shown in *figure 5.13*.

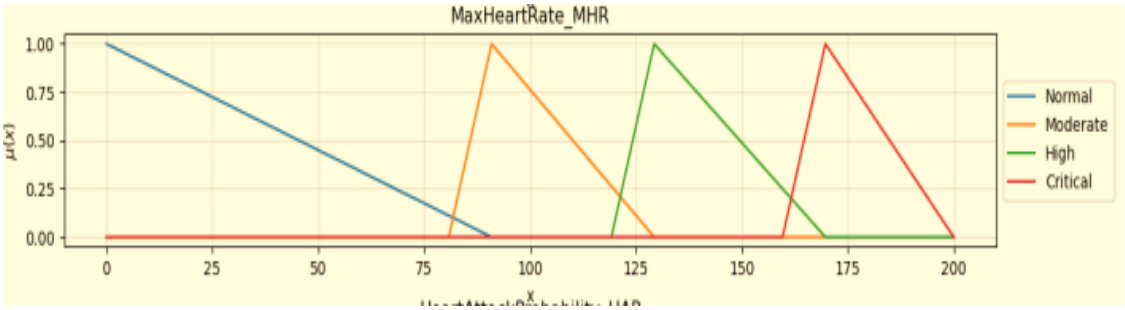


Figure 5.12: MF for input MHR

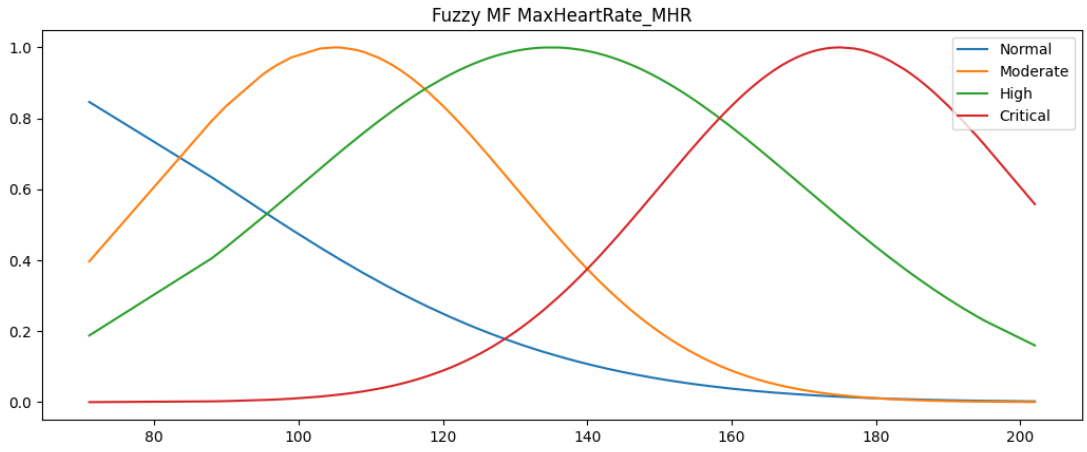


Figure 5.13: Gaussian MF for MHR

- *CPD:* Input has 2 fuzzy sets "normal and high" covered as "triangular" MF. Range is depicted in *figure 5.14* and *figure 5.15*.

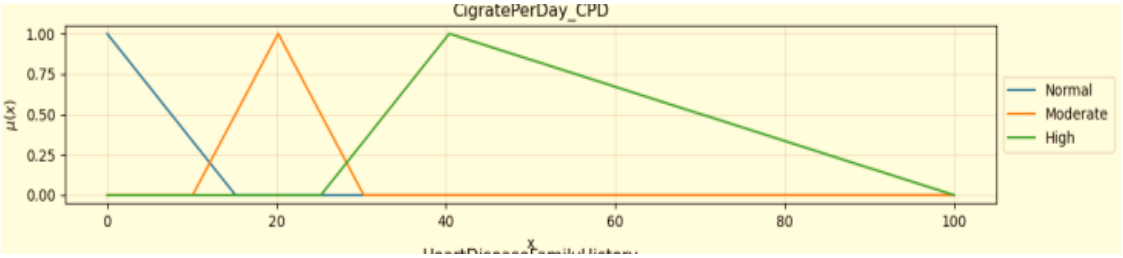


Figure 5.14: MFs for input CPD

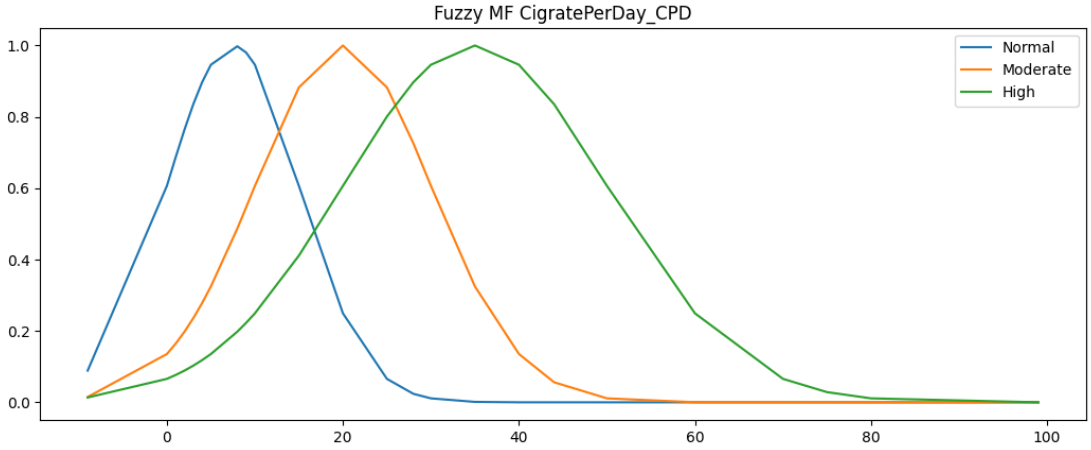


Figure 5.15: Gaussian MF for CPD

- *Heart Disease Family History:* Input has binary fuzzy sets "yes and no" with "triangular" MF. Range is depicted in *figure 5.16.*

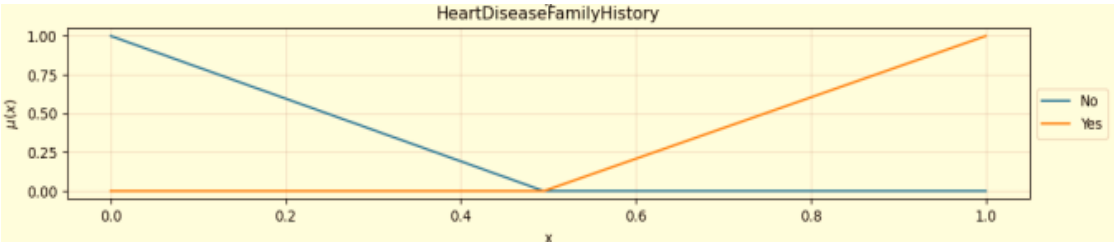


Figure 5.16: MFs for HeartDiseaseFamiliyHistory

- *Is Heart Patient:* Input has binary fuzzy sets "yes and no" with "triangular" MF. Figure 5.17 shows the range for the same.

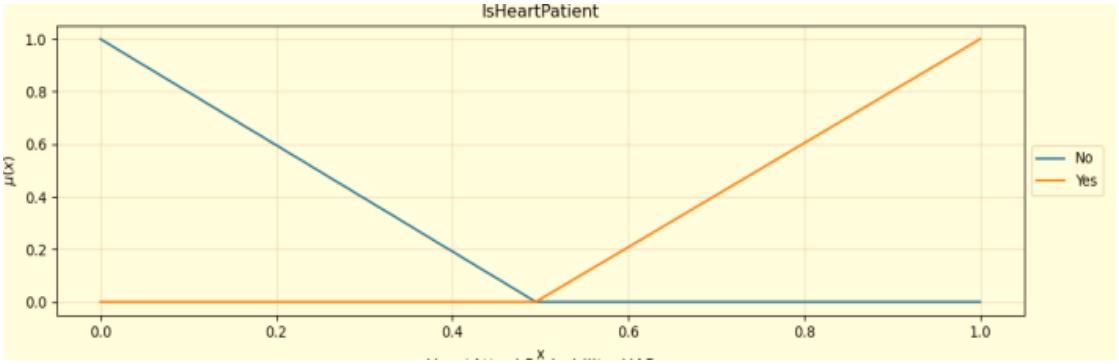


Figure 5.17: MFs graph for input feature IsHeartPatient

- *Heart Attack Probability:* The output param has 4 fuzzy sets – "normal, moderate, high and critical", range shown in *figure 5.18*. Membership function is specified as "triangular".

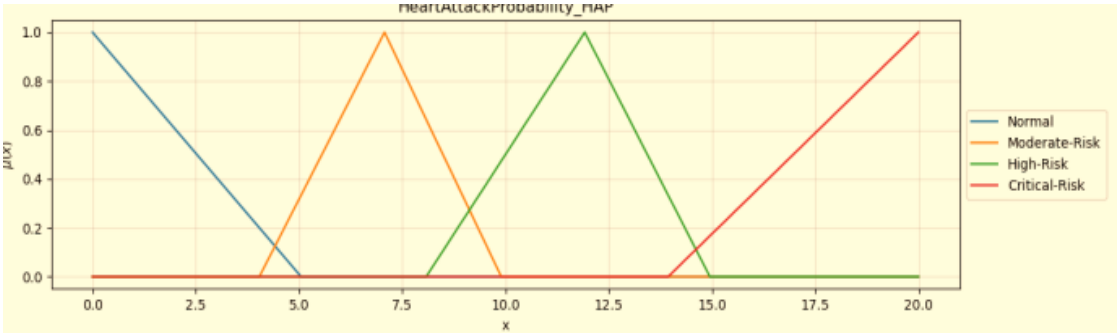


Figure 5.18: MFs for HAP

5.5.2.5 Fuzzification

The subsequent phase in research work entails the process of fuzzification, a pivotal step wherein the precisely normalized input values are transfigured into fuzzy representations. This intricate transformation is orchestrated by leveraging the rich reservoir of knowledge encapsulated within the consecrated areas of available knowledge base, which serves as the bastion of fuzzy logic principles. This transformation is achieved through a series of mathematical operations that assign each crisp input value a degree of membership to each fuzzy set defined within the system. Through the deft application of fuzzy logic paradigms, each input value is filled with a nuanced association with the constituent fuzzy sets, thereby stimulating a

comprehensive understanding of its contextual relevance and semantic associations within the domain of discourse. The fuzzification process exemplifies the essence of computational intelligence, seamlessly bridging the chasm between crisp numerical inputs and linguistically interpretable fuzzy representations. For example, consider a crisp input value representing the blood pressure of a patient. Through fuzzification, this crisp value would be assigned a degree of membership to each of the fuzzy sets defined for blood pressure, such as "low," "normal," "high," and "critical." The degree of membership indicates the extent to which the crisp input value belongs to each fuzzy set, based on its proximity to the boundaries defined by the membership functions. Through its adept manipulation of linguistic variables and membership functions, fuzzification infuses the framework with the requisite adaptability and resilience to navigate the complex workings of real-world data.

5.5.2.6 Inference engine

At this critical juncture of the experiment, we reach the pinnacle of our analytical journey, where exhaustive computations and meticulous formulae converge to yield meaningful conclusions. This stage heralds the application of the Mamdani system, a fundamental component of our research framework. Within the Mamdani paradigm, we deploy the max-min method, an established technique revered for its efficacy in generating decisive outputs.

> *"Max-Min" inference method:* In the Max-Min inference method, the determination of the rule connector, whether it's "and" or "or," dictates the course of action. If the connector is "and," the minimum value is selected, while if it's "or," the maximum value is chosen. This process is crucial as it influences how the input variables interact to generate the output. Following this selection, the next step involves truncating the triangular shape of the fuzzy set and retaining only the trapezoidal portion. This truncation process is essential for refining the fuzzy output and ensuring that it accurately reflects the underlying membership function. By focusing on the most relevant segment of the fuzzy set, the Max-Min method enhances the precision and reliability of the inference process. Ultimately, this method plays a vital role in transforming the fuzzy

input into a crisp output, providing valuable insights for decision-making and analysis.

5.5.2.7 Defuzzification

Once the input has been fuzzified into fuzzy values and the aggregation stage is concluded, the subsequent phase involves defuzzifying the fuzzified values to obtain a crisp output. In the Mamdani system, defuzzification is a crucial step following the aggregation of fuzzy values. In this method, various defuzzification techniques can be employed to convert the aggregated fuzzy set into a crisp output. One common approach is the "centroid" method, which calculates the "center of gravity" of the fuzzy set to determine the crisp output. This method considers the area under the fuzzy set curve, with greater emphasis placed on the regions where the membership values are higher. Alternatively, the height method can be utilized, wherein the crisp output is determined based on the maximum membership value within the aggregated fuzzy set. This method prioritizes the highest membership value as the representative output, effectively capturing the most significant contribution to the final result. However, in the case of Sugeno, the need for a defuzzification method is obviated. Instead, the function output directly yields the desired output for the solution. This distinction underscores the inherent simplicity and efficiency of the Sugeno method, as it streamlines the computational process by directly generating the crisp output without the need for additional steps. By bypassing the defuzzification process, Sugeno offers a straightforward and expedient solution, making it an attractive option for certain applications where computational efficiency is paramount. The Mamdani approach is often favoured over the Sugeno approach in cases where the output of the fuzzy inference system requires a more interpretable and human-understandable form. This is because the Mamdani method produces linguistic fuzzy sets as output, which can directly convey qualitative information in natural language. Additionally, the Mamdani method is preferred when the system requires handling of uncertain or imprecise inputs, as it allows for a more flexible representation of fuzzy relationships between variables. This makes it suitable for applications where the rules and relationships governing the system are not well-defined or may be subject to change over time.

5.6 Result and Discussion

Experiment has been performed on both classical and hybrid system and result are noted. Various set of features were employed to machine learning methodologies and the outcomes were compared. Cross-validation was also employed to ensure the reliability of the results. Based on the confusion matric formula discussed in *section 5.4.1*, scores have been computed from the model outcome. The following table presents the experimental results for both the classical and proposed GANN method. Below *figure 5.19* presents the ROC curve for the given dataset.

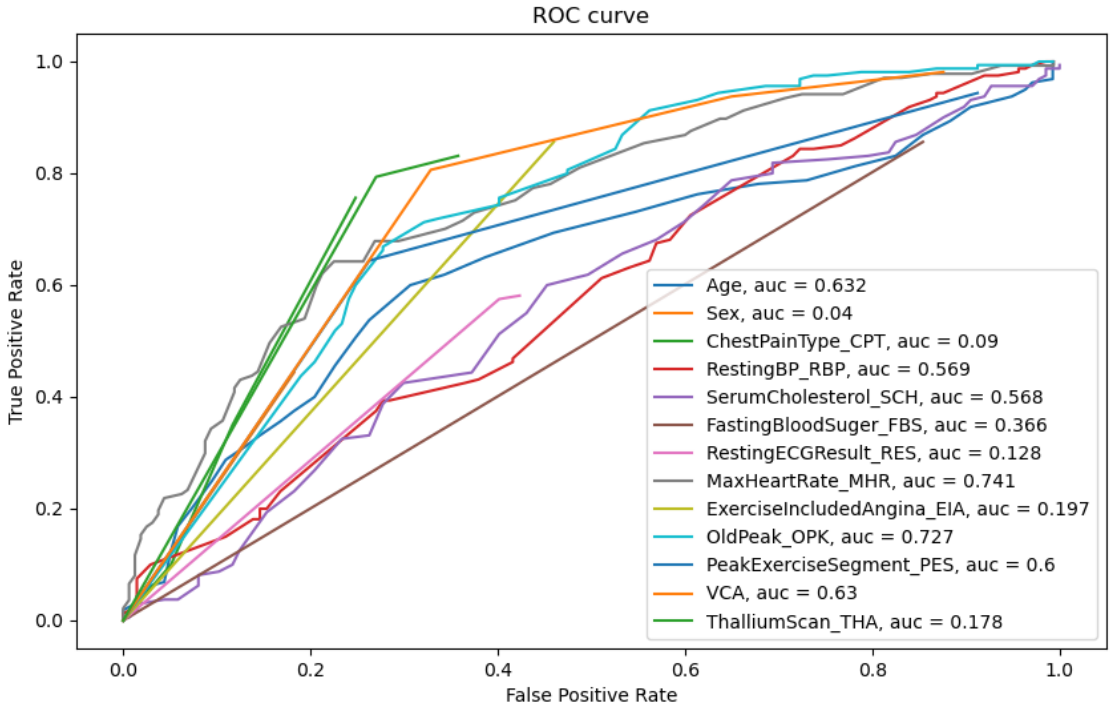


Figure 5.19: ROC curve for Claveland dataset

Below, *table 5.5* shows the prediction result for different classification methods applying Chi-Square feature selection. Here, the result clearly shows that naïve bayes outperform over other classical methods. Chi-Square feature selection is specifically designed to select features that are strongly correlated with the outcome variable and selecting only the most informative features, the logistic regression model can achieve better accuracy by focusing on the most important predictors. Both LR and NB are relatively simple and fast to train compared to other models. KNN requires computing

distances between each sample in the training set and the test set and SVM, on the other hand, can be slow to train. Here in table 5.5, it can be clearly notice that accuracy of SVM is very low, about 65% and for LR and NB it is more than 85%.

Table 5.5: Prediction result with different classification methods using Chi-Square feature-selection

ML Techniques	**Feature List**	**Performance**
Decision tree	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA	Accuracy: 75.20% Sensitivity: 86.77% Specificity: 65.33%
KNN	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA	Accuracy: 73.33% Sensitivity: 83.53% Specificity: 63.37%
Logistic Regression	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA, ExerciseIncludedAngina_EIA, Age, SerumCholesterol_SCH	Accuracy: 83.33% Sensitivity: 90.10% Specificity: 77.62%
Naïve Bayes	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA	Accuracy: 85.22% Sensitivity: 93.13% Specificity: 76.87%
Random Forest	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA, ExerciseIncludedAngina_EIA, Age	Accuracy: 81.67% Sensitivity: 86.63% Specificity: 77.67%

SVM	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA	Accuracy: 65.28% Sensitivity: 82.50% Specificity: 54.87%

Below, *table 5.6* shows the prediction result for different classification methods applying mRMR feature selection. Here also NB and LR outperform over other methods. mRMR feature selection is designed to select features that are both informative and complementary, is of maximum relevance and own minimum redundancy, which means that the selected features are not only informative on their own but also provide unique information when combined with other features. LR and NB both can take benefit from this property as they are linear models that can take advantage of the complementary nature of the selected features. In the table, it can be seen that NB give better results with feature "SCH, RestingECGResult_RES, FBS, ExerciseIncludedAngina_EIA, OldPeak_OPK, VCA, PeakExerciseSegment_PES, ThalliumScan_THA ChestPainType_CPT, Sex" whereas LR shows better result with lesser features as well "SerumCholesterol_SCH, VCA, ThalliumScan_THA, FastingBloodSuger_FBS, PES, ExerciseIncludedAngina_EIA". KNN and SVN remains at lower side in terms of performance and other models like DT or RF gives satisfactory percentage which is more than 80%.

Table 5.6: Result applying feature selection (mRMR)

Technique	**Feature list**	**Model performance**
Decision tree	ThalliumScan_THA, ChestPainType_CPT, VCA, ExerciseIncludedAngina_EIA, PES, SerumCholesterol_SCH, FastingBloodSuger_FBS, Sex	ACC: 80.0% SENS: 86.67% SPEC: 73.33%
KNN	Sex, RestingECGResult_RES, OldPeak_OPK,	ACC: 66.67% SENS: 73.33%

	ExerciseIncludedAngina_EIA, RestingBP_RBP, ThalliumScan_THA, VCA, PES, ChestPainType_CPT, MHR SCH, FastingBloodSuger_FBS, Age	SPEC: 60 %
Logistic Regression	SerumCholesterol_SCH, VCA, ThalliumScan_THA, FastingBloodSuger_FBS, PES, ExerciseIncludedAngina_EIA	Accuracy: 85.0% Sensitivity: 96.67% Specificity: 73.33%
Naïve Bayes	SCH, RestingECGResult_RES, FBS, ExerciseIncludedAngina_EIA, OldPeak_OPK, VCA, PeakExerciseSegment_PES, ThalliumScan_THA ChestPainType_CPT, Sex	Accuracy: 85.0% Sensitivity: 93.33% Specificity: 76.67%
Random Forest	SerumCholesterol_SCH, VCA, PES, FBS, ChestPainType_CPT, ExerciseIncludedAngina_EIA, ThalliumScan_THA	Accuracy: 81.67% Sensitivity: 86.67% Specificity: 76.67%
SVM	SerumCholesterol_SCH, RES, FastingBloodSuger_FBS, ChestPainType_CPT, Sex, ThalliumScan_THA, EIA, OldPeak_OPK, RestingBP_RBP, VCA,	Accuracy: 56.67% Sensitivity: 80.0% Specificity: 33.33%

	PeakExerciseSegment_PES, Age, MaxHeartRate_MHR	

Below, *table 5.7* shows the prediction result for different classification methods applying feature selection and cross validation. Here NB and LR outperform over other methods though the perform of models have improved as compared to previous one. NB give good result with features MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA, ExerciseIncludedAngina_EIA, Age, SerumCholesterol_SCH, RestingBP_RBP, ChestPainType_CPT, RestingECGResult_RES, PeakExerciseSegment_PES which is nearby 85%. Apart from this other model performance are nearby 80% only.

Table 5.7: Experiment result using 5-Fold Cross Validation, Chi-Square feature-selection

Technique	**Feature list**	**Result**
Decision tree	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA' 'ExerciseIncludedAngina_EIA, Age, SerumCholesterol_SCH, RestingBP_RBP' 'ChestPainType_CPT	Accuracy: 73.33% Sensitivity: 66.67% Specificity: 80.0%
KNN	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA	Accuracy: 73.33% Sensitivity: 83.33% Specificity: 63.33%
Logistic Regression	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA'	Accuracy: 81.67% Sensitivity: 90.0% Specificity: 73.33%

	'ExerciseIncludedAngina_EIA, Age	
Gaussian_NaiveBayes	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA' 'ExerciseIncludedAngina_EIA, Age, SerumCholesterol_SCH, RestingBP_RBP' 'ChestPainType_CPT, RestingECGResult_RES, PeakExerciseSegment_PES	Accuracy: 85.0% Sensitivity: 93.33% Specificity: 76.67%
Random Forest	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA' 'ExerciseIncludedAngina_EIA, Age, SerumCholesterol_SCH	Accuracy: 78.33% Sensitivity: 90.0% Specificity: 66.67%
SVM	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA	Accuracy: 65.0% Sensitivity: 80.0% Specificity: 50.0%

Below, *table 5.8* shows the prediction result for different classification methods along with GANN applying feature selection. Here proposed model GANN outperform along with NB over other methods. GANN achieved 94% accuracy with features "SerumCholesterol_SCH, MaxHeartRate_MHR, RestingECGResult_RES, FastingBloodSuger_FBS and RestingBP_RBP" where experiments executed on batch size 100 for 2000 epoch having 2-hidden layered neural network structure (table 5.9 list the parameter setup for GANN) which is more than other methods in that NB outperformed and remains at 85% with feature set "MaxHeartRate_MHR, VCA,

OldPeak_OPK, ThalliumScan_THA, ExerciseIncludedAngina_EIA, Age, SerumCholesterol_SCH, RestingBP_RBP, ChestPainType_CPT, RestingECGResult_RES, PeakExerciseSegment_PES". Apart from them other method performance remains nearby 80% or below.

Table 5.8: Result comparison of GANN with other classical methods

Method	**Feature details**	**Model performance**
GANN	FastingBloodSuger_FBS, RestingBP_RBP, SerumCholesterol_SCH, MaxHeartRate_MHR, RestingECGResult_RES	Accuracy: 94.67% Sensitivity: 93.23% Specificity: 76.77%
KNN	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA	Accuracy: 73.33% Sensitivity: 83.33% Specificity: 63.33%
Linear Regression	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA' 'ExerciseIncludedAngina_EIA, Age	Accuracy: 81.67% Sensitivity: 90.0% Specificity: 73.33%
Gaussian_NaiveBayes	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA' 'ExerciseIncludedAngina_EIA, Age, SerumCholesterol_SCH, RestingBP_RBP'	Accuracy: 85.0% Sensitivity: 93.33% Specificity: 76.67%

	'ChestPainType_CPT, RestingECGResult_RES, PeakExerciseSegment_PES	
Random Forest	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA' 'ExerciseIncludedAngina_EIA, Age, SerumCholesterol_SCH	Accuracy: 78.33% Sensitivity: 90.0% Specificity: 66.67%
SVM	MaxHeartRate_MHR, VCA, OldPeak_OPK, ThalliumScan_THA	Accuracy: 65.0% Sensitivity: 80.0% Specificity: 50.0%

Below in *table 5.9*, a comprehensive overview of the parameter configurations employed during the training of the GANN model has been highlighted. This table encapsulates an array of diverse combinations that were systematically explored. Through the evaluation process, we identified the optimal set of parameters which ultimately served as the blueprint for the model depicted below. This iterative approach to parameter selection ensures that the GANN model is finely-tuned and adept at producing accurate predictions.

Table 5.9: Comprehensive Overview of Hyperparameters for the GANN Model

Hyperparameter	**Values**
Hidden layers	2
Neurons inside hidden layer	8, 4
Activation method used inside Hidden Layer	ReLu

Activation method on Output Layer	Sigmoid
Learning-rate	0.001
Loss Function	Binary Cross-Entropy
Number of Epoch	2000
Batch side	100
Mutation%	3%

In *figure 5.20*, the performance of GANN over 2000 epochs has been depicted. Initially, the model was in its learning phase, exhibiting a less effective performance. However, with subsequent epochs, the generation of new offspring contributed significantly to the learning process, which is evident in the plotted curve.

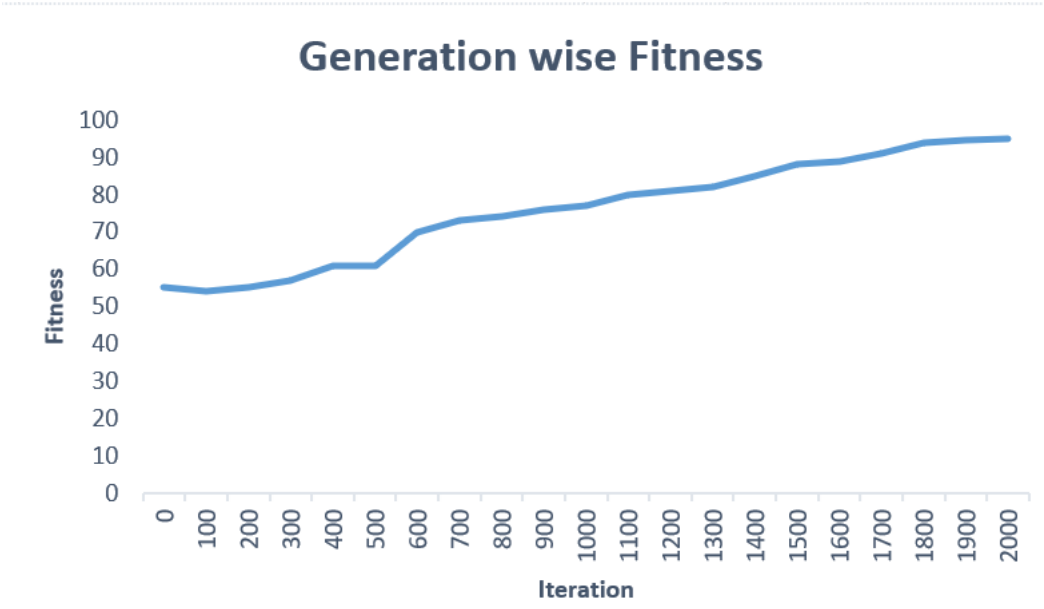


Figure 5.20: Graph showing GANN test results demonstrating fitness for 2000 iteration

In *figure 5.21*, the fitness results during the training of GNFIS are depicted. It's evident that initially, the learning rate was slow, and the model accuracy was quite low.

However, as the training progressed, around the 600-epoch mark, the model started to yield improved results.

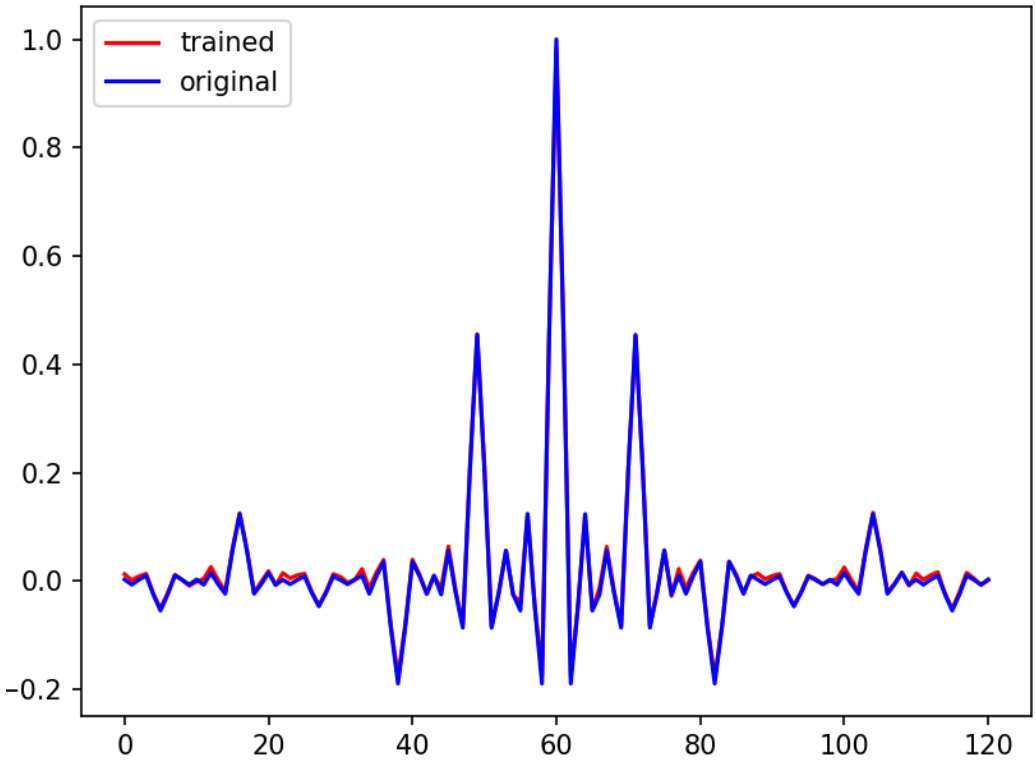


Figure 5.21: Diagram shows the fitness result of GNFIS

Below, *table 5.10* presents the outcomes obtained from GANFIS, alongside a comparative analysis with previous research efforts by other scholars. The results unequivocally demonstrate that the proposed system outperforms with an impressive accuracy of 94%.

Table 5.10: Result comparison with existing state-of-the-art work

Ref	**Author**	**Year**	**Method**	**Dataset**	**Result**
[8]	Negar Ziasabounchi "International Journal of Electrical & Computer Sciences"	2014	ANFIS, GA	Cleveland dataset	92.30%

[9]	Oluwarotimi Williams Samuel et al. “Expert Systems with Applications”	2017	ANN, Fuzzy_AHP	Cleveland dataset	91.10%
[10]	A.V. Senthil Kumar “Journal of Artificial Intelligence”	2012	ANFIS, MATLAB	Cleveland dataset	91.83%
[12]	Zeinab Arabasadi et al. “Computer Methods and Programs in Biomedicine”	2017	ANN, GA	Z-Alizadeh Sani	93.85%
[13]	Kaan and Ahmet “Procedia Computer Science”	2017	ANN-Fuzzy_AHP, GARFNN	Cleveland dataset	91.10%
[14]	G. S. G. Thippa Reddy et al.	2020	GA, FL	Cleveland dataset	90%
[31]	MAbushariah et al. “Journal of Software Engineering and Applications”	2014	MLP, ANN, ANFIS, MATLAB	Cleveland dataset	87%
Proposed work		2022	Genetical Neural Fuzzy Inference System (GANFIS)	Cleveland dataset	94%

Below *table 5.11* shows the test cases calculating the probability percentage result for heart attack using proposed system.

Table 5.11: Result demonstration of Inference System with test cases

Test Cases	**Result**
"RestingBP_RBP":210, **"SerumCholesterol_SCH"**:290, **"FastingBloodSugar_FBS"**:0, **"RestingECGResult_RES"**:0, **"RestingHeartRate_RHR"**:100, **"MaxHeartRate_MHR"**:185, **"CigratePerDay_CPD"**:4, **"HeartDiseaseFamilyHistory"**:0, **"IsHeartPatient"**:0	57.57% "MODERATE-HIGH"
"RestingBP_RBP":125, **"SerumCholesterol_SCH"**:215, **"FastingBloodSugar_FBS"**:0, **"RestingECGResult_RES"**:2, **"RestingHeartRate_RHR"**:75, **"MaxHeartRate_MHR"**:175, **"CigratePerDay_CPD"**:0, **"HeartDiseaseFamilyHistory"**:0, **"IsHeartPatient"**:0	34.96% "LOW-MODERATE"
"RestingBP_RBP":95, **"SerumCholesterol_SCH"**:190, **"FastingBloodSugar_FBS"**:0, **"RestingECGResult_RES"**:0, **"RestingHeartRate_RHR"**:90, **"MaxHeartRate_MHR"**:160, **"CigratePerDay_CPD"**:10,	34.84% "LOW-MODERATE"

“HeartDiseaseFamilyHistory”:0, **“IsHeartPatient”**:0	
“RestingBP_RBP”:110, **“SerumCholesterol_SCH”**:170, **“FastingBloodSugar_FBS”**:0, **“RestingECGResult_RES”**:2, **“RestingHeartRate_RHR”**:63, **“MaxHeartRate_MHR”**:158, **“CigratePerDay_CPD”**:55, **“HeartDiseaseFamilyHistory”**:1, **“IsHeartPatient”**:1	33.35% “LOW-MODERATE”
“RestingBP_RBP”:138, **“SerumCholesterol_SCH”**:240, **“FastingBloodSugar_FBS”**:0, **“RestingECGResult_RES”**:0, **“RestingHeartRate_RHR”**:87, **“MaxHeartRate_MHR”**:215, **“CigratePerDay_CPD”**:35, **“HeartDiseaseFamilyHistory”**:1, **“IsHeartPatient”**:1	57.57% “MODERATE-HIGH”
“RestingBP_RBP”:145, **“SerumCholesterol_SCH”**:285, **“FastingBloodSugar_FBS”**:0, **“RestingECGResult_RES”**:2, **“RestingHeartRate_RHR”**:70, **“MaxHeartRate_MHR”**:168, **“CigratePerDay_CPD”**:55,	57.57% “MODERATE-HIGH”

“HeartDiseaseFamilyHistory”:1, **“IsHeartPatient”**:1	
“RestingBP_RBP”:130, **“SerumCholesterol_SCH”**:270, **“FastingBloodSugar_FBS”**:1, **“RestingECGResult_RES”**:0, **“RestingHeartRate_RHR”**:103, **“MaxHeartRate_MHR”**:158, **“CigratePerDay_CPD”**:50, **“HeartDiseaseFamilyHistory”**:1, **“IsHeartPatient”**:1	85.35% “CRITICAL”
“RestingBP_RBP”:140, **“SerumCholesterol_SCH”**:296, **“FastingBloodSugar_FBS”**:0, **“RestingECGResult_RES”**:0, **“RestingHeartRate_RHR”**:109, **“MaxHeartRate_MHR”**:148, **“CigratePerDay_CPD”**:50, **“HeartDiseaseFamilyHistory”**:1, **“IsHeartPatient”**:0	57.57% “MODERATE-HIGH”
“RestingBP_RBP”:80, **“SerumCholesterol_SCH”**:180, **“FastingBloodSugar_FBS”**:0, **“RestingECGResult_RES”**:0, **“RestingHeartRate_RHR”**:55, **“MaxHeartRate_MHR”**:100, **“CigratePerDay_CPD”**:0,	0% “NORMAL or LOW”

“HeartDiseaseFamilyHistory”:0, **“IsHeartPatient”**:0	

5.7 Summary

In this chapter we discussed the different scenarios experimented using different classical methods i.e., SVM, DT, RF, KNN, LR, NB as well as proposed hybrid method. During experiment difference variance has been applied and results were noted. Result clearly shows naive bayes outperformed in nearly all variance but when we experimented with GANN, the result remains unbeatable. GANN algorithm gives better result for datasets having small number of samples so is better for heart disease with small dataset like Cleveland. Hybrid systems gives better result than traditional classic models. Classifiers in combination of feature selection methods gives better result than alone. Result also shows that reduced size of features can improve the efficiency of model which has been measured as accuracy. This can also help in improving the execution time of the classification models. Techniques like ANN and SVM are widely used in health disease prediction due to their flexibility and requirement of smaller training datasets. Generation of more accurate results will be possible by further enhancement in existing hybrid systems converting it into intelligent hybrid systems. One thing we noted in all experiment that each classical model and GANN improved its efficiency when applying the variance.

Chapter 6: Model Integration and Testing with Web Service

6.1 Introduction

In previous chapters there has been a discussion to complete the proposed framework which started with introduction to basic concepts related to heart disease and machine learning and went through literature review, research framework explanation, data collection approaches and experimental scenarios and result covering multiple scenarios to cover the framework. This chapter explores the vital process of integrating the trained model into a web service for public accessibility. It encompasses the discussion of API exposure techniques and the deployment process, ensuring its availability for public utilization.

6.2 Model integration with FastAPI

Once the model undergoes rigorous training and is primed for prediction, the subsequent crucial step is to encapsulate it in a manner that allows seamless integration into various applications. The optimal approach involves constructing a web service, embedding the trained model, and subsequently exposing it through an endpoint. In this advanced stage of our research, we have employed the widely acclaimed Python library, FastAPI, to facilitate this process. The process has been divided into stages like “create packaging trained model”, “unpackaging packaged model”, “creating web service”, “exposing end point” and “starting web service”, same has been discussed below.

- *Packaging the trained model:* In the initial stage, the thoroughly trained model is packaged, ensuring it is prepared for efficient integration into the upcoming web service. The “*joblib*” Python library, known for its efficacy in system-readable bundling, has been employed for this purpose ensuring all required details are captured and bundled properly. After correct packaging the model is stored, creating a “*pickle (.pkl)*” file..

- *Unpackaging the package:* Subsequently, before integrating the model into the web service, it must be reloaded within the system. Here, we utilize the same "joblib" Python library for seamless unpackaging, ensuring that the model is in a state ready for integration.
- *Building endpoint:* This crucial step involves the actual creation of the web API and the exposure of endpoints. For this task, the "*FastAPI*" library, known for its adeptness in this domain, is harnessed to establish an efficient and effective web service. The endpoint created so far, acts as the gateway through which users can access the predictive capabilities of the model.
- *Initiating the environment:* With the packaged model primed for reloading and integration into the system, and an endpoint established for revelation, the final step involves initiating the environment and making the API accessible to the masses. Since our research endeavour is conducted locally, the selected library handles the initialization within the local environment, specifically on localhost using port 4000. This setup allows for thorough endpoint testing. The Python library facilitating this process is "uvicorn", ensuring a seamless transition from development to practical deployment.

  Environment start on local port URL http://127.0.0.1:4000 and endpoints can be accessed on URL http://127.0.0.1:4000/docs. In the later phase this can be deployed and can be accessed via public URL.

6.3 Testing API

With the API now operational, the system is prepared for testing. This marks a crucial milestone in the research process signifying the transition from development to evaluation and validation, *figure 6.1* shows the API definition loaded on the web.

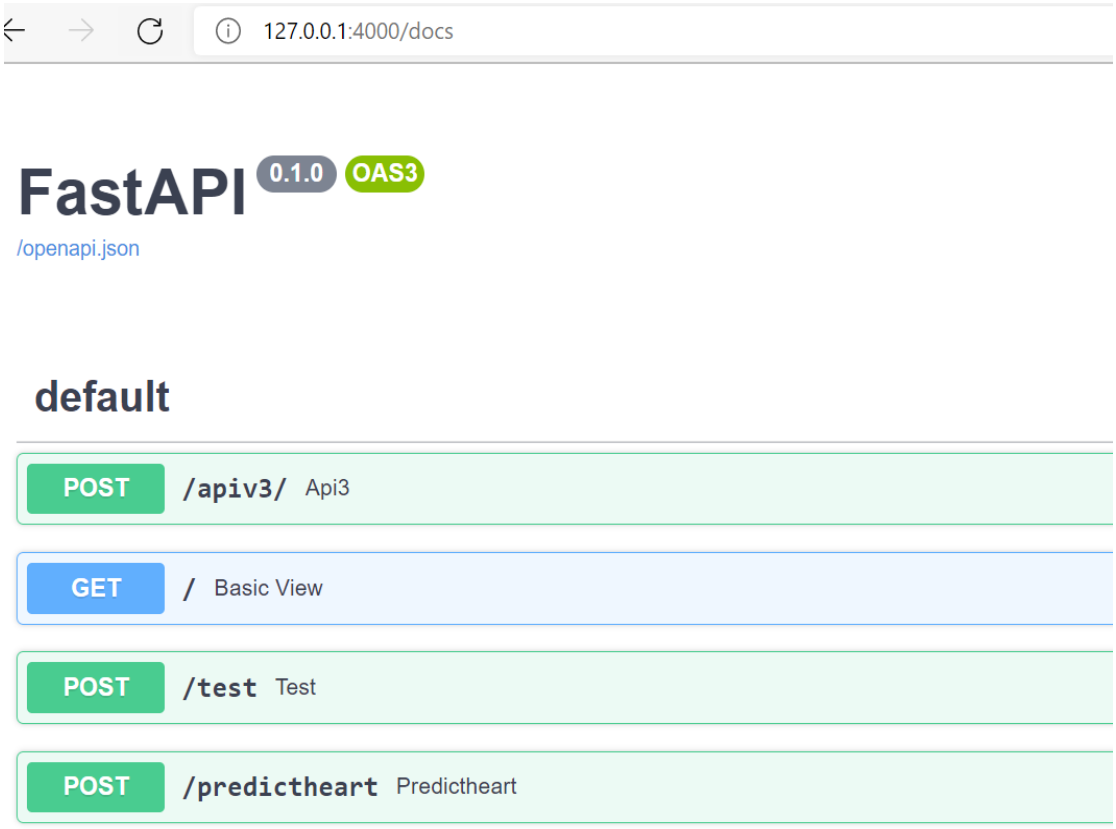


Figure 6.1: Web interface displaying loaded API definition

Here endpoint "/predictheart" is a post call which means that it is secure and cannot directly called from any application. Below *figure 6.2* shows the request detail for the endpoint.

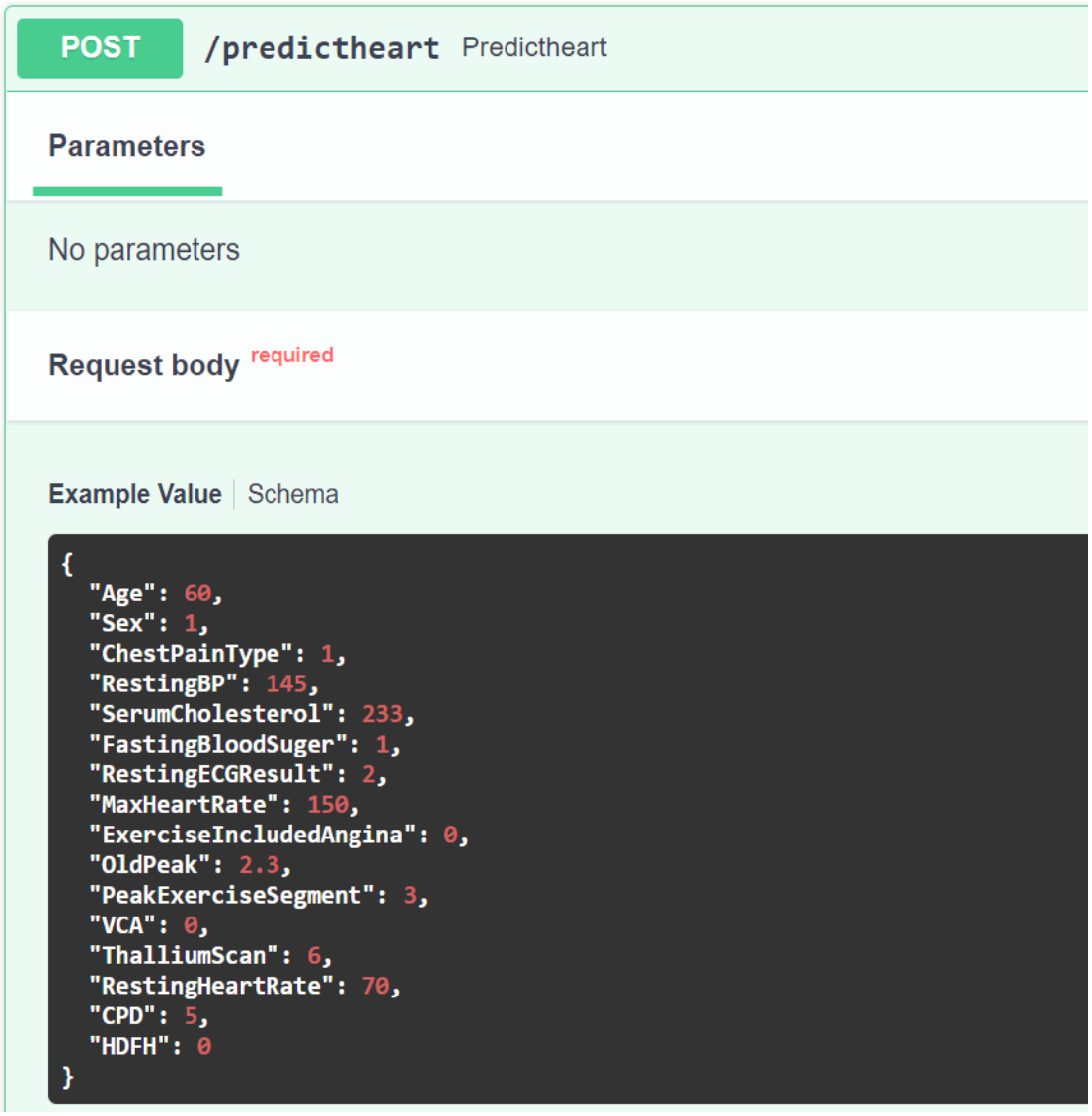


Figure 6.2: API request for the endpoint

Below, *figure 6.3* and *6.4* shows the test request and prediction response.

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:4000/predictheart' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "Age": 60,
  "Sex": 1,
  "ChestPainType": 1,
  "RestingBP": 145,
  "SerumCholesterol": 233,
  "FastingBloodSuger": 1,
  "RestingECGResult": 2,
  "MaxHeartRate": 150,
  "ExerciseIncludedAngina": 0,
  "OldPeak": 2.3,
  "PeakExerciseSegment": 3,
  "VCA": 0,
  "ThalliumScan": 6,
  "RestingHeartRate": 70,
  "CPD": 5,
  "HDFH": 0
}'
```

Request URL

```
http://127.0.0.1:4000/predictheart
```

Figure 6.3: Curl request with URL

A set of medical parameters having 16 features has been added as a request to the API, are pivotal in facilitating the prediction process. Below is the response for the same.

Server response

Code	Details
200	Response body


```
{
  "mdata": {
    "Age": 60,
    "Sex": 1,
    "ChestPainType_CPT": 1,
    "RestingBP_RBP": 145,
    "SerumCholesterol_SCH": 233,
    "FastingBloodSuger_FBS": 1,
    "RestingECGResult_RES": 2,
    "MaxHeartRate_MHR": 150,
    "ExerciseIncludedAngina_EIA": 0,
    "OldPeak_OPK": 2.3,
    "PeakExerciseSegment_PES": 3,
    "VCA": 0,
    "ThalliumScan_THA": 6
  },
  "predict": "1",
  "risk_percentage": "57.81226836834753"
}
```


Response headers


```
content-length: 329
content-type: application/json
date: Wed,04 May 2022 16:33:39 GMT
server: uvicorn
```


Figure 6.4: Endpoint response for the request

In the figure above, the response to the provided request is displayed in two sections:

- *predict*: This section predicts the existence of heart disease in individuals.
- *risk_percentage*: This section indicates the likeliness percentage of a heart attack.

In this specific case, the prediction indicates the existence of HD, and the associated risk percentage is 57.81. This implies that the individual has a 57% likelihood of experiencing a heart attack based on the provided medical parameters and their current health condition.

6.4 Test Cases

The following *table 6.1* displays a selection of test cases alongside their respective responses obtained through the implemented API. It provides a clear illustration of the requests made and the corresponding outcomes as processed by the API.

Table 6.1: API Test cases request/response

Request	**Response**
{ "Age": 65, "Sex": 1, "ChestPainType": 1, "RestingBP": 140, "SerumCholesterol": 235, "FastingBloodSuger": 1, "RestingECGResult": 2, "MaxHeartRate": 155, "ExerciseIncludedAngina": 0, "OldPeak": 2.3, "PeakExerciseSegment": 3, "VCA": 0, "ThalliumScan": 6, "RestingHeartRate": 80, "CPD": 8, "HDFH": 0 }'	**"predict": "1", "risk_percentage": "57.81226836834753"**
{ "Age": 40, "Sex": 1, "ChestPainType": 1, "RestingBP": 85, "SerumCholesterol": 190, "FastingBloodSuger": 0, "RestingECGResult": 2, "MaxHeartRate": 110, "ExerciseIncludedAngina": 0, "OldPeak": 2, "PeakExerciseSegment": 3, "VCA": 0, "ThalliumScan": 6, "RestingHeartRate": 80, "CPD": 10, "HDFH": 0 }'	**"predict": "1", "risk_percentage": "34.90275702665084"**

{ "Age": 26, "Sex": 1, "ChestPainType": 1, "RestingBP": 95, "SerumCholesterol": 205, "FastingBloodSuger": 0, "RestingECGResult": 1, "MaxHeartRate": 118, "ExerciseIncludedAngina": 0, "OldPeak": 1, "PeakExerciseSegment": 1, "VCA": 0, "ThalliumScan": 4, "RestingHeartRate": 75, "CPD": 2, "HDFH": 0 }'	**"predict": "0", "risk_percentage": "34.90275702665084"**

6.5 Deployment

This section discusses the deployment of exposed API on cloud environment. Once the API is ready and tested in local environment, it can be deployed to public platform using various available medium so that it could be available for public use. There is various deployment medium like the service can be deployed in in-house infra support or it can be deployed as a service using cloud environment. Now a days cloud environment is very popular few of them are Azure cloud, AWS cloud or Google cloud. The number is not limited to these and the list is very long and many big organizations is entering into this section due to huge popularity of cloud for example one of the big organizations named IBM is also showing its presence in cloud technology and providing its services to end users. In this section we will not discuss all of them but discuss one of it provided by Microsoft that is Azure Cloud.

6.5.1 Deployment using Azure cloud

Azure cloud is very popular cloud platform which is being opted by many big organizations due to the level of services it provides along with the security measurement, time and cost flexibility, nearly 0% non-availability of environment and 24X7 infra support. Below sub sections briefly discuss the key highlights of Azure cloud platform and way of deployment we can use to deploy the service.

6.5.1.1 Deployment as a service

Azure cloud provide the feasibility to deploy the API as a service in workspace. For that we just need to create a resource group, provide the costing estimation, do some custom settings like location and then with the help of few clicks it can deploy the API as a service which shall be publicly available just after a setup process of 2-3 minutes. It also provides an URL which can customize with our own domain name as well.

6.5.1.2 Deployment as a function app

An API can also be deployed as a function app inside Azure cloud. Function app has totally different concept as it has its own environment and a SME need not to worry about the infra related to function app. Function app also provides a flexibility to use the API as a part of messaging or directly as a http service.

6.5.1.3 Deployment as a logic app

An API can also be deployed as a part of logic app wherein multiple flow can be associated before or after using that API. This way the API can work with the association with other functionality of application like the output of the API can be use directly in other flows and can generate other output which could be required for another flow. This way it can help in creating a flow like structure and a complex system can be build.

6.5.1.4 Deployment as a docker image

The API can also be deployed as a docker image which make it cross-platform compatible. This way the image can be install in any environment like windows or LINUX without making any changes to existing code.

6.5.1.5 Deployment in Kubernetes

Kubernetes is a new way to deploy the web service and it provide environment and security which can be avail by the web services. One can directly deploy the docker image (discussed in section 6.4.1.5) in Kube and run the URL.

6.5.1.6 Use of DevOps for continuous deployment

DevOps is the new technology which maintain the deployment of web service and help in automate the deployment and build process for example if any code changes is made in the API created so far, it can be directly get deployed using DevOps service without using any human interference with few settings.

6.5.1.7 Terraforms in deployment

Terraform from HashiCorp is an open-source tool which is termed as IaC or Infrastructure-as-Code has been used for managing and provisioning cloud infrastructure [79] is the new way of deployment can help DevOps service to be stronger than previous and help in automated deployment with the help of terraform scripts. These scripts can even make an image or can create infra related things like adding resource group as per environment i.e., dev, uat or production and finally can deploy the API on provided environment.

6.6 Summary

In this chapter we discussed the details on integrating trained model with web service to use further in different applications or any other researches. For that we used very popular python library called FastAPI, joblib and uvicorn which helped in integration part. We discussed that how the trained model can be packaged into pickle file and later be reloaded again when needed during API call. We also discussed the process of API creation, initializing the environment and executing the endpoints by passing request and get the response. Later we discussed the deployment of web service so that it could be available in public domain.

Chapter 7: Conclusion and Future Scope

7.1 Conclusion

In the chapter, an insight of the research work and the thesis has been provided and along with this, it shares the recommendations for the future scope of the area. This thesis is structured seven chapters and each chapter discuss the various stages of research work proposed in the thesis. The research work proposes a framework for *"prediction of heart disease and finding the probability of heart attack"* based on the medical parameters using Machine Learning. Complete research work has been split into two parts wherein part 1 focuses on prediction of heart disease and in second part deals ins finding the heart attack probability. Thesis starts with chapter 1 where we discussed about heart disease along with the symptom, cause, preventive measures and etc. We also discussed the basic of machine learning algorithms. In second chapter we reviewed more than hundreds research papers from various journals from Scopus, SCI, IEEE and other indexed research papers. This helped in having background of existing works related to this area, finding research gaps, framing research objectives, getting idea of model efficiencies already achieved by other research scholars. Then we moved to chapter 3 and discussed research objective and framework to achieve these objectives. In this chapter we discussed in detail about the work to be done and methods which can help in achieving the goals. As we already said that we have framed the framework in such a way that work will be accomplished in two parts, so in this chapter we discussed both the stages which can lead in completing the framework and to build the desired model and hence completing the apart of research objectives. Next, in chapter 4, we discussed the dataset in details which is to be used throughout the research work. Here we are using a popular secondary HD dataset which is accessible and downable from UCI repository and the name is Cleveland dataset. In this chapter we discussed positive and negative point of this dataset and the medical parameter we chosen for the experiments. In this chapter we also discussed the steps in data pre-processing like importing dataset, encoding values, selecting features, scaling data and

so on. These steps are very important in building the efficient model which can outperform and save training and installing cost as well. In chapter 5 we discussed all the experimental scenarios used to cover both the two stages in the study. In this chapter we discussed the detailing of experiment, methods used, result obtained and the ways to calculate result like confusion matrix, evaluation formula for accuracy, sensitivity, specificity and so on. In chapter 6 we discussed the test cases, API integration and deployment related stuff which can help in future expansion of the model build so far for various purposes like to develop an application which can utilize the API and serve the people for social cause. Chapter 7 enlighten the summary of all chapters discussed so far in the thesis.

7.2 Future Scope

This research work strongly demands further work to expand the current work by overcoming the limitation of datasets used in this work, adding more medical parameter to the framework so that optimistic result could be maximized, more features to current work could be added like building model for suggesting precaution to lowering the threshold of medical parameter. This is also a suggestion to monitor the model on regular basis so that performance could be monitored regularly to ensure that it continues to perform well. Another suggestive future work is on the importance of incorporating explanations in machine learning algorithms used in clinical diagnosis systems. While ML algorithms have been recognized as an efficient technique in identifying the medical conditions, there is often a lack of transparency in the reasoning behind their predictions. This lack of transparency can lead mistrust from healthcare professionals, ultimately affecting patient care. The study should explore the different methods for incorporating explanations in ML algorithms, such as using rule-based systems and discusses their potential impact on improving the interpretability and trustworthiness of clinical diagnosis systems. In the study, we did not focus much on deployment of web service created for model so deployment strategies should also be made so that built API could be available publicly and serve in different application for solving various purpose. Machine learning models can be used in real time environments by deploying them as part of a software system that can process data and

provide output in real-time. This involves integrating the model with the software system, which can be done using various techniques such as APIs, microservices, or containerization. But when it comes to using these models in real-time environments, there are several challenges to consider for example, the model must be able to handle large data, able to process it quickly, adapt and learn from new data in real-time, as new information may become available that was not present during the model's initial training and finally, there is the challenge of ensuring the model is secure and protected against potential attacks too so it is very important to develop strategies as a future scope to address them so that model's effectiveness and reliability could be insured.

BIBLIOGRAPHY

1. WHO Cardiovascular Disease. Available online: https://www.who.int/health-topics/cardiovascular-diseases/(accessed on 1 February 2022)
2. S. Nashif, R. Rakib Raihan, R. Islam, and H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," World Journal of Engineering and Technology, vol. 6, pp. 854-876, 2018.
3. A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," Mobile Information Systems, vol. 2018, no. Special Issue, pp. 21, 2018.
4. M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol. 179, pp. 104992, 2019.
5. R. K. Jha, S. Henge, and A. Sharma, "Optimal Machine Learning Classifiers for Prediction of Heart Disease," SERSC, 2020.
6. A. D. Dolatabadi, S. E. Z. Khadem, and B. M. Asl, "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM," Computer Methods and Programs in Biomedicine, vol. 138, pp. 117-126, 2017.
7. M. Aghamohammadi, M. Madan, J. K. Hong, and I. Watson, "Predicting Heart Attack Through Explainable Artificial Intelligence," in Computational Science - ICCS 2019, vol. 11537, pp. 633-645.
8. N. Ziasabounchi and I. Askerzade, "ANFIS Based Classification Model for Heart Disease Prediction," International Journal of Electrical & Computer Sciences, vol. 14, no. 2, 2014.

9. O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," Expert Systems with Applications, vol. 68, pp. 163-172, 2017.
10. A. S. Kumar, "Diagnosis of Heart Disease using Fuzzy Resolution Mechanism," Journal of Artificial Intelligence, vol. 5, pp. 47-55, 2012.
11. A. Yazdani and K. Ramakrishnan, "Performance Evaluation of Artificial Neural Network Models," in Springer ICIBEL 2015, pp. 179-182.
12. Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Computer Methods and Programs in Biomedicine, vol. 141, pp. 19-26, 2017.
13. K. Uyar and A. Iihan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," Procedia Computer Science, vol. 120, pp. 588-593, 2017.
14. G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," Evol. Intel., vol. 13, pp. 185–196, 2020.
15. M. Akgul, O. E. Sönmez, and T. Ozcan, "Diagnosis of Heart Disease Using an Intelligent Method: A Hybrid ANN – GA Approach," INFUS 2019, pp. 1250-1257.
16. N. G. B. Amma, "Cardiovascular disease prediction system using genetic algorithm and neural network," International Conference on Computing, Communication and Applications 2012, pp. 1-5.
17. S. Nikam, P. Shukla, and M. Shah, "Cardiovascular Disease Prediction using Genetic Algorithm and Neuro-fuzzy System," International Journal of Latest Trends in Engineering and Technology, vol. 8, Issue (2), pp.104-110, 2017.
18. M.A. Jabbar, B.L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technology, vol. 10, pp.85-94, 2013.
19. M.A. Jabbar, B.L. Deekshatulu, and P. Chandra, "An Evolutionary Algorithm for Heart Disease Prediction", Wireless Networks and Computational Intelligence. ICIP 2012. Communications in Computer and Information Science, vol 292, pp.378-389, 2012.

20. F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques," in The Impact of Digital Technologies on Public Health in Developed and Developing Countries. ICOST 2020, vol. 12157, pp. 299-306.
21. I. Yekkala and S. Dixit, "A Novel Approach for Heart Disease Prediction Using Genetic Algorithm and Ensemble Classification," in IntelliSys 2020, Advances in Intelligent Systems and Computing, vol. 1251, 2020, pp. 468-489.
22. F. Ali, S. Sappagh, S.M.R. Islam, D. Kwak, A. Ali, M. Imran, and K. Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," Information Fusion, vol. 63, pp. 208-222, 2020.
23. L.A. Demidova, M.M. Egin, and R.V. Tishkin, "A Self-tuning Multiobjective Genetic Algorithm with Application in the SVM Classification," Procedia Computer Science, vol. 150, pp. 503-510, 2019.
24. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D. Hungarian Institute of Cardiology. Budapest: Andras Janosi, "Heart Disease Data Set - UCI. Available online: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
25. S.K. Henge and B. Rama, "Five Layered-Neural Fuzzy Closed Loop Hybrid Control System with Compound Bayesian Decision Making Process for Classification cum Identification of Mixed Connective Conjunct Consonants and Numerals," in Advances in Intelligent Systems and Computing, vol. 553, pp. 619-629, 2017.
26. S.K. Henge and B. Rama, "Neural Fuzzy Closed Loop Hybrid System for Classification, Identification of Mixed Connective Consonants and Symbols with Layered Methodology," Intelligent Control and Energy Systems (ICPEICES) 2016, pp. 1-6.
27. B. Singh and S.K. Henge, "Neural Fuzzy Inference Hybrid System with SVM for Identification of False Singling in Stock Market Prediction for Profit Estimation," in Advances in Intelligent Systems and Computing, vol. 1197, pp. 221-227, 2020.
28. Y. Hayashi, K. Nakajima, and K. Nakajima, "A rule extraction approach to explore the upper limit of hemoglobin during anemia treatment in patients with predialysis chronic kidney disease," Informatics in Medicine Unlocked, vol. 17, 2019.

29. Y. Li, et al., "Combining Convolutional Neural Network and Distance Distribution Matrix for Identification of Congestive Heart Failure," IEEE Access, vol. 6, pp. 39734-39744, 2018.
30. T. Santhanam and E. P. Ephzibah, "Heart Disease Prediction Using Hybrid Genetic Fuzzy Model," Indian Journal of Science and Technology, vol. 8, no. 9, pp. 797-803, 2015.
31. M. Abushariah, M.A.M. et al. "Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) Approaches," Journal of Software Engineering and Applications, vol. 7, issue 12, 2014.
32. Heart attack and stroke symptom - Atherosclerosis and cholesterol. Available online. Atherosclerosis | American Heart Association
33. Understanding blood pressure readings. Available online. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
34. Reading the new blood pressure guidelines. Available online. https://www.health.harvard.edu/heart-health/reading-the-new-blood-pressure-guidelines
35. Managing Heart Failure Symptoms. Available online. Managing Heart Failure Symptoms | American Heart Association
36. Supraventricular tachycardia. Available online. https://www.mayoclinic.org/diseases-conditions/supraventricular-tachycardia/symptoms-causes/syc-20355243
37. Fuzzification. Available online. https://www.sciencedirect.com/topics/engineering/fuzzification
38. Support vector machine, wikipedia. Available online: https://en.wikipedia.org/wiki/Support_vector_machine
39. Decision Tree - Wiki," wikipedia. Available online. https://en.wikipedia.org/wiki/Decision_tree
40. Decision Tree Classification Algorithm - Javapoint. Available online. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

41. Random Forests, Available online. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
42. K-Nearest Neighbors Algorithm - Wiki, wikipedia, Available online. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
43. What is Logistic Regression, wiki, Available online. https://en.wikipedia.org/wiki/Logistic_regression
44. Naive Bayes classifier - Wiki, wikipedia, Available online. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
45. Understanding Random Forest. Available online. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest
46. Logistic Regression in Machine Learning - Javapoint. Available online. https://www.javatpoint.com/logistic-regression-in-machine-learning
47. Smoking and Heart Disease, webmd. Available online. https://www.webmd.com/heart-disease/smoking-heart-disease
48. How smoking affects heart health. Available online. https://www.fda.gov/tobacco-products/health-effects-tobacco-use/how-smoking-affects-heart-health
49. How High Blood Pressure Can Lead to a Heart Attack. Available online. https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-a-heart-attack
50. Heart Disease and Lowering Cholesterol. Available online. https://www.webmd.com/heart-disease/guide/heart-disease-lower-cholesterol-risk
51. Heart Disease and Diabetes. Available online. https://www.webmd.com/diabetes/heart-blood-disease
52. Diabetes and Your Heart. Available online. https://www.cdc.gov/diabetes/library/features/diabetes-and-heart.html
53. A Layman's Guide to Deep Neural Networks. Available online. https://towardsdatascience.com/a-laymans-guide-to-deep-neural-networks-ddcea24847fb
54. Activation Functions in Neural Networks | by SAGAR SHARMA | Towards Data Science, Available online. https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

55. Implement sigmoid function using Numpy. Available online. https://www.geeksforgeeks.org/implement-sigmoid-function-using-numpy/
56. Warning signs for heart attack. Available online. https://www.heart.org/en/health-topics/heart-attack/warning-signs-of-a-heart-attack
57. Introduction to Genetic Algorithms — Including Example Code Available online. https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3
58. Genetic Algorithms. Available online. https://www.geeksforgeeks.org/genetic-algorithms/
59. Genetic Algorithms. Available online. https://en.wikipedia.org/wiki/Genetic_algorithm
60. Fuzzy Inference. Available online. https://www.sciencedirect.com/topics/engineering/fuzzy-inference
61. P. Tahmasebi, A. Hezarkhani, "A hybrid neural networks-fuzzy logic-genetic algorithm for grade estimation," Computers & Geosciences, vol. 42, pp. 18-27, 2012.
62. Arup Kumar Nandi. GA-Fuzzy Approaches: Application to Modeling of Manufacturing Process, *CSIR-CMERI* **2013,** doi 10.1007/978-3-642-25859-6_4
63. Fuzzy Logic - Inference System Available online. https://www.tutorialspoint.com/fuzzy_logic/fuzzy_logic_inference_system.htm
64. B. Khoshnevisan, S. Rafiee, M. Omid and H. Mousazadeh, "Development of an intelligent system based on ANFIS for predicting wheat grain yield on the basis of energy inputs," in Information Processing in Agriculture, vol. 1, pp. 14-22, 2014.
65. R. Misir and R. K. Samanta, "A Study on Performance of UCI Hungarian Dataset Using Missing Value Management Techniques," International Journal of Computer Sciences and Engineering, vol. 5, no. 2, 2017.
66. Heart Health and Aging. Available online. https://www.nia.nih.gov/health/heart-health-and-aging
67. Z. Gao, Z. Chen, A. Sun and X. Deng, "Gender differences in cardiovascular disease," Medicine in Novel Technology and Devices, vol. 4, 2019.
68. Heart rate as a risk factor for cardiovascular disease. Available online. https://pubmed.ncbi.nlm.nih.gov/19615487

69. K. Fox et al., "Resting heart rate in cardiovascular disease," Journal of the American College of Cardiology, vol. 50, no. 9, pp. 823-830, Aug. 2007.
70. F. Custodis et al., "Heart rate: A global target for cardiovascular disease and therapy along the cardiovascular disease continuum," Journal of Cardiology, vol. 62, no. 3, pp. 183-187, 2013.
71. High blood pressure dangers: Hypertension's effects on your body. Available online. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868
72. How High Blood Pressure Can Lead to a Heart Attack. Available online. https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-a-heart-attack
73. Feature scaling. Available online. https://en.wikipedia.org/wiki/Feature_scaling
74. Data Preprocessing in Data Mining. Available online. https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/
75. Cross-validation (statistics). Available online. https://en.wikipedia.org/wiki/Cross-validation_(statistics)
76. Importance of Cross Validation: Are Evaluation Metrics enough? Available online. https://www.analyticsvidhya.com/blog/2021/05/importance-of-cross-validation-are-evaluation-metrics-enough
77. Sanjay Yadav, Sanyam Shukla; Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. *2016 IEEE 6th International Conference on Advanced Computing (IACC),* **2016**, doi: 10.1109/IACC.2016.25
78. Create your first function in the Azure portal. Available online. https://docs.microsoft.com/en-us/azure/azure-functions/functions-create-function-app-portal
79. Overview of Terraform on Azure - What is Terraform? Available. Online. https://docs.microsoft.com/en-us/azure/developer/terraform/overview
80. A. Davari Dolatabadi, M. Ghaffari, H.R. Marateb, and M. Khosravi, "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM," Computer Methods and Programs in Biomedicine, vol. 138, pp. 117-126, 2017.

81. Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," in Computer Methods and Programs in Biomedicine, vol. 130, pp. 54-64, 2016.
82. M. Abdar et al., "A new machine learning technique for an accurate diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol. 179, 2019.
83. C. K. Shashikant R., "Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter," Appl. Comput. Informatics, vol. 19, pp. 23-45, Jun. 2019.
84. S. N. Qasem and M. Alsaidan, "A New Hybrid Intelligent System for Prediction of Medical Diseases," International Journal of Advanced Computer Science and Applications, 2018.
85. M. S. Amin, M. S. Islam, S. M. Z. Islam and S. B. Uddin, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, vol. 36, pp. 82-93, 2019.
86. S. Nashif, R. Raihan, M. R. Islam, and M. H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," World Journal of Engineering and Technology, vol. 6, no. 4, pp. 854-873, 2018.
87. R. Nichenametla, T. Maneesha, S. Hafeez, and H. Krishna, "Prediction of Heart Disease Using Machine Learning Algorithms," International Journal of Engineering & Technology, pp. 363-366, 2018.
88. A. Adeli and M. Neshat, "A fuzzy expert system for heart disease diagnosis," Proceedings of the International Multiconference of Engineers and Computer Scientists, vol. 1, pp. 17-19, 2010.
89. D. Hassan, H.I. Hussein and M.M. Hassan, "Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis," Biomedical Signal Processing and Control, vol. 75, 2022, article no. 104019.
90. K. Uma Maheswari and J. Jasmine, "Neural Network based Heart Disease Prediction," International Journal of Engineering Research & Technology, vol. 5, issue 17, 2017.

91. C.V. Aravinda, Meng Lin, K.R. Udaya Kumar Reddy, and G. Amar Prabhu, "A deep learning approach for the prediction of heart attacks based on data analysis," in Deep Learning for Medical Applications with Unique Data, 2022, pp. 1-18.
92. Bhuvaneswari Amma N G, "An intelligent approach based on Principal Component Analysis and Adaptive Neuro Fuzzy Inference System for predicting the risk of cardiovascular diseases," 2013 Fifth International Conference on Advanced Computing (ICoAC), 2014, pp. 364-369.
93. Yoshimi Fukuoka and Yoo Jung Oh, “Perceived Heart Attack Likelihood in Adults with a High Diabetes Risk”, Heart & Lung, vol. 52, pp. 42-47, 2022.
94. Yoshimi Fukuoka, JiWon Choi, Melinda S. Bender, Prisila Gonzalez, Shoshana Arai, "Family history and body mass index predict perceived risks of diabetes and heart attack among community-dwelling Caucasian, Filipino, Korean, and Latino Americans - DiLH Survey", Diabetes Research and Clinical Practice, vol. 109, issue 1, pp. 157-163, 2015.
95. J. Abdollahi, B. Nouri-Moghaddam "Feature Selection for Medical Diagnosis: Evaluation for Using a Hybrid Stacked-Genetic Approach in the Diagnosis of Heart Disease", Journal of Medical Ethics and History of Medicine, vol 14, pp. 1-11, 2021.
96. Sudha V. K. and Kumar D., "Hybrid CNN and LSTM Network For Heart Disease Prediction," SN COMPUT. SCI., vol. 4, no. 1, p. 172, 2023, https://doi.org/10.1007/s42979-022-01598-9.
97. A. A. Altae and A. E. Rad, "Diagnosing heart disease by a novel hybrid method: Effective learning approach," Informatics in Medicine Unlocked, vol. 40, pp. 101275, 2023.
98. M. Raihan et al., "Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design," in 2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 299-303, 2016.
99. N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," Health and Technology, vol. 11, pp. 49-62, 2020. https://doi.org/10.1007/s12553-020-00499-2.

100. D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Computer Science, vol. 1, p. 345, 2020. https://doi.org/10.1007/s42979-020-00365-y.
101. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
102. D. K. Ravish, K. J. Shanthi, N. R. Shenoy and S. Nisargh, "Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural networks", Proc. Int. Conf. Contemp. Comput. Inform. (IC3I), pp. 1-6, Nov. 2014.
103. K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857.
104. S. Mohapatra, S. Maneesha, P. K. Patra, S. Mohanty, "Heart Diseases Prediction based on Stacking Classifiers Model," Procedia Computer Science, vol. 218, pp. 1621-1630, 2023. https://doi.org/10.1016/j.procs.2023.01.140.
105. C. Sharma, S. Sharma, M. Kumar and A. Sodhi, "Early Stroke Prediction Using Machine Learning," 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, 2022, pp. 890-894, doi: 10.1109/DASA54658.2022.9765307.
106. S. Mushtaq, K. S. Saini and S. Bashir, "Machine Learning for Brain Stroke Prediction," 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 2023, pp. 401-408.
107. G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 6, 2021.
108. Chicco, D., Jurman, G., "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Medical Informatics and Decision Making, vol. 20, p. 16, 2020. https://doi.org/10.1186/s12911-020-1023-5.

109. A. Newaz, N. Ahmed, and F. S. Haq, "Survival prediction of heart failure patients using machine learning techniques," Informatics in Medicine Unlocked, vol. 26, 100772, 2021. https://doi.org/10.1016/j.imu.2021.100772.
110. F. R. Bühler, K. Vesanen, J. T. Watters, and P. Bolli, "Impact of smoking on heart attacks, strokes, blood pressure control, drug dose, and quality of life aspects in the International Prospective Primary Prevention Study in Hypertension," American Heart Journal, vol. 115, no. 1, part 2, pp. 282-288, 1988. https://doi.org/10.1016/0002-8703(88)90651-5.
111. Heart failure clinical records. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5Z89R.
112. R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets," Procedia Computer Science, vol. 65, pp. 459-468, 2015.
113. S. K. Satapathy, A. Patel, P. Yadav, Y. Thacker, D. Vaniya and D. Parmar, "Machine Learning Approach for Estimation and Novel Design of Stroke Disease Predictions using Numerical and Categorical Features," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-6, doi: 10.1109/ICONAT57137.2023.10080722.
114. G. Sugendran and S. Sujatha, "Earlier identification of heart disease using enhanced genetic algorithm and fuzzy weight based support vector machine algorithm" Measurement: Sensors, vol. 28, pp. 100814, 2023.
115. C. M. Chethan Malode, K. Bhargavi, B. G. Gunasheela, G. Kavana and R. Sushmitha, "Soft set and Fuzzy Rules Enabled SVM Approach for Heart Attack Risk Classification Among Adolescents," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697650.
116. E. Maraj and S. Kuka, "Prediction of Coronary Heart Disease Using Fuzzy Logic: Case Study in Albania," 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), Prague, Czech Republic, 2022, pp. 1-6, doi: 10.1109/ICECET55527.2022.9872569.
117. A. S. Noor, P. A. Venkatachalam, and F. H. Ahmad, "Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System,"

Proceedings of the International Conference on Man-Machine Systems (ICoMMS), Batu Ferringhi, Penang, Malaysia, 2009, pp. 11–13.

118. L. J. Muhammad and E. A. Algehyne, "Fuzzy based expert system for diagnosis of coronary artery disease in Nigeria," Health and Technology, vol. 11, pp. 319-329, 2021. https://doi.org/10.1007/s12553-021-00531-z.

119. P.K. Anooj, "Clinical Decision Support System: Risk Level Prediction of Heart Disease using Weighted Fuzzy Rules," Journal of King Saud University - Computer and Information Sciences, vol. 24, no. 1, pp. 27-40, 2012.

120. A. Czmil, "Comparative Study of Fuzzy Rule-Based Classifiers for Medical Applications," Sensors, vol. 23, no. 2, pp. 992, Jan. 2023, doi: 10.3390/s23020992.

121. Hossain, S., Sarma, D., Chakma, R.J., Alam, W., Hoque, M.M., Sarker, I.H, "A Rule-Based Expert System to Assess Coronary Artery Disease Under Uncertainty." Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, 2020, vol 1235, pp. 143–159. Springer, Singapore.

122. G. Casalino, G. Castellano, C. Castiello, V. Pasquadibisceglie and G. Zaza, "A Fuzzy Rule-Based Decision Support System for Cardiovascular Risk Assessment," in International Workshop on Fuzzy Logic and Applications, Fuzzy Logic and Applications, 2019, vol. 11291, pp. 97-108.

123. H. K. Chang, C. T. Wu, J. H. Liu and J. S. R. Jang, "Using Machine Learning Algorithms in Medication for Cardiac Arrest Early Warning System Construction and Forecasting", 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taichung, Taiwan, 2018, pp. 1-4, doi: 10.1109/TAAI.2018.00010.

124. Hameed A.Z., Ramasamy B., Shahzad M.A. et al., "Efficient hybrid algorithm based on genetic with weighted fuzzy rule for developing a decision support system in prediction of heart diseases", J Supercomputer, vol. 77, pp. 10117–10137, 2021, https://doi.org/10.1007/s11227-021-03677-9.

125. Sharma, P., Saxena, K., "Application of fuzzy logic and genetic algorithm in heart disease risk level prediction," International Journal of System Assurance Engineering and Management, vol. 8, no. Suppl 2, pp. 1109-1125, 2017, https://doi.org/10.1007/s13198-017-0578-8.

126. Nagarajan R. and Thirunavukarasu R., "A neuro-fuzzy based healthcare framework for disease analysis and prediction", Multimedia Tools and Applications, vol. 81, pp. 11737-11753, 2022, https://doi.org/10.1007/s11042-022-12369-2.
127. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.
128. S. M. Mousavi, S. Abdullah, S. T. A. Niaki, and S. Banihashemi, "An intelligent hybrid classification algorithm integrating fuzzy rule-based extraction and harmony search optimization: Medical Diagnosis Applications," Knowledge-Based Systems, vol. 220, p. 106943, 2021. [Online]. Available: https://doi.org/10.1016/j.knosys.2021.106943.
129. O. Terrada, B. Cherradi, A. Raihani and O. Bouattane, "A fuzzy medical diagnostic support system for cardiovascular diseases diagnosis using risk factors," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, Morocco, 2018, pp. 1-6, doi: 10.1109/ICECOCS.2018.8610649.
130. A. Paul, P. Shill, M. Rabin, et al., "Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease," Applied Intelligence, vol. 48, pp. 1739-1756, 2018.
131. M. F. A. Hakim, N. Fajriati, and R. N. Pratama, “Heart Disease Diagnosis Using Tsukamoto Fuzzy Method”, JAIST, vol. 5, no. 1, pp. 12-22, Apr. 2023.
132. O. Taylan, A. Alkabaa, H. Alqabbaa, E. Pamukçu, and V. Leiva, “Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods,” Biology, vol. 12, no. 1, p. 117, Jan. 2023, doi: 10.3390/biology12010117.
133. M. O. Omisore, O. W. Samuel, and E. J. Atajeromavwo, "A Genetic-Neuro-Fuzzy inferential model for diagnosis of tuberculosis," Applied Computing and Informatics, vol. 13, no. 1, pp. 27-37, 2017.
134. M. A. Khan and F. Algarni, "A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS," in IEEE Access, vol. 8, pp. 122259-122269, 2020, doi: 10.1109/ACCESS.2020.3006424.

135. G. Manogaran, R. Varatharajan, and M. K. Priyan, "Hybrid Recommendation System for Heart Disease Diagnosis based on Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System," Multimedia Tools and Applications, 2018, vol. 77, pp. 4379–4399, doi: https://doi.org/10.1007/s11042-017-5515-y.
136. F. Hu et al., "Walking Compared With Vigorous Physical Activity and Risk of Type 2 Diabetes in Women", Journal of Cardiopulmonary Rehabilitation and Prevention, vol. 282, no. 15, pp. 1433-1439, 1999.
137. Abbasi SH and Ponce De Leon A et al. "Gender Differences in the Risk of Coronary Artery Disease in Iran", Iranian Journal of Public Health, vol 41, no. 3, pp. 36-47, 2012.
138. Azimi S, Khalili D, Hadaegh F, Mehrabi Y, Yavari P, Azizi F. "Direct Estimate of Population Attributable Fraction of Risk Factors for Cardiovascular Diseases: Tehran Glucose and Lipid Study", Iranian Journal of Epidemiology, vol 7, no. 4, pp. 9-18, 2012.
139. N. Pirani and F. F. Khiavi, "Population attributable fraction for cardiovascular diseases risk factors in selected countries: a comparative study," Mater Sociomed, vol. 29, no. 1, pp. 35-39, Mar. 2017.
140. J. T. Salonen et al., "Smoking, blood pressure and serum cholesterol as risk factors of acute myocardial infarction and death among men in Eastern Finland," European Heart Journal, vol. 2, no. 5, pp. 365-373, October 1981.
141. A. Rezaianzadeh et al., "Incidence and risk factors of cardiovascular disease among population aged 40–70 years: a population-based cohort study in the South of Iran," Trop Med Health, vol. 51, p. 35, 2023.
142. S. Rusanov, "Hydrodynamic theory of atherosclerosis formation in humans - Reaction to spasm. Cylindrical cholesterol plaque is the cause of heart attack and stroke", Medical Research Archives, vol. 11, no 3, 2023. ISSN 2375-1924. doi: https://doi.org/10.18103/mra.v11i3.3663.

List of Publications

S. No.	Title of Paper	Author (s)	Name of Journal	Type of Paper	Journal Indexing	Published Date
1	Comparative assessment of Machine-Learning based methodologies and algorithms with accuracy, sensitivity and specificity for prediction of Heart Disease	Rahul Kumar Jha, Santosh Kumar Henge	International Conference on Information and Communication Technology for Competitive Strategies (ICTCS 2019)	Conference Proceeding	Scopus	05-May-20
2	Optimal Machine Learning Classifiers for Prediction of Heart Disease	Rahul Kumar Jha, Santosh Kumar Henge	International Journal of Control and Automation	Full length article	Scopus	25-Mar-20
3	Heart Disease Prediction and	Rahul Kumar Jha, Santosh	International Conference on Intelligent and	Conference	Scopus	24-Aug-21

	Hybrid GANN	Kumar Henge, Ashok Sharma	Fuzzy Systems & INFUS 2021	Proceeding		
4	Neural Fuzzy Hybrid Rule based Inference System with Test Cases for Prediction of Heart Attack Probability	Rahul Kumar Jha, Santosh Kumar Henge, Ashok Sharma, Sanjeev Kumar Mandal, Amit Sharma, Supriya Sharma, Afework Aemro Berhanu, Vijay Kumar	Mathematical Problems in Engineering	Full length article	SCI	29-Sep-2022
5	Personating GA based Neural Fuzzy Hybrid System for Computing	Rahul Kumar Jha, Santosh Kumar Henge, Sanjeev Kumar	International Journal of Advanced Computer Science and Applications	Full length article	SCOPUS	31-July-2023

	HA Probability	Mandal, C Menaka, Deepak Mehta, Aditya Upadhyay, Ashok Kumar Saini, Neha Mishra				