ABSTRACT

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you're born with (congenital heart defects), among others. According to World Health Organization (WHO), cardiovascular disease (CVD) is one of the lethal diseases leads to the most number of deaths worldwide. Cardiovascular disease prediction aids practitioners in making more accurate health decisions for their patients. Early detection can aid people in making lifestyle changes and, if necessary, ensuring effective medical care. Machine learning (ML) is a plausible option for reducing and understanding heart symptoms of disease using the device vital.This project proposes a Light Gradient Boosting Machine (LightGBM)technique as the backbone of computer-aided diagnostic tools for more accurately forecasting heart disease risk levels. LightGBM modeling is a promising classification approach for predicting medication adherence in CVD patients. This predictive model helps stratify the patients so that evidence-based decisions can be made and patients managed appropriately. The chi-square statistical test is performed to select specific attributes from the Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you're born with (congenital heart defects), among others. According to World Health Organization (WHO), cardiovascular disease (CVD) is one of the lethal diseases leads to the most number of deaths worldwide. Cardiovascular disease prediction aids practitioners in making more accurate health decisions for their patients. Early detection can aid people in making lifestyle changes and, if necessary, ensuring effective medical care. Machine learning (ML) is a plausible option for reducing and understanding heart disease.

CHAPTER 1

INTRODUCTION

Cardiovascular System

The cardiovascular system is sometimes called the blood-vascular, or simply the circulatory, system. It consists of the heart, which is a muscular pumping device, and a closed system of vessels called arteries, veins, and capillaries. As the name implies, blood contained in the circulatory system is pumped by the heart around a closed circle or circuit of vessels as it passes again and again through the various "circulations" of the body.

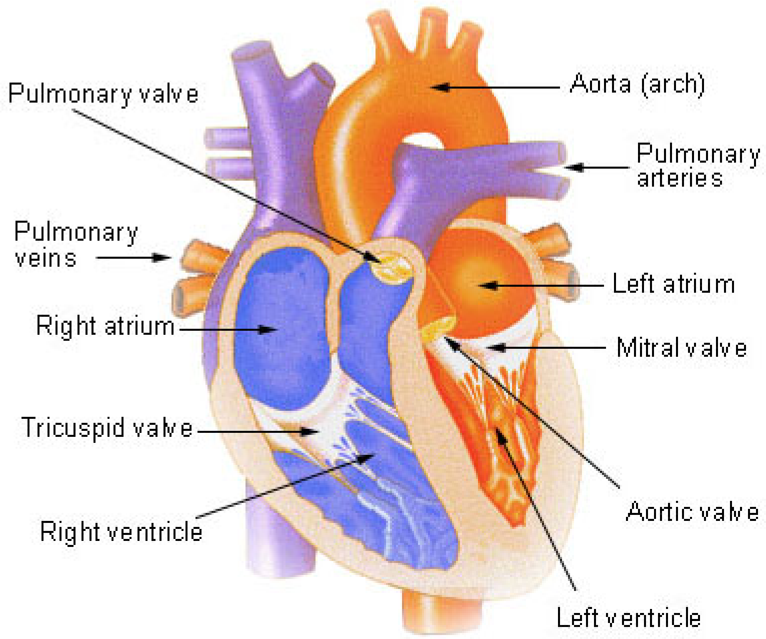
As in the adult, survival of the developing embryo depends on the circulation of blood to maintain homeostasis and a favorable cellular environment. In response to this need, the cardiovascular system makes its appearance early in development and reaches a functional state long before any other major organ system. Incredible as it seems, the primitive heart begins to beat regularly early in the fourth week following fertilization.The vital role of the cardiovascular system in maintaining homeostasis depends on the continuous and controlled movement of blood through the thousands of miles of capillaries that permeate every tissue and reach every cell in the body. It is in the microscopic capillaries that blood performs its ultimate transport function. Nutrients and other essential materials pass from capillary blood into fluids surrounding the cells as waste products are removed.

Numerous control mechanisms help to regulate and integrate the diverse functions and component parts of the cardiovascular system in order to supply blood to specific body areas according to need. These mechanisms ensure a constant internal environment surrounding each body cell regardless of differing demands for nutrients or production of waste products.

Heart

The heart is a muscular pump that provides the force necessary to circulate the blood to all the tissues in the body. Its function is vital because, to survive, the tissues need a continuous supply of oxygen and nutrients, and metabolic waste products have to be removed. Deprived of these necessities, cells soon undergo irreversible changes that lead to death. While blood is the transport medium, the heart is the organ that keeps the blood moving through the vessels. The normal adult heart pumps about 5 liters of blood every minute throughout life. If it loses its pumping effectiveness for even a few minutes, the individual's life is jeopardized.

Internal View of the Heart



Structure of the Heart

The human heart is a four-chambered muscular organ, shaped and sized roughly like a man's closed fist with two-thirds of the mass to the left of midline.The heart is enclosed in a pericardial sac that is lined with the parietal layers of a serous membrane. The visceral layer of the serous membrane forms the epicardium.

Layers of the Heart Wall

Three layers of tissue form the heart wall. The outer layer of the heart wall is the epicardium, the middle layer is the myocardium, and the inner layer is the endocardium.

Chambers of the Heart

The internal cavity of the heart is divided into four chambers:

- Right atrium
- Right ventricle
- Left atrium
- Left ventricle

The two atria are thin-walled chambers that receive blood from the veins. The two ventricles are thick-walled chambers that forcefully pump blood out of the heart. Differences in thickness of the heart chamber walls are due to variations in the amount of myocardium present, which reflects the amount of force each chamber is required to generate.

The right atrium receives deoxygenated blood from systemic veins; the left atrium receives oxygenated blood from the pulmonary veins.

Valves of the Heart

Pumps need a set of valves to keep the fluid flowing in one direction and the heart is no exception. The heart has two types of valves that keep the blood flowing in the correct direction. The valves between the atria and ventricles are called atrioventricular valves (also called cuspid valves), while those at the bases of the large vessels leaving the ventricles are called semilunar valves.

The right atrioventricular valve is the tricuspid valve. The left atrioventricular valve is the bicuspid, or mitral, valve. The valve between the right ventricle and pulmonary trunk is the pulmonary semilunar valve. The valve between the left ventricle and the aorta is the aortic semilunar valve.

When the ventricles contract, atrioventricular valves close to prevent blood from flowing back into the atria. When the ventricles relax, semilunar valves close to prevent blood from flowing back into the ventricles.

Coronary Artery Disease

Heart is a strong muscular pump that is responsible for moving about 3,000 gallons of blood through your body every day. Like other muscles, your heart requires a continuous supply of blood to work properly. Your heart muscle gets the blood it needs to do its job from the coronary arteries.

What is coronary artery disease?

Coronary artery disease is the narrowing or blockage of the coronary arteries, usually caused by atherosclerosis. Atherosclerosis (sometimes called "hardening" or "clogging" of the arteries) is the buildup of cholesterol and fatty deposits (called plaques) on the inner walls of the arteries. These plaques can restrict blood flow to the heart muscle by physically clogging the artery or by causing abnormal artery tone and function.Without an adequate blood supply, the heart becomes starved of oxygen and the vital nutrients it needs to work properly. This can cause chest pain called angina. If the blood supply to a portion of the heart muscle is cut off entirely, or if the energy demands of the heart become much greater than its blood supply, a heart attack (injury to the heart muscle) may occur.

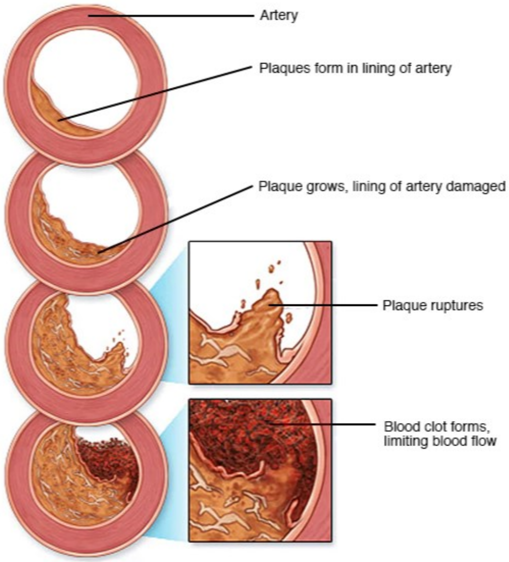
Symptoms

If your coronary arteries narrow, they can't supply enough oxygen-rich blood to your heart — especially when it's beating hard, such as during exercise. At first, the decreased blood flow may not cause any coronary artery disease symptoms. As plaque continues to build up in your coronary arteries, however, you may develop coronary artery disease signs and symptoms, including:

- Chest pain (angina). You may feel pressure or tightness in your chest, as if someone were standing on your chest. This pain, referred to as angina, usually occurs on the middle or left side of the chest. Angina is generally triggered by physical or emotional stress.
- The pain usually goes away within minutes after stopping the stressful activity. In some people, especially women, this pain may be fleeting or sharp and felt in the neck, arm or back.
- Shortness of breath. If your heart can't pump enough blood to meet your body's needs, you may develop shortness of breath or extreme fatigue with exertion.
- Heart attack. A completely blocked coronary artery will cause a heart attack. The classic signs and symptoms of a heart attack include crushing pressure in your chest and pain in your shoulder or arm, sometimes with shortness of breath and sweating.

Women are somewhat more likely than men are to experience less typical signs and symptoms of a heart attack, such as neck or jaw pain. Sometimes a heart attack occurs without any apparent signs or symptoms.

Causes


Artery
Plaques form in lining of artery
Plaque grows, lining of artery damaged
Plaque ruptures
Blood clot forms, limiting blood flow

Development of atherosclerosis

Atherosclerosis is a process in which blood, fats such as cholesterol and other substances build up on your artery walls. Eventually, deposits called plaques may form. The deposits may narrow or block your arteries. These plaques can also rupture, causing a blood clot.

Coronary artery disease is thought to begin with damage or injury to the inner layer of a coronary artery, sometimes as early as childhood. The damage may be caused by various factors, including:

- Smoking
- High blood pressure
- High cholesterol
- Diabetes or insulin resistance
- Sedentary lifestyle

Once the inner wall of an artery is damaged, fatty deposits (plaque) made of cholesterol and other cellular waste products tend to accumulate at the site of injury in a process called atherosclerosis. If the surface of the plaque breaks or ruptures, blood cells called platelets will clump at the site to try to repair the artery. This clump can block the artery, leading to a heart attack.

Problem Identified

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either it is expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today’s world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

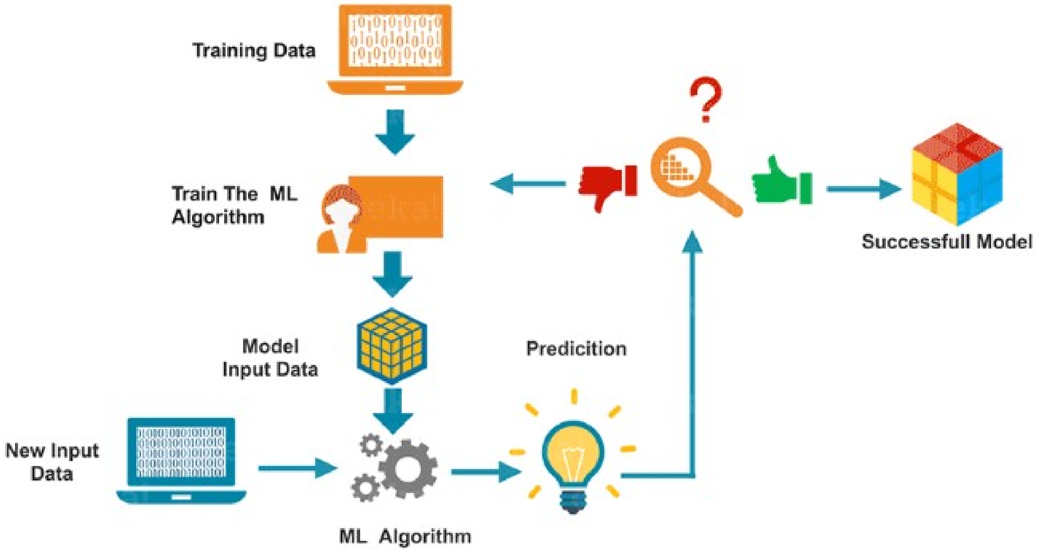
We propose themachine learning algorithm Support Vector Machine (SVM) for CHD interpretation. As is known to all, SVM can learn a hierarchical feature representation from raw data automatically, so it does need any handcrafted features by experts.

Machine Learning

Machine learning is a branch of AI. Other tools for reaching AI include rule-based engines, evolutionary algorithms, and Bayesian statistics. While many early AI programs, like IBM's Deep Blue, which defeated Garry Kasparov in chess in 1997, were rule-based and dependent on human programming, machine learning is a tool through which computers have the ability to teach themselves, and set their own rules. In 2016, Google's DeepMind beat the world champion in Go by using machine learning–training itself on a large data set of expert moves.

Machine Learning Work

Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it makes a prediction on the basis of the model.

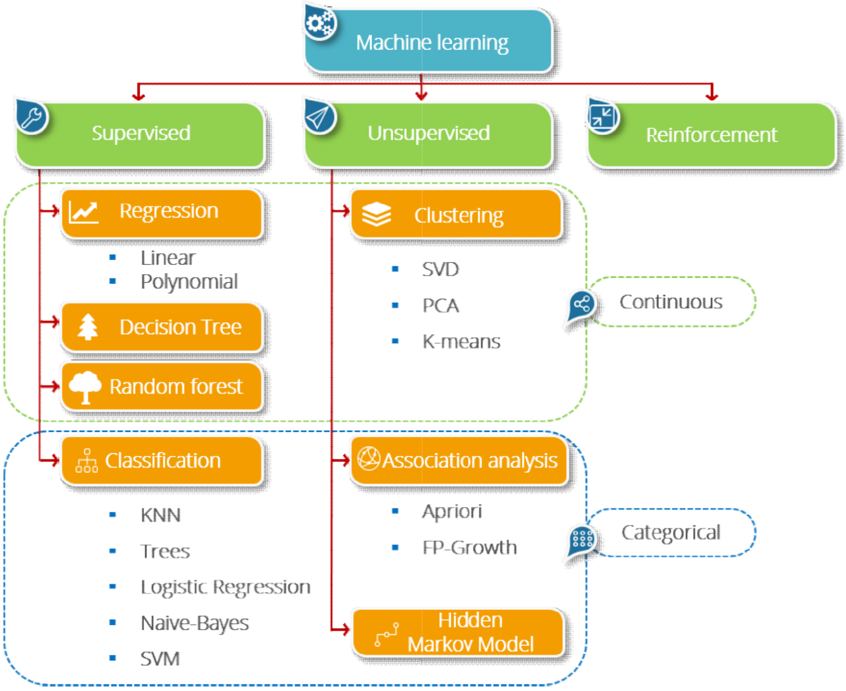


The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is trained again and again with an augmented training data set. This is just a very high-level example as there are many factors and other steps involved.

Types of Machine Learning

Machine learning is sub-categorized to three types:

- Supervised Learning – Train Me!
- Unsupervised Learning – I am self-sufficient in learning
- Reinforcement Learning – My life My rules! (Hit & Trial)



Supervised Learning: More Control, Less Bias

Supervised machine learning algorithms apply what has been learned in the past to new data using labelled examples to predict future events. By analysing a known training dataset, the learning algorithm produces an inferred function to predict output values. The system can provide targets for any new input after sufficient training. It can also compare its output with the correct, intended output to find errors and modify the model accordingly.

Unsupervised Learning: Speed and Scale

Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. At no point does the system know the correct output with certainty. Instead, it draws inferences from datasets as to what the output should be.

Reinforcement Learning: Rewards Outcomes

Reinforcement machine learning algorithms are a learning method that interacts with its environment by producing actions and discovering errors or rewards. The most relevant characteristics of reinforcement learning are trial and error search and delayed reward. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context to maximize its performance. Simple reward feedback — known as the reinforcement signal — is required for the agent to learn which action is best.

Machine Learning Use Cases

Companies are already using machine learning in several ways, including:

Recommendation algorithms: The recommendation engines behind Netflix and YouTube suggestions, what information appears on your Facebook feed, and product recommendations are fuelled by machine learning. “They want to learn, like on Twitter, what tweets we want them to show us, on Facebook, what ads to display, what posts or liked content to share with us.”

Image analysis and object detection: Machine learning can analyse images for different information, like learning to identify people and tell them apart — though facial recognition algorithms are controversial. Business uses for this vary. Shulman noted that hedge funds famously use machine learning to analyse the number of cars in parking lots, which helps them learn how companies are performing and make good bets.

Fraud detection: Machines can analyse patterns, like how someone normally spends or where they normally shop, to identify potentially fraudulent credit card transactions, log-in attempts, or spam emails.

Automatic helplines or chatbots: Many companies are deploying online chatbots, in which customers or clients don't speak to humans, but instead interact with a machine. These algorithms use machine learning and natural language processing, with the bots learning from records of past conversations to come up with appropriate responses.

Self-driving cars: Much of the technology behind self-driving cars is based on machine learning, deep learning in particular.

Medical imaging and diagnostics: Machine learning programs can be trained to examine medical images or other information and look for certain markers of illness, like a tool that can predict cancer risk based on a mammogram.

Objective

The objective of developing a machine learning intelligence framework for heart disease diagnosis is to potentialize the system in predicting the heart disease in order to increase the survival rate of patients by the accurate, precise and early detection of disease.

CHAPTER 2

SYSTEM ANALYSIS

Existing System

- **Naive Bayes**

The Naïve Bayes Classifier technique is mainly applicable when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model recognizes the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Naive Bayes or Bayes' Rule is the basis for many machine learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data.

- **Logistic Regression**

Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used to for prediction and also calculating the probability of success.

- **Decision Tree**

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making.

- **Random Forest**

  Random Forest is essentially an ensemble of unpruned classification trees. It gives excellent performance on a number of practical problems, largely because it is not sensitive to noise in the data set, and it is not subject to overfitting. It works fast, and generally exhibits a substantial performance improvement over many other tree-based algorithms. Random forests are built by combining the predictions of several trees, each of which is trained in isolation. There are three main choices to be made when constructing a random tree.

- **XGboost**

  eXtreme Gradient Boosting (Xgboost)- Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The name xgboost, though, actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms.

Disadvantages

- One algorithm may work well on a specific dataset while it cannot show a good performance on some others.
- So, selecting a suitable algorithm for a specific dataset is a big challenge in bioinformatics.
- Consequently, selecting good feature selection or classification algorithms is also a big challenge in this field.
- ML/DM algorithms commonly need massive datasets to be trained.
- These datasets must be inclusive and unbiased with high quality.
- Datasets also need time to be collected.

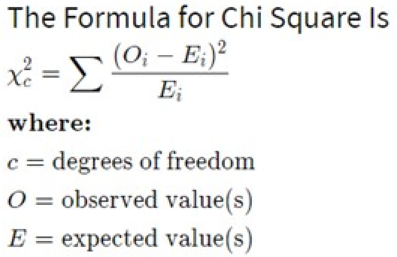
- ML/DM algorithms need time to be trained and tested enough to be able to generate results with high confidence.
- These algorithms need a lot of resources and equipment.
- ML/DM algorithms face the verification problem. It is difficult to prove that the prediction made by them work correctly for all scenarios.
- Another disadvantage of ML/DM algorithm is their high error-susceptibility. If they are trained with biased or incorrect data, they end up with imprecise outputs.
- This may lead to a chain of errors that mislead treatment methods. When these errors get noticed, it takes some times to diagnose the source of these errors and even needs more time to correct them.

Proposed System

The proposed diagnostic system used

- **x2 statistical model for features refinement**

A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.



The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

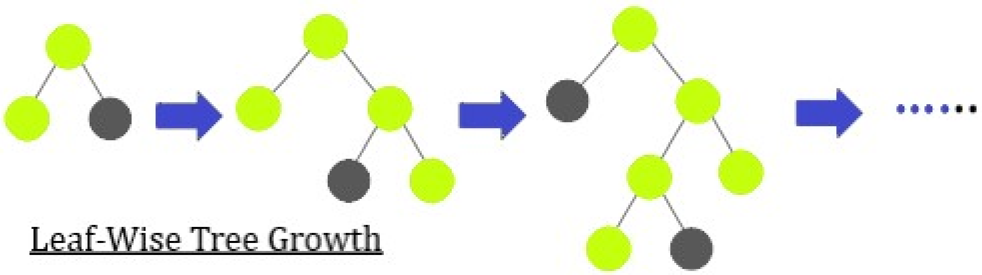
c = degrees of freedom

O = observed value(s)

E = expected value(s)

- **LightGBM**

LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.LightGBM is called “Light” because of its computation power and giving results faster. It takes less memory to run and is able to deal with large amounts of data. It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks.



The two techniques of GOSS and EFB described below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks

Gradient-based One Side Sampling Technique for LightGBM:

Different data instances have varied roles in the computation of information gain. The instances with larger gradients (i.e., under-trained instances) will contribute more to the information gain. GOSS keeps those instances with large gradients (e.g., larger than a predefined threshold, or among the top percentiles), and only randomly drop those instances with small gradients to retain the accuracy of information gain estimation. This treatment can lead to a more accurate gain estimation than uniformly random sampling, with the same target sampling rate, especially when the value of information gain has a large range.

LightGBM is not for a small volume of datasets. It can easily overfit small data due to its sensitivity. It can be used for data having more than 10,000+ rows. There is no fixed threshold that helps in deciding the usage of LightGBM. It can be used for large volumes of data especially when one needs to achieve a high accuracy.

LightGBM Parameters

It is very important to get familiar with basic parameters of an algorithm that you are using. LightGBM has more than 100 parameters that are given in the documentation of LightGBM, but there is no need to study all of them. Let us see what are such different parameters.

- **Control Parameters**

Max depth: It gives the depth of the tree and also controls the overfitting of the model. If you feel your model is getting over fitted lower down the max depth.

Min_data_in_leaf: Leaf minimum number of records also used for controlling overfitting of the model.

Feature_fraction: It decides the randomly chosen parameter in every iteration for building trees. If it is 0.7 then it means 70% of the parameter would be used.

Bagging_fraction: It checks for the data fraction that will be used in every iteration. Often, used to increase the training speed and avoid overfitting.

Early_stopping_round: If the metric of the validation data does show any improvement in last early_stopping_round rounds. It will lower the imprudent iterations.

Lambda: It states regularization. Its values range from 0 to 1.

Min_gain_to_split: Used to control the number of splits in the tree.

- **Core Parameters**

Task: It tells about the task that is to be performed on the data. It can either train on the data or prediction on the data.

Application: This parameter specifies whether to do regression or classification.LightGBM default parameter for application is regression.

Binary: It is used for binary classification.

Multiclass: It is used for multiclass classification problems.

Regression: It is used for doing regression.

Boosting: It specifies the algorithm type.

rf : Used for Random Forest.

Goss: Gradient-based One Side Sampling.

Num_boost_round: It tells about the boosting iterations.

Learning_rate: The role of learning rate is to power the magnitude of the changes in the approximate that gets updated from each tree's output. It has values: 0.1,0.001,0.003.

Num_leaves: It gives the total number of leaves that would be present in a full tree, default value: 31

- **Metric Parameter**

It takes care of the loss while building the model. Some of them are stated below for classification as well as regression.

Mae: Mean absolute error.

Mse: Mean squared error.

Binary_logloss: Binary Classification loss.

Multi_logloss: Multi Classification loss.

- **Parameter Tuning**

Parameter Tuning is an important part that is usually done by data scientists to achieve a good accuracy, fast result and to deal with overfitting. Let us see quickly some of the parameter tuning you can do for better results.

num_leaves: This parameter is responsible for the complexity of the model. Its values should be ideally less than or equal to 2. If its value is more it would result in overfitting of the model.

Min_data_in_leaf: Assigning bigger value to this parameter can result in under fitting of the model. Giving it a value of 100 or 1000 is sufficient for a large dataset.

Max_depth: To limit the depth of the tree max_depth is used.

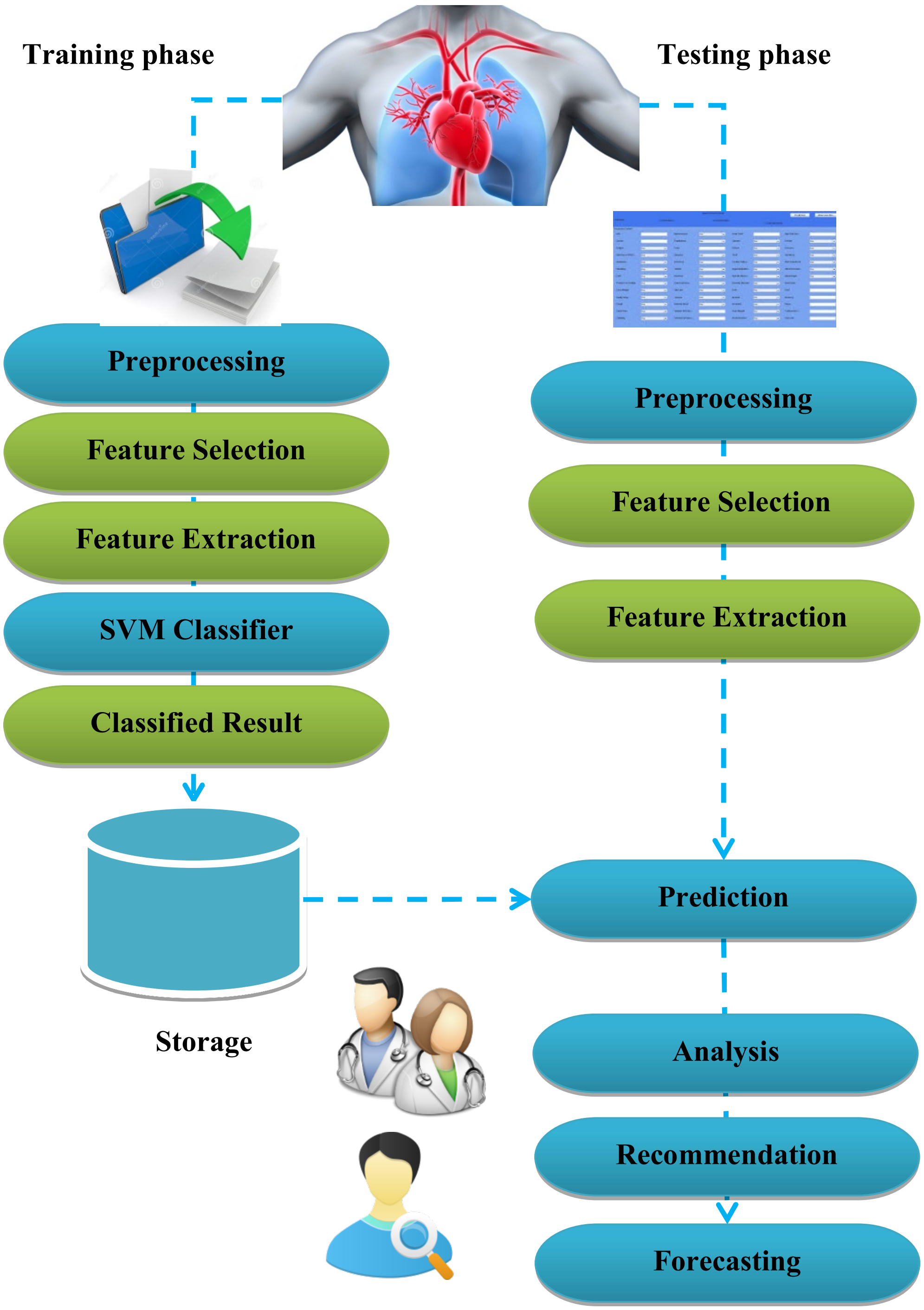
Advantages

- It may result in early detection that leads to a decrease in mortality rate.
- ML can provide a priori probability of disease and use this probability to selectively target patients for angiography. Tis can save in cost and time for other patients. The side effects of angiography are also eliminated for them.
- Using ML can extract hidden patterns in the collected data. Tis may lead to finding new methods for early detection in many diseases like CAD.
- This is a novel method that builds on current research to derive quick and precise diagnostics.
- The method significantly outperforms other published research in this area due to its superior accuracy.
- intention is to accurately classify the presence of diseases.
- sufficiently sensitive and specific to correctly identify unhealthy patients as abnormal while ensuring healthy patients are not misclassified.
- In the case of deep learning, the accuracy, sensitivity and specificity of these systems may equal or even surpass human experts.

CHAPTER 3

SYSTEM DESIGN

System Architecture



CHAPTER 5

SYSTEM IMPLEMENTATION

Dataset Descript

Dataset Introduction

In this project, we collected a heart disease dataset known as Cleveland heart disease database from an online machine learning and data mining repository of the University of California, Irvine (UCI). it covers the role of various subsystems/modules/classes along with implementation details listing of the code for the major functionalities.

Database schema

Table Database Description

Field	**Description**	**Range and Values**
Age	Age of the patient	0-100 in years
Sex	Gender of the patient	0-1 (1:Male 0:Female)
Chest Pain	Type of chest pain	1-4 (1: Typical Angina, 2: Atypical Angina, 3: Non-anginal, 4: Asymptotic)
Resting Blood Pressure	Blood pressure during rest	mm Hg
Cholesterol	Serum Cholesterol	mg / dl
Fasting Blood Sugar	Blood sugar content before food intake if >120 mg/dl	0-1 (0: False, 1: True)
ECG	Resting Electrocardiographic results	0-1 (0: Normal, 1: Having ST-T wave)
Max Heart Rate	Maximum heart beat rate.	Beats/min
Exercise	Has pain been	0-1 (0: No, 1: Yes)

Induced Angina	induced by exercise	
Old Peak	ST depression induced by exercise relative to rest	0-4
Slope of Peak Exercise	Slope of the peak exercise ST segment	1-3 (1: Up sloping, 2: Flat, 3: Down sloping)
Ca	Number of vessels colored by fluoroscopy	0-3
Thal		3- normal 6-Fixed Defect 7- Reversible Defect
Num	Diagnostics of Heart Disease	0-1 (0: <50% Narrowing 1: >50% Narrowing)

The dataset consists of 303 individuals data. There are 14 columns in the dataset, which are described below.

1. ***Age***: displays the age of the individual.
2. ***Sex***: displays the gender of the individual using the following format :
   1 = male
   0 = female
3. ***Chest-pain type***: displays the type of chest-pain experienced by the individual using the following format :
   1 = typical angina
   2 = atypical angina
   3 = non — anginal pain
   4 = asymptotic
4. ***Resting Blood Pressure***: displays the resting blood pressure value of an individual in mmHg (unit)

5. ***Serum Cholestrol***: displays the serum cholesterol in mg/dl (unit)
6. ***Fasting Blood Sugar***: compares the fasting blood sugar value of an individual with 120mg/dl.
   If fasting blood sugar > 120mg/dl then : 1 (true)
   else : 0 (false)
7. ***Resting ECG*** : displays resting electrocardiographic results
   0 = normal
   1 = having ST-T wave abnormality
   2 = left ventricular hyperthrophy
8. ***Max heart rate achieved*** : displays the max heart rate achieved by an individual.
9. ***Exercise induced angina*** :
   1 = yes
   0 = no
10. ***ST depression induced by exercise relative to rest***: displays the value which is an integer or float.
11. ***Peak exercise ST segment*** :
    1 = upsloping
    2 = flat
    3 = downsloping
12. ***Number of major vessels (0–3) colored by flourosopy*** : displays the value as integer or float.
13. ***Thal*** : displays the thalassemia :
    3 = normal
    6 = fixed defect
    7 = reversible defect
14. ***Diagnosis of heart disease*** : Displays whether the individual is suffering from heart disease or not :
    0 = absence
    1, 2, 3, 4 = present.

5.1.3. Why these parameters:

In the actual dataset, we had 76 features but for our study, we chose only the above 14 because:

1. **Age**: Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.
2. **Sex**: Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman's risk is similar to a man's although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.
3. **Angina (Chest Pain)**: Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.
4. **Resting Blood Pressure**: Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.
5. **Serum Cholesterol**: A high level of low-density lipoprotein (LDL) cholesterol (the "bad" cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers your risk of a heart attack.
6. **Fasting Blood Sugar**: Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body's blood sugar levels to rise, increasing your risk of a heart attack.
7. **Resting ECG**: For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening

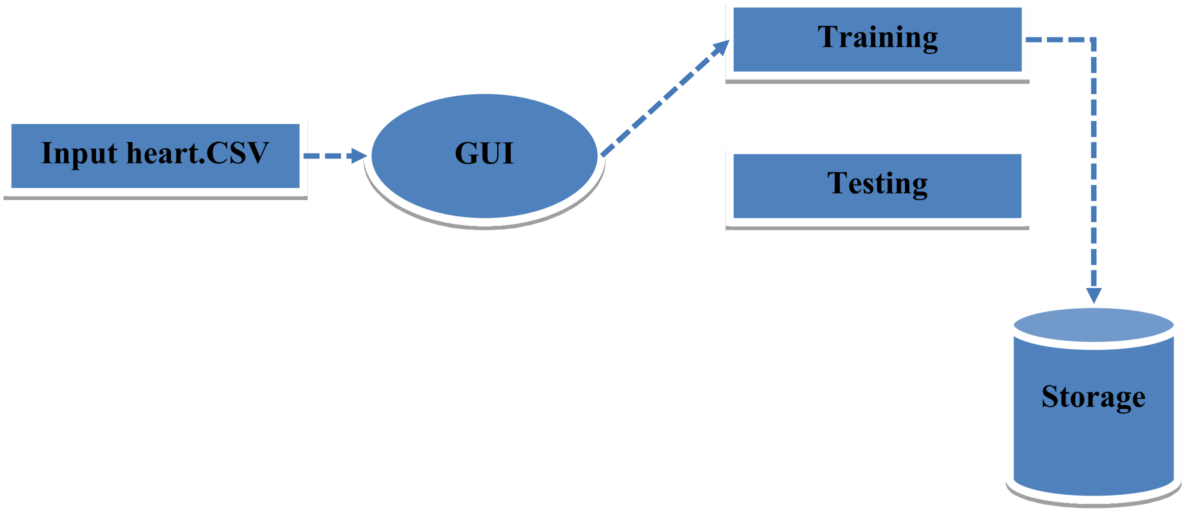
with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.

8. **Max heart rate achieved**: The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.
9. **Exercise induced angina**: The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the center of your chest but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands. o Types of Angina a. Stable Angina / Angina Pectoris b. Unstable Angina c. Variant (Prinzmetal) Angina d. Microvascular Angina.
10. **Peak exercise ST segment**: A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an ‘equivocal’ test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.

Problem Description

Data Set Acquisition

Data set used for implementation of proposed work is taken from UCI repository (https://archive.ics.uci. Edu/ml/datasets/Heart Disease). Nature of the data set is defined by its dimensionality. It consists of 13 columns which represents features of the data set. A complete overview has been shown in the figure given below as like other data set it may consist of certain redundancy, noise and missing values for talking case of it proper steps need to be taken. In next subsection pre-processing is explained in brief. Attributes of dataset need to be taken care of and their respective value need to standardize.



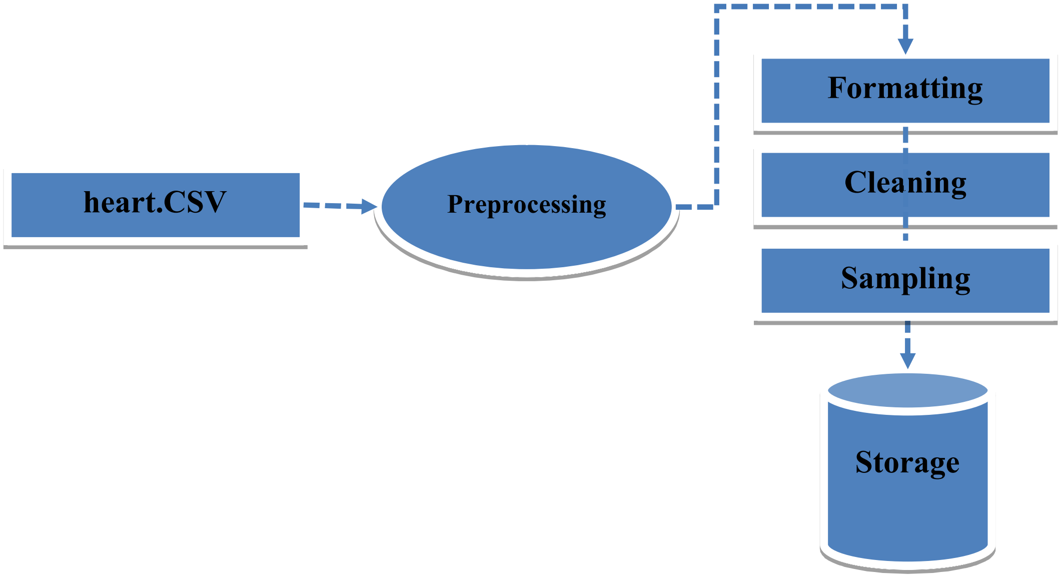
Data set Preprocessing

Every dataset consists of various types of anomalies such as missing values, redundancy orany other problem for removing this problem there is need of certain step called as processing data. Pre-processing step is needed to overcome from such problem. There arethree pre-processing steps:

1. Formatting: The data set is used for implementation is taken from UCI repository, it may contain certain attributes whose names are not clear in the (dataset name) also contain certain unrelated attribute which is not useful for the greater performance of proposed work . An attribute name as “Thal” has been removed from dataset by using following command in R, Dataset$Thal<-Null

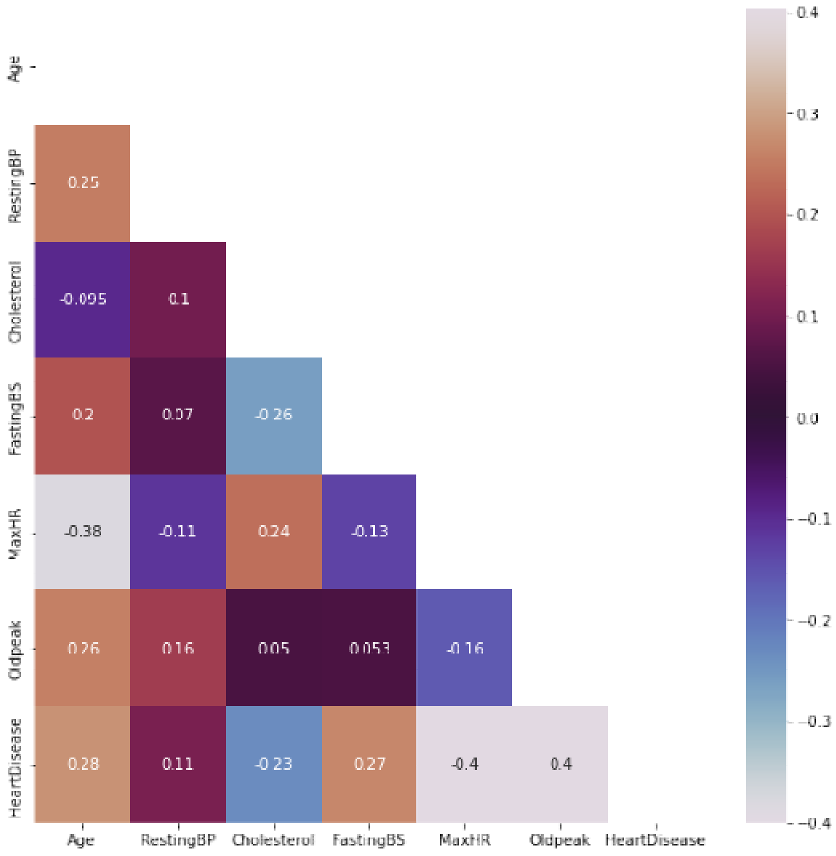
2. Cleaning: This part of pre-processing belongs to remove or fixing of missing out entry in the data frame. Row containing these incomplete columned to be removed also for removing certain redundant entries in data frame this step is recommend

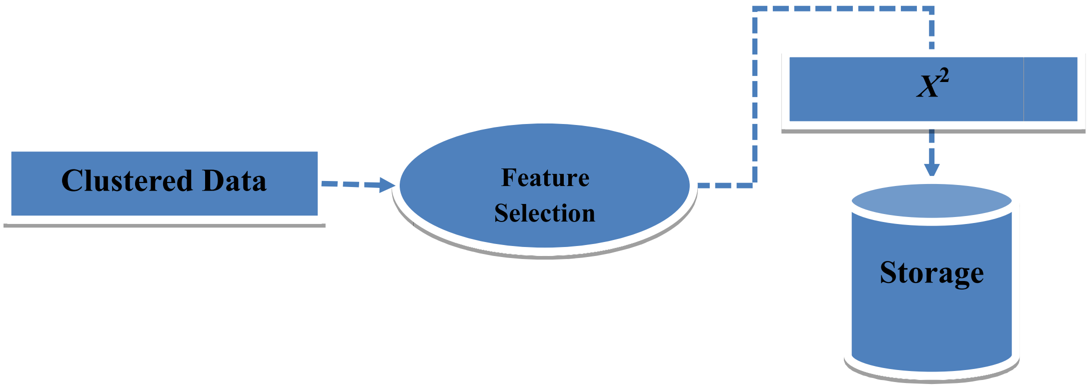
3. Sampling: Sampling is also done on the dataset to enhance the performance of the algorithm on sample data set may lead algorithm to take longer time.



Feature Selection

In feature selection, irrelevant features areeliminated and the most important or relevant features areapplied to the network. Thus, if we supply all features toLightGBM, some features may be noisy and if they are learned inthe training process, they may degrade generalization of thenetwork although the network will show good performanceon the training data. That is why large number of features arealso considered one of the main causes of over fitting. Thus,searching out optimal subset of features by eliminating noisyfeatures can help LightGBM to show good performance on bothtraining and testing data.In this module, we use X^2 statistical model to eliminate irrelevantfeatures. In the feature's elimination process, we computeX^2 statistics between each non-negative feature F_i andclass i.e., y. The X^2 model performs X^2 test that measuresdependence between the features and class. Hence, the modelis capable of eliminating those features which are more likelyto be independent of class. Because, these features can beregarded as irrelevant forclassification.





LightGBM CVDClassification

The LightGBM is a gradient-based boosting approach which makes use of tree-based learning methods. LightGBM is a machine learning algorithm framework based on boosting ensemble learning technology proposed by Microsoft in 2017. Compared with GBDT (Gradient Boosting Decision Tree) algorithm and XGBoost algorithm, lightGBM algorithm can not only achieve the same prediction performance, but also has more obvious advantages in training speed and memory consumption. On the basis of GBDT algorithm and XGBoost algorithm, lightGBM algorithm integrates several optimization strategies, such as histogram algorithm, grandient-based one-side sampling (GOSS) algorithm, exclusive feature bundling (EFB) strategy, leaf-wise strategy, supporting category feature strategy and supporting efficient parallel strategy. These optimization strategies make lightGBM become a classification model with high efficiency, low consumption, more accuracy and more convenient.

Weight	Feature
0.0838 ± 0.0122	num_major_vessels
0.0676 ± 0.0337	chest_pain_type
0.0171 ± 0.0424	st_slope
0.0018 ± 0.0177	max_heart_rate_achieved
0 ± 0.0000	thalassemia
0 ± 0.0000	exercise_induced_angina
0 ± 0.0000	rest_electrocardiographic

Weight	Feature
0 ± 0.0000	fasting_blood_sugar
0 ± 0.0000	resting_blood_pressure
0 ± 0.0000	sex
-0.0045 ± 0.0057	st_depression
-0.0063 ± 0.0092	cholesterol
-0.0072 ± 0.0044	age

The parameter boosting type was set to GBDT (Gradient Boosting Decision Tree) algorithm when we used the lightGBM algorithm. The method where lightGBM algorithm plays the function of feature combination for lr classifier is to use the output of lightGBM algorithm as the input of lr algorithm for training to obtain the final prediction output. This is the application of stacking ensemblelearning technology. It is worth noting that the input of lr algorithm is not the category labels (such as 0 or 1) predicted by lightGBM algorithm for each instance, but the index of leaf node of each instance on each decision tree.

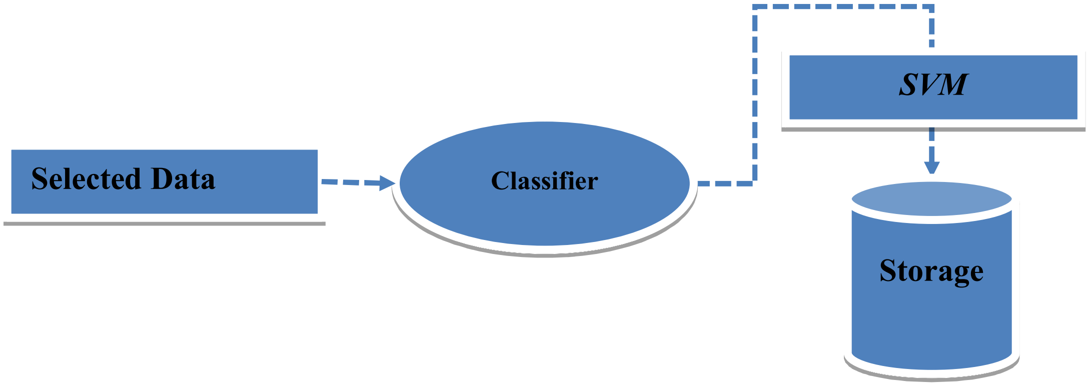
These indexes need to be processed by One-Hot Encoding before inputting lr algorithm. The specific process is as follows:

(1) LightGBM model is applied to the training set, and the trained LightGBM classifier is obtained. (2) After the training, the index of leaf node on each decision tree in the lightGBM model is output for each instance of the training set. After all iterations, all indexes of each instance form a new set of features.

At this time, a set of m × n-dimensional dataset is formed, where m is the size of the training set, and n is the number of weak estimators (decision trees) in the lightGBM model. (3) The m × n-dimensional dataset obtained in step (2) is encoded by One-Hot Encoding, and a sparse matrix Mtrain with m × n × l-dimensional is obtained, where l is the number of leaf nodes per decision tree in the lightGBM model. The sparse matrix Mtrain is the training set of lr algorithm.

(4) lr model is applied to the sparse matrix Mtrain, and the trained lr classifier is obtained.

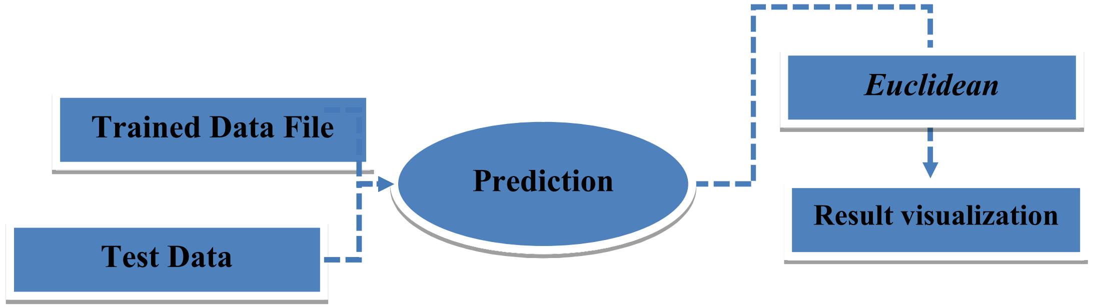
(5) Similarly, the testing set is processed by steps (1)–(3) to obtain the sparse matrix Mtest, and the sparse matrix Mtest is entered into the trained lr classifier to obtain the final prediction.



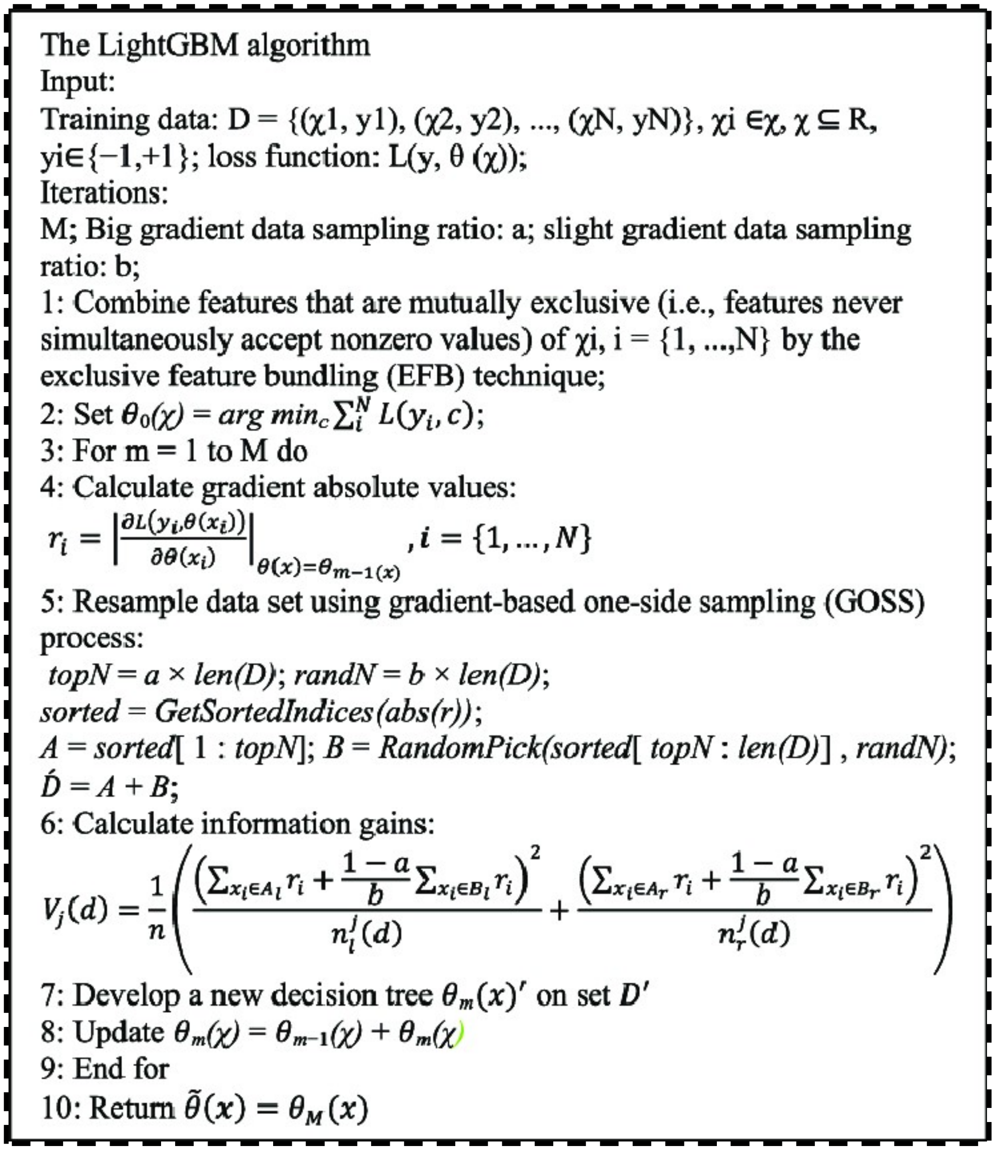
Prediction

In this module, the Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric.

The Euclidean distance between points p and q is thelength of the line segment connecting them: p.qIn Cartesian coordinates,if p = (p , p ,..., p) and q = (q , q ,..., q) are two points inEuclidean n space, then the distance from p to q, or from q to p is given by the following heterogenous value difference metric.



Algorithm – LightGBM



```
The LightGBM algorithm
Input:
Training data: D = {(χ1, y1), (χ2, y2), ..., (χN, yN)}, χi ∈χ, χ ⊆ R,
yi∈{−1,+1}; loss function: L(y, θ (χ));
Iterations:
M; Big gradient data sampling ratio: a; slight gradient data sampling
ratio: b;
1: Combine features that are mutually exclusive (i.e., features never
simultaneously accept nonzero values) of χi, i = {1, ...,N} by the
exclusive feature bundling (EFB) technique;
2: Set $\theta_0(\chi) = arg\ min_c \sum_i^N L(y_i, c)$;
3: For m = 1 to M do
4: Calculate gradient absolute values:
$r_i = \left|\frac{\partial L(y_i,\theta(x_i))}{\partial\theta(x_i)}\right|_{\theta(x)=\theta_{m-1(x)}}, i = \{1, \dots, N\}$
5: Resample data set using gradient-based one-side sampling (GOSS)
process:
 $topN = a \times len(D)$; $randN = b \times len(D)$;
$sorted = GetSortedIndices(abs(r))$;
$A = sorted[\,1 : topN]$; $B = RandomPick(sorted[\,topN : len(D)]\,, randN)$;
$\acute{D} = A + B$;
6: Calculate information gains:
$V_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_l} r_i + \frac{1-a}{b}\sum_{x_i\in B_l} r_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i\in A_r} r_i + \frac{1-a}{b}\sum_{x_i\in B_r} r_i\right)^2}{n_r^j(d)}\right)$
7: Develop a new decision tree $\theta_m(x)'$ on set $D'$
8: Update $\theta_m(\chi) = \theta_{m-1}(\chi) + \theta_m(\chi)$
9: End for
10: Return $\tilde{\theta}(x) = \theta_M(x)$
```

CHAPTER 6

SYSTEM REQUIREMENTS

Hardware Requirements

- Processors:
  - Intel® Core™ i5 processor 4300M at 2.60 GHz or 2.59 GHz (1 socket, 2 cores, 2 threads per core), 8 GB of DRAM
- Disk space: 320 GB
- Operating systems: Windows® 10, macOS*, and Linux*

Software Requirements

- **Server Side**: Python 3.7.4(64-bit) or (32-bit)
- **Client Side** : HTML, CSS, Bootstrap
- **IDE** : Flask 1.1.1
- **Back end** : MySQL 5.
- **Server** : Wampserver 2i
- **OS** : Windows 10 64 –bit or Ubuntu 18.04 LTS “Bionic Beaver”

LANGUAGE SPECIFICATION

Python 3.7.4

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). This tutorial gives enough understanding on Python programming language.

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages. Python is a MUST for students and working professionals to become a great Software Engineer specially when they are working in Web Development Domain.

Python is currently the most widely used multi-purpose, high-level programming language. Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc. The biggest strength of Python is huge collection of standard library which can be used for the following:

- Machine Learning
- GUI Applications (like Kivy, Tkinter, PyQtetc.)
- Web frameworks like Django (used by YouTube, Instagram, Dropbox)
- Image processing (like OpenCV, Pillow)
- Web scraping (like Scrapy, BeautifulSoup, Selenium)
- Test frameworks
- Multimedia
- Scientific computing
- Text processing and many more..

Pandas

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

Pandas is mainly used for data analysis and associated manipulation of tabular data in Data frames. Pandas allows importing data from various file formats such as comma-separated values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features. The development of pandas introduced into Python many comparable features of working with Data frames that were established in the R programming language. The panda's library is built upon another library NumPy, which is oriented to efficiently working with arrays instead of the features of working on Data frames.

NumPy

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

Seaborn

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.Visualization is the central part of Seaborn which helps in exploration and understanding of data.

Seaborn offers the following functionalities:

- Dataset oriented API to determine the relationship between variables.
- Automatic estimation and plotting of linear regression plots.
- It supports high-level abstractions for multi-plot grids.
- Visualizing univariate and bivariate distribution.

Scikit Learn

scikit-learn is a Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license.

Scikit-learn (formerly scikits. learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

MySQL

MySQL tutorial provides basic and advanced concepts of MySQL. Our MySQL tutorial is designed for beginners and professionals. MySQL is a relational database management system based on the Structured Query Language, which is the popular language for accessing and managing the records in the database. MySQL is open-source and free software under the GNU license. It is supported by Oracle Company.MySQL database that provides for how to manage database and to manipulate data with the help of various SQL queries. These queries are: insert records, update records, delete records, select records, create tables, drop tables, etc. There are also given MySQL interview questions to help you better understand the MySQL database.

MySQL is currently the most popular database management system software used for managing the relational database. It is open-source database software, which is supported by Oracle Company. It is fast, scalable, and easy to use database management system in comparison with Microsoft SQL Server and Oracle Database. It is commonly used in conjunction with PHP scripts for creating powerful and dynamic server-side or web-based enterprise applications. It is developed, marketed, and supported by MySQL AB, a Swedish company, and written in C programming language and C++ programming language. The official pronunciation of MySQL is not the My Sequel; it is My Ess Que Ell. However, you can pronounce it in your way. Many small and big companies use MySQL. MySQL supports many Operating Systems like Windows, Linux, MacOS, etc. with C, C++, and Java languages.

The Apache Web Server

addition to PHP, MySQL, JavaScript, and CSS, there's actually a fifth hero in the dynamic Web: the web server. In the case of this book, that means the Apache web server. We've discussed a little of what a web server does during the HTTP server/client exchange, but it actually does much more behind the scenes. For example, Apache doesn't serve up just HTML files—it handles a wide range of files, from images and Flash files to MP3 audio files, RSS (Really Simple Syndication) feeds, and more. Each element a web client encounters in an HTML page is also requested from the server, which then serves it up. But these objects don't have to be static files, such as GIF images. They can all be generated by programs such as PHP scripts. That's right: PHP can even create images and other files for you, either on the fly or in advance to serve up later. To do this, you normally have modules either precompiled into Apache or PHP or called up at runtime. One such module is the GD library (short for Graphics Draw), which PHP uses to create and handle graphics.

Apache also supports a huge range of modules of its own. In addition to the PHP module, the most important for your purposes as a web programmer are the modules that handle security. Other examples are the Rewrite module, which enables the web server to handle a varying range of URL types and rewrite them to its own internal requirements, and the Proxy module, which you can use to serve up often-requested pages from a cache to ease the load on the server. Later in the book, you'll see how to actually use some of these modules to enhance the features provided by the core technologies we cover. About Open Source Whether or not being open source is the reason these technologies are so popular has often been debated, but PHP, MySQL, and

Apache are the three most commonly used tools in their categories. What can be said, though, is that being open-source means that they have been developed in the community by teams of programmers writing the features they themselves want and need, with the original code available for all to see and change. Bugs can be found and security breaches can be prevented before they happen. There's another benefit: all these programs are free to use. There's no worrying about having to purchase additional licenses if you have to scale up your website and add more servers. And you don't need to check the budget before deciding whether to upgrade to the latest versions of these products.

WampServer

WampServer is a Windows web development environment. It allows you to create web applications with Apache2, PHP and a MySQL database. Alongside, PhpMyAdmin allows you to manage easily your database.

WAMPServer is a reliable web development software program that lets you create web apps with MYSQL database and PHP Apache2. With an intuitive interface, the application features numerous functionalities and makes it the preferred choice of developers from around the world. The software is free to use and doesn't require a payment or subscription.

Bootstrap 4

Bootstrap is a free and open-source tool collection for creating responsive websites and web applications. It is the most popular HTML, CSS, and JavaScript framework for developing responsive, mobile-first websites.

It solves many problems which we had once, one of which is the cross-browser compatibility issue. Nowadays, the websites are perfect for all the browsers (IE, Firefox, and Chrome) and for all sizes of screens (Desktop, Tablets, Phablets, and Phones). All thanks to Bootstrap developers -Mark Otto and Jacob Thornton of Twitter, though it was later declared to be an open-source project.

Easy to use: Anybody with just basic knowledge of HTML and CSS can start using Bootstrap

Responsive features: Bootstrap's responsive CSS adjusts to phones, tablets, and desktops

Mobile-first approach: In Bootstrap, mobile-first styles are part of the core framework

Browser compatibility: Bootstrap 4 is compatible with all modern browsers (Chrome, Firefox, Internet Explorer 10+, Edge, Safari, and Opera)

Using an IDE

As good as dedicated program editors can be for your programming productivity, their utility pales into insignificance when compared to Integrated Developing Environments (IDEs), which offer many additional features such as in-editor debugging and program testing, as well as function descriptions and much more.

Web Framework

Web Application Framework or simply Web Framework represents a collection of libraries and modules that enables a web application developer to write applications without having to bother about low-level details such as protocols, thread management etc.

Flask

Flask is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.

Flask is often referred to as a micro framework. It aims to keep the core of an application simple yet extensible. Flask does not have built-in abstraction layer for database handling, nor does it have formed a validation support. Instead, Flask supports the extensions to add such functionality to the application. Although Flask is rather young compared to most Python frameworks, it holds a great promise and has already gained popularity among Python web developers. Let's take a closer look into Flask, so-called "micro" framework for Python.

Flask was designed to be easy to use and extend. The idea behind Flask is to build a solid foundation for web applications of different complexity. From then on you are free to plug in any extensions you think you need. Also you are free to build your own modules. Flask is great for all kinds of projects. It's especially good for prototyping.

Flask is part of the categories of the micro-framework. Micro-framework are normally framework with little to no dependencies to external libraries. This has pros and cons. Pros would be that the framework is light, there are little dependency to update and watch for security bugs, cons is that some time you will have to do more work by yourself or increase yourself the list of dependencies by adding plugins. In the case of Flask, its dependencies are:

WSGI-Web Server Gateway Interface (WSGI) has been adopted as a standard for Python web application development. WSGI is a specification for a universal interface between the web server and the web applications.

Werkzeug-It is a WSGI toolkit, which implements requests, response objects, and other utility functions. This enables building a web framework on top of it. The Flask framework uses Werkzeug as one of its bases.

Jinja2 Jinja2 is a popular templating engine for Python. A web templating system combines a template with a certain data source to render dynamic web pages.

CHAPTER 7
SYSTEM TESTING

Testing

In this phase of methodology, testing was carried out on the several application modules. Different kind of testing was done on the modules which are described in the following sections. Generally, tests were done against functional and non-functional requirements of the application following the test cases. Testing the application again and again helped it to become a reliable and stable system.

- **UsabilityTesting**

This was done to determine the usability of the application that was developed. This helped to check whether the application would be easy to use or what pitfalls would the users come through. This was used to determine whether the application is user friendly. It was used to ascertain whether a new user can easily understand the application even before interacting with it so much. The major things checked were: the system flow from one page to another, whether the entry points, icons and words used were functional, visible and easily understood byuser.

- **FunctionalTesting**

Functional Testing is defined as a type of testing which verifies that each function of the software application operates in conformance with the requirement specification. This testing mainly involves black box testing and it is not concerned about the source code of theapplication. Functional tests were done based on different kind of features and modules of the application and observed that whether the features are met actual project objectives and the modules are hundred percent functional. Functional tests, as shown in the following Table-1 to Table-5, were done based on use cases to determine success or failure of the system implementation and design. For each use case, testing measures were set with results being considered successful or unsuccessful. Below are the tables which are showing some of the major test cases along with their respective testresults.

Table 1: Signup/Registration Test Case

Identifier	**Test Case-1**
Test Case	Signup
Description	To register new account in the application.
Pre-requisite	1) Username and email must not exist previously.
Test procedure	1) Select Sign Up from themenu. Fill in username, email, and password and retype password accordingly. 3) Click on Sign Upbutton
Expected Result	1) User can register to the applicationsuccessfully. Username, email and password stored in the user table in the database.
Pass/Fail	Pass

Table 2: Login Test Case

Identifier	**Test Case-2**
Test Case	Login
Description	To login new account in the application
Pre-requisite	1) Registration must be done previously.

Test procedure	1) Select Log In from themenu. 2) Fill in username and passwordaccordingly. 3) Click on Log In button.
Expected Result	1) User can login to the applicationsuccessfully. 2) User should access the application features which areallowed
Pass/Fail	Pass

Table 3: Logout Test Case

Identifier	**Test Case-3**
Test Case	Logout
Description	To log out from the application
Pre-requisite	1) Must exist as logged in user already
Test procedure	1) Select Log Out from the menu.
Expected Result	1) User can logout from the applicationsuccessfully. User should no longer access the application features until he/she again login to theapplication.
Pass/Fail	Pass

Table 4: Model Building Test Case

Identifier	**Test Case-4**
Test Case	Model Building
Description	Machine learning model building for classification
Pre-	1) User is logged in to the system and enter required

requisite	inputs
Test procedure	1) Use LigthGBM to build model
Expected Result	1) An efficient model will be developed.
Pass/Fail	Pass

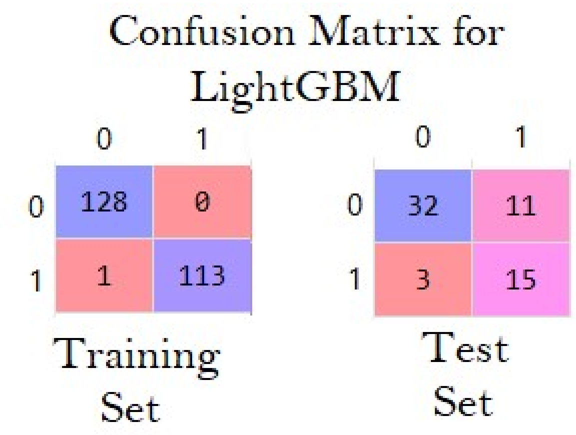
Table 5: Heart Disease Detection Test Case

Identifier	**Test Case-5**
Test Case	Heart Disease Detection
Description	Detect Heart Disease
Pre-requisite	1) User is logged in to the system and enter required inputs
Test procedure	1)Input Symptom parameters 2) Click on the Predictbutton
Expected Result	1) Automatically extract features and classifyit. 2) Show theresult
Pass/Fail	Pass

CHAPTER 8
RESULTS AND DISCUSSION

Evaluation Metrics in order to comprehensively evaluate the classification performance and effectiveness of our proposed method, we applied accuracy, recall, F1 score, precision, specificity, ROC and AUC evaluation metrics. For the sake of expression of the significance and calculation formula of these evaluation metrics, we introduced the confusion matrix (See Table 5) first. The confusion matrix is a specific matrix used to visually present the performance of the algorithm. The confusion matrix of binary classification consists of two rows and two columns. Rows represent the true labels of the two classes in the dataset (denoted y_{true}). Columns represent the predicted label of the two classes acquired by the model (denoted as y_{pre}). As shown in Table 5, the confusion matrix of binary classification includes four indicators: TN, FN, FP and TP. The four indicators are defined as follows. We specified that the label of positive class is 1 and the label of negative class is 0.

- TN (true negative) refers to the number of correctly predicted samples in the samples with the real class label of 0.
- FN (false negative) refers to the number of incorrectly predicted samples in the samples with the real class label of 1.
- FP (false positive) refers to the number of incorrectly predicted samples in the samples the real class label of 0.
- TP (true positive) refers to the number of correctly predicted samples in the samples with the real class label of 1.



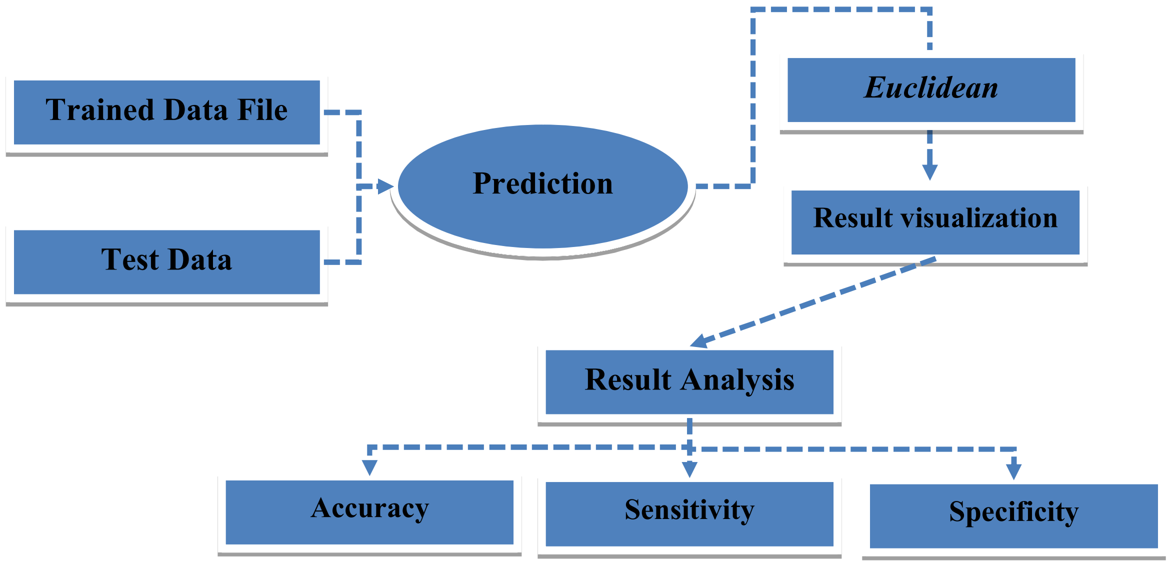
Confusion Matrix for LightGBM

Training Set

	0	1
0	128	0
1	1	113

Test Set

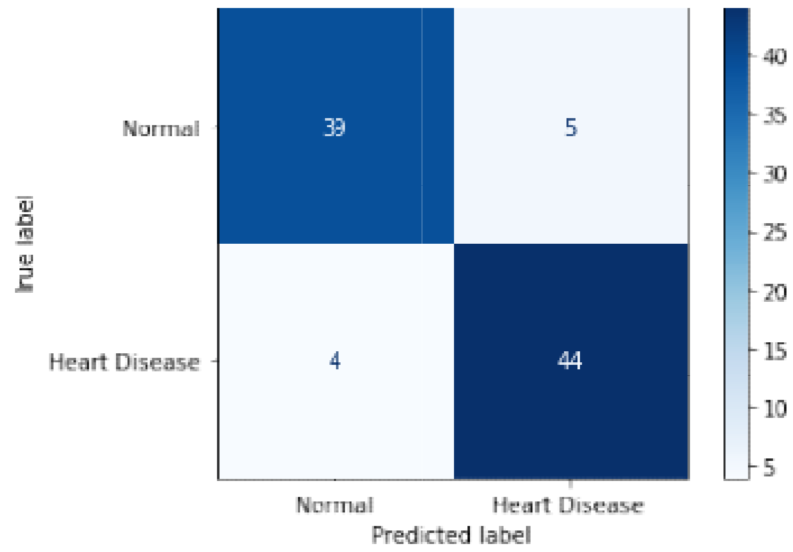
	0	1
0	32	11
1	3	15



Accuracy

Accuracy refers to the proportion of samples that can be correctly predicted by the model in all samples. The calculation equation of accuracy is as follows. TN, TP, FN and FP refer to true negative, true positive, false negative and false positive, respectively.

Accuracy = TN + TP TN + TP + FN + FP (1)



Accuracy is one of the most frequently used and most important model performance evaluation metrics. However, in the dataset with a class imbalance problem, due to the influence of majority class samples, the accuracy is often difficult to accurately measure the classification ability of the model. Therefore, in the dataset with a class imbalance problem, in addition to accuracy, more evaluation indicators need to be applied.

Recall

Recall refers to the proportion of samples that can be correctly predicted by the model in all samples with positive real class labels. Recall is an important indicator to measure the ability of model to identify positive samples. In medical models, it is necessary to pay attention to recall. Recall is calculated according to the following equation:

Recall = TP TP + FN (2)

where TP and FN are true positive and false negative, respectively. In medical application, the cost of undiagnosed positive cases and wrongly diagnosed negative cases is different. The former may cause loss of life, while the latter may lead to excessive treatment. Compared with the former, the latter costs less. At the time of diagnosis, doctors and patients pay more attention to the detection of positive cases. Therefore, the recall is one of the important indicators to judge whether the model can be applied in practice.

Precision

Precision, like recall, is an important indicator to measure the ability of the model to correctly predict positive samples. Precision refers to the proportion of samples with positive real class labels among all samples predicted as positive by the model. According to the definition of precision, its calculation formula is as follows. TP and FP in the formula are true positive and false positive, respectively.

Precision = TP TP + FP (3)

F1 Score

Sometimes the performance of the model evaluated by recall and precision may be show the opposite result, that is, one index has a good result but the other index has a poor result, so the ability of the model cannot be evaluated accurately. Therefore, F1 score is introduced. F1 score combines the results of recall and precision, and is the weighted harmonic mean of recall and precision. Only when the results of recall and precision are good, the F1 score will be higher. The higher the F1 score is, the better the model classification effect is. The following is the calculation formula of F1 score.

F1 = 2 × Precision × Recall Precision + Recall (4)

Specificity

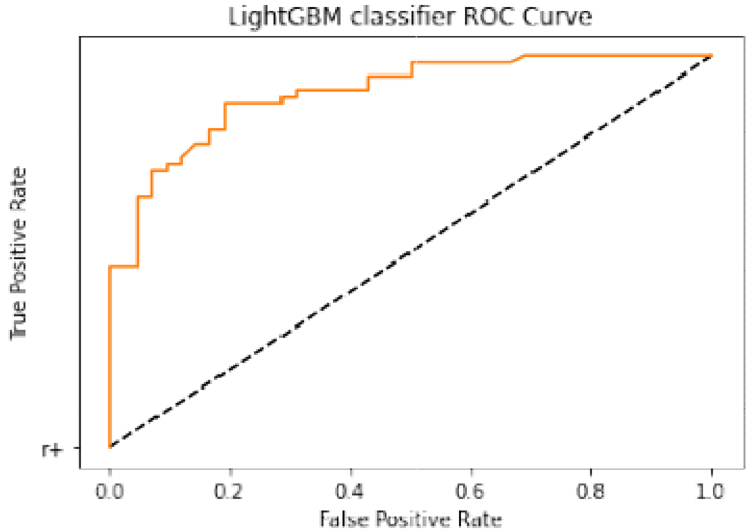
Specificity refers to the proportion of samples that can be correctly predicted by the model in all samples with negative real class labels. Specificity measures the ability of the model to recognize negative samples. Specificity is calculated according to the following formula. TN and FP correspond to true negative and false positive, respectively.

Specificity = TN TN + FP (5)

ROC and AUC

Area under curve (AUC) is the area under the receiver operating characteristic (ROC) curve. The ROC curve is drawn with the false positive rate (FPR) as x-axis and the truepositive rate (TPR) as y-axis. ROC curve intuitively reflects the relationship between specificity and recall. The value of AUC is between 0 and 1, when the value of x-axis (i.e., the false positive rate (FPR) of the model) is closer to 0, and the value of y-axis (i.e., the true positive rate (TPR) of the model) is closer to 1, the value of AUC is closer to 1. The closer the AUC value is to 1, the higher the prediction performance of the classifier.

Classifiers	**Accuracy**	**Recall**	**F1**	**Precision**	**Specificity**	**AUC**
lightGBM	**0.97 ± 0.060**	**0.922 ± 0.068**	**0.980 ± 0.038**	**0.963 ± 0.041**	**0.881 ± 0.095**	**0.93 ± 0.05**



The best model is the LIGHTGBM tree model. The accuracy is 96.7%, with an f1-Score, recall and precision of 97.5%.

CHAPTER 9

CONCLUSION & FUTURE ENHANCEMENT

Conclusion

In this project, the machine learning based support vector machine classification and prediction models were developed and evaluated based on diagnostic performance of coronary heart disease in patients using sensitivity, specificity, precision, FScore, AUC, DOR, 95% confidence interval for DOR, and K-S test. The developed machine learning classification and prediction models were built with a multilayer perceptron equipped with linear and non-linear transfer functions, regularization and dropout, and a binary sigmoid classification using machine learning technologies to create a strong and enhanced classification and prediction model. The developed LightGBM based classification and prediction models were trained and tested using the holdout method and 28 input attributes based on the clinical dataset from patients at the Cleveland Clinic. Based on the testing results, the developed machine learning models achieved diagnostic accuracy for heart disease of 83.67%, probability of misclassification error of 16.33%, sensitivity of 93.51%, specificity of 72.86%, precision of 79.12%, F-score of 0.8571, AUC of 0.8922, the K-S test of 66.62%, DOR of 38.65, and 95% confidence interval for the DOR of this test of [38.65, 110.28]. These results exceed those of currently published research. Therefore, the developed machine learning classification and prediction models can provide highly reliable and accurate diagnoses for coronary heart disease and reduce the number of erroneous diagnoses that potentially harm patients. Thus, the models can be used to aid healthcare professionals and patients throughout the world to advance both public health and global health, especially in developing countries and resource-limited areas where there are fewer cardiac specialists available.

REFERENCES

[1] K. Polaraju, D. Durga Prasad, “Prediction of Heart Disease using Multiple Linear Regression Model”, International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017. [2] Marjia Sultana, Afrin Haider, “Heart Disease Prediction using WEKA tool and 10-Fold cross-validation”, The Institute of Electrical and Electronics Engineers, March 2017.

[3] Dr.S.SeemaShedole, KumariDeepika, “Predictive analytics to prevent and https://www.researchgate.net/punlication/316530782, January 2016.

[4] Ashok kumarDwivedi, “Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation”, Springer, 17 September 2016.

[5] MeghaShahi, R. Kaur Gurm, “Heart Disease Prediction System using Data Mining Techniques”, Orient J. Computer Science Technology, vol.6 2017, pp.457-466.

[6] Mr. ChalaBeyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018. [7] R. Sharmila, S. Chellammal, “A conceptual method to enhance the prediction of heart diseases using the data techniques”, International Journal of Computer Science and Engineering, May 2018.

[8] Jayami Patel, Prof. TejalUpadhay, Dr. Samir Patel, “Heart disease Prediction using Machine Learning and Data mining Technique”, March 2017.

[9] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, “Efficient Heart Disease Prediction System”, 2016, pp.962-969.

[10] K.Gomathi, Dr.D.ShanmugaPriyaa, “Multi Disease Prediction using Data Mining Techniques”, International Journal of System and Software Engineering, December 2016, pp.12-14.

[11] Mr.P.Sai Chandrasekhar Reddy, Mr.PuneetPalagi, S.Jaya, “Heart Disease Prediction using ANN Algorithm in Data Mining”, International Journal of Computer Science and Mobile Computing, April 2017, pp.168- 172.

[12] Ashwini Shetty A, Chandra Naik, “Different Data Mining Approaches for Predicting Heart Disease”, International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277- 281.

[13] Jaymin Patel, Prof. TejalUpadhyay, Dr.Samir Patel, “Heart Disease Prediction using Machine Learning and Data Mining Technique”, International Journal of Computer Science and Communication, September 2015-March 2016, pp.129-137.

[14] Boshra Brahmi, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164- 168.

[15] NouraAjam, “Heart Disease Diagnoses using Artificial Neural Network”, The International Insitute of Science, Technology and Education, vol.5, No.4, 2015, pp.7-11.

[16] PrajaktaGhadge, VrushaliGirme, KajalKokane, PrajaktaDeshmukh, “Intelligent Heart Disease Prediction System using Big Data”, International Journal of Recent Research in Mathematics Computer Science and Information Technology, vol.2, October 2015 - March 2016, pp.73-77.

[17] S.Prabhavathi, D.M.Chitra, “Analysis and Prediction of Various Heart Diseases using DNFS Techniques”, International Journal of Innovations in Scientific and Engineering Research, vol.2, 1, January 2016, pp.1-7.