CHAPTER 1

INTRODUCTION

1.1 HEART DISEASES

The cardiovascular system, also known as the circulatory system, consists of the heart and the various blood vessels throughout the body. The heart is the largest and most important muscle in the circulatory system. Its job is to transport blood throughout the body. It also collaborates with other systems in the body to regulate blood pressure and heart rate. The coronary arteries that run along the surface of the heart supply the heart with nutrients and oxygen.

A range of illnesses that affect the heart and vascular system collectively are referred to as cardiovascular disease. It's a leading global killer and public health concern (Roth et al., 2020). Different forms of heart disease include Coronary Artery Disease (CAD), arrhythmias, congestive heart failure, and valvular heart disease.

CAD is brought on by cholesterol or fat deposits building up inside the coronary arteries and obstructing the flow of blood to the heart. Possible consequences of CAD include angina (chest pain), heart attacks, and sudden cardiac death. CAD risk factors include but are not limited to: high blood pressure, high cholesterol, smoking, obesity, diabetes, and lack of physical activity.

Congestive heart failure is a medical disorder in which the heart fails to pump enough blood to meet the body's needs. It may be driven on by a number of factors, including cardiac muscle injury caused by heart attacks, exorbitant blood pressure, or particular infections. Fatigue, shortness of breath, fluid retention, and a decreased capacityfor exercise are possible symptoms.

Arrhythmias are abnormal heartbeats that might hinder the heart from pumping normally. They can result in tachycardia, bradycardia, or atrial fibrillation, which are all abnormal heartbeats. While some arrhythmias are not dangerous, some can be serious and require medical attention.

The heart valves suffer damage by Valvular heart disease, which makes it challenging for them to function as intended. Aortic stenosis or mitral valve prolapse

might result from this disease. It can be congenital or acquired as a result of ageing, infections, or other factors. Fatigue, chest pain, breathlessness, and fluid retention are possible symptoms.

Adopting a healthy lifestyle which entails regular exercise, eating a nutritious diet low in fats and oils, abstaining from using tobacco products, controlling stress, andkeeping up a healthy weight are all essential elements of preventing heart disease. Intervention and early risk factor identification have been shown to minimize the incidence of cardiovascular disease.

1.2 HEART DISEASE - CAUSES AND RISK FACTORS

The heart is a muscle that circulates blood throughout thebody (Shahi et al., 2017). When the heart muscle contracts, blood is pushed from the heart chambers into the aorta (the body's main artery), where it is oxygenated and fed to the rest of the body. Extra oxygen is required for the heart to do its function. The heart muscle relies on the oxygenated blood carried by the coronary arteries.

Coronary Heart Disease (CHD) results from atherosclerosis, a process that causes accumulation of plaque on the artery walls, eventually narrowing or blocking the coronary arteries. Numerous factors lead to the risk of acquiring CHD. While some of these potential risks can be managed with exercise, modifications to diet, and/or medication, others cannot.

Factors causing risk for cardiovascular disease include but are not limited to: high blood pressure, high blood cholesterol levels, smoking, diabetes, obesity, lack of physical activity, poor nutrition, and stress. Factors like age, gender, genetics, and ethnicity cannot be changed.

Large cohort studies like the Framingham Heart Study (https: // www.nih.gov /sites /default /files /about-nih /impact/ framingham-heart-study.pdf) and the Third National Health and Nutrition Examination Survey (NHANES III) (https:// www.cdc.gov /nchs /nhanes / about_nhanes.htm) have discovered a robust correlation and predictive value for dyslipidemia, hypertension, tobacco use, and insulin resistance. Sixty percent to ninety percent of coronary heart disease events happened to people who had at least one risk factor.

Some main risk aspects of CHD are as follows:

- A history of cardiac illness in the family:
  Most people understand that cardiovascular disease often runs in families. Anyone with a history of cardiovascular illness in their family is at a higher risk for acquiring heart disease themselves.

- Smoking:
  For the heart, the brain, and the circulatory system, tobacco smoking is a leading risk factor. More than 40% of all deaths can be attributed to smoking-related heart and blood vessel diseases. After giving off smoking for a year or more, a person's risk of suffering a heart attack drops significantly.

- Cholesterol:
  Abnormal blood lipid levels are a known risk factor for cardiovascular disease. Cholesterol is a type of lipid found in the blood and in every living cell in the human body. An increased risk of atherosclerosis and cardiovascular disease is connected with elevated levels of Low Density Lipoprotein (LDL) cholesterol and triglycerides, the primary source of fat in the body.

- Hypertension:
  The medical condition of hypertension, sometimes known as high blood pressure, is commonly misunderstood. When our blood pressure is too high, the walls of our blood vessels can easily be damaged and stretched out of shape. It also increases the likelihood of suffering a heart attack, stroke, or heart failure.

- Obesity:
  To describe a person's health who is significantly over their ideal body weight, the term "obesity" is employed. Health issues like cardiovascular disease, high blood pressure, stroke, diabetes, and others are further exacerbated by obesity.

- Inadequate physical activity:
  The risk of CAD is higher in people who don't get enough exercise. Both diabetes and hypertension are independent risk factors for coronary artery disease, and both are made worse by a lack of physical activity.

1.3 DIAGNOSIS OF CHD

A physical examination is required if CHD is suspected (Bösner et al., 2010). Complex parameters and present symptoms will be evaluated. In order to make a correct diagnosis of CHD, more testing is required.

- **Blood tests**

  Electrolyte, blood cell, and hormone levels should all be determined by first taking a blood sample from the patient. Heart issues will be diagnosed by testing for specific enzymes and proteins.

- **Electrocardiogram (ECG)**

  Resting electrical activity of the heart can be measured with an ECG. Electrodes are placed on the chest, arms, and legs to record the electrical activity of the heart. The alterations are consistent with inadequate oxygen supply to the cardiac muscle.

- **Exercise ECG/ Exercise Tolerance Test (ETT)**

  This evaluation anticipates the heart's reaction to exercise and stress. The workout consists of 12 minutes of treadmill walking or bike riding at varying speeds and inclines. Shortness of breath, pain in the chest, jaw, or arm, and continuously monitored ECG and blood pressure readings.

- **Echocardiography**

  In patients with possible cardiac disease, echocardiography is an essential tool for assessing wall motion anomalies. Myocardial infarction can be detected earlier with the help of this instrument, which displays localized wall motion abnormalities. Patients with heart failure should always be evaluated for their ejection fraction as part of their care and monitoring.

  Cardiac catheterization, often known as an angiography, is a procedure used to examine blockages and narrowings in the heart's arteries. The effectiveness of cardiac function is also evaluated.

- **Nuclear Isotope Imaging**

  In order to create a nuclear isotope image, a radioactive chemical known as a tracer is injected into a patient's veins. The tracer is photographed virtually as it

passes the heart. The image is analyzed to determine the heart's pumping efficiency and whether or not any blood arteries are blocked. Imaging with nuclearisotopes includes procedures like Single Photon Emission Computed Tomography (SPECT) and Multigated Radionuclide Angiography.

1.4 TREATMENT OF CHD

Treatment for CHD (LaRosa et.al, 2001) is to discover the symptoms and thereby control and minimize the complexity of further problems. The precise condition and degree of severity of heart disease will determine the available treatments. Treatments could involve adaptations to one's way of living, drugs, surgical treatments, or minimally invasive techniques like angioplasty and stenting. Heart transplantation might be required in extreme circumstances. The treatments for heart disease are as follows:

- **Lifestyle changes**

  One of the most effective ways to treat CHD is to make significant changes to one's way of life. If the patients are diagnosed with CHD, this factor reduces the complexity of the disease

  .

- **Medicines**

  Several medicines are utilized to treat CHD. Both of these work to lower blood pressure and/or dilate narrowed arteries. There are negative reactions to some heart medications. The symptoms of heart disease can worsen if the patient abruptly stops taking their medication without first consulting with their doctor. The typesof medicines used in CHD treatment are as follows:

a) Antiplatelet medications

By preventing blood from clotting and thinning the blood, antiplatelet medicines can reduce one of several risk factors for heart attacks.

b) Statins

If the cholesterol level is high, cholesterol-lowering medicine named as statins are prescribed. Statins inhibit cholesterol synthesis and lower blood levels of LDL, hence lowering the risk of cardiovascular disease and heart stroke.

c) Beta-blockers

Angina and hypertension can be avoided with the help of beta-blockers such atenolol, bisoprolol, metoprolol, and nebivolol. They are effective because they counteract the effects of some hormones that raise blood pressure and lower heart rate.

d) Nitrates

Blood vessel dilation is achieved by nitrates. Nitrates have been called vasodilators by medical professionals. Glyceryl trinitrate and isosorbide mononitrate, for example, can be taken orally in the form of pills or sprays, or applied topically in the form of skin patches.

Nitrates work by widening the space between the walls of the blood vessels, allowing more blood to flow through. The heart discomfort and elevated blood pressure are both alleviated.

e) Angiotensin-converting enzyme (ACE) inhibitors

ACE inhibitors are used to treat hypertension all over the world. Inhibiting the hormone angiotensin II's function, which results in dilated blood arteries, is one of their primary functions. Blood circulation is improved with ACE inhibitors, too. While on ACE inhibitors, blood pressure levels must be maintained periodically, and kidney function must be checked with periodic blood tests. Dry cough and dizziness are typical ACE drug adverse effects.

f) Calcium channel blockers

Blood pressure is lowered by calcium channel blockers because they allow the smooth muscles that line the inside of the blood vessels to relax. This results in expanded arteries, which reduces blood pressure.

g) Diuretics

Diuretics, also known as "water pills," are drugs that cause the kidneys to excrete extra water and salt.

1.5 PREVENTION OF CHD

Methods to avoid or lessen the likelihood of CHD (Kotseva et al., 2017) are:

- Eat healthy food and also balance the diet
- Consume only a limited level of Alcohol
- Quit smoking
- Regularly exercise for 30 minutes every day
- Maintaining a healthy weight
- Manage cholesterol levels in the blood
- Effectively manage health problems including diabetes and hypertension
- Recognize the potential dangers

Research and medical advancements continue to improve our understanding of heart disease, leading to better prevention strategies, diagnostic tools, and treatment options. However, in order to counteract this pervasive and sometimes fatal disorder, it is still necessary to raise awareness about the importance of heart health and maintain a heart-healthy lifestyle.

Blood tests, ECGs, and Holter monitors are all viable options for diagnosing such conditions. Integrating and analyzing all this medical big data, which is collected and stored in various databases but provides no value on its own, can generate diagnostic information that can save lives and reduce costs using AI, ML, and Data Mining techniques. ML offers a fresh approach to addressing problems and responding to difficultqueries.

1.6 MACHINE LEARNING OVERVIEW

To put it simply, ML is the process of teaching a computer program, or "model," how to draw accurate inferences from a collection of data. An ML model is a representation of the mathematical relationship between the data elements used by an ML system to make predictions. Many large databases have been analyzed using ML methods in an effort to unearth underlying patterns.

The foundation of both AI and Data Science, ML is a rapidly developing subfield of Computer Science and Statistics. Success in ML in recent years can be attributed to a number of factors, including the development of novel learning algorithm theories and the

ever-increasing availability of vast quantities of data and inexpensive processing. More evidence-based decision-making is being fostered by the widespread adoption of ML- based approaches across many scientific, technological, and industrial domains, including healthcare, financial modeling, law enforcement, biomedicine, data governance, manufacturing, education, and marketing.

1.6.1 TYPES OF MACHINE LEARNING

ML intelligence can be done in one of four ways: Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, and Reinforcement Learning.

Supervised Learning:

Like being trained by a teacher or supervised by an employer, the name suggests a similar learning environment. Training (or labeled) data is separated from test data. The model will use the information it has learned from the training data to label the test data.

Unsupervised Learning:

Unsupervised learning refers to the process through which a model learns by itself,without the intervention of a human teacher or trainer. This technique clusters the data points based on their similarity. Unsupervised learning techniques are frequently employed to categorize data into groups and discover correlations between variables.

Semi-Supervised Learning:

Incorporating aspects of both supervised and unsupervised learning, semi-supervised learning (SSL). In order to improve model performance, it uses more unlabeled data and less labeled data. SSL works well for clustering and anomaly detection. Semi-supervised learning, however, does not perform well for all tasks. The strategy could fail if the labeled data's sample size isn't reflective of the distribution as a whole.

Reinforcement Learning:

In order to learn from its own actions and rewards, this sort of ML employs a trial-and- error methodology. The algorithms interact with an environment and learn from theconsequences of their actions. They try to maximize the reward or minimize the penalty, over time.

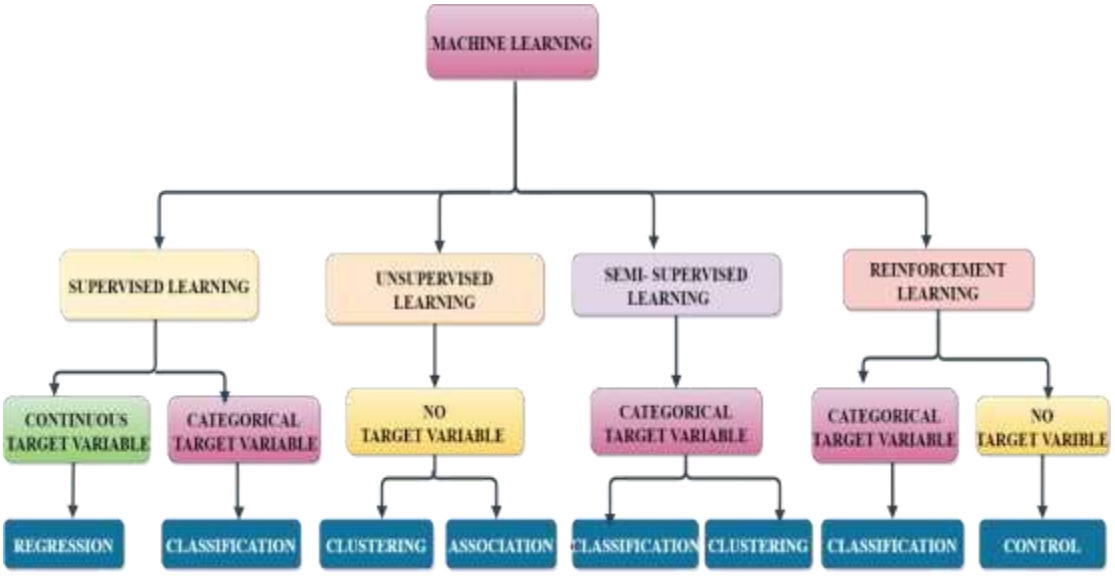

Figure 1.1 Classification of Machine Learning Techniques

Following this, a high-level summary of a few ML techniques will be provided.

1.6.1.1 CLASSIFICATION

Classification is a task in ML where the objective is to give labels or categories to input data based on those data's attributes. There are several types of classification algorithms, each with its own characteristics and suitability for different types of data and problem domains. Figure 1.1 shows the classification of ML techniques. Common ML classification algorithms are as follows:

Logistic Regression (LR):

LR is a common and easy method of categorization. It simulates the relationship between the input features and the likelihood of belonging to a given class. given that it takes the results of linear regression and converts them into a probability using a logistic function. It is appropriate for situations involving binary classification.

Decision Trees (DT):

Decision trees (DT) are decision support, tree like models which divide the input space based on the feature values. A class label is represented by a leaf node, and a feature test by an internal node. They can be used for both regression and classification.

They can be extended to ensemble techniques like Gradient Boosting and Random Forests (RF).

Support Vector Machine (SVM):

SVM is a robust classification technique that is used on high- dimensional datasets to find the optimum hyperplane. SVM seeks to reduce classification errors while maximizing the margin between classes. Through the use of several kernels, it can solve classification problems that are both linear and non-linear.

Naive Bayes (NB):

NB classifiers, which are based on Bayes' theorem, assume feature independence. Despite its imprecise premises, NB classifiers have proven useful in many practicalcontexts. They are frequently employed in text classification and spam filtering and are especially successful when dealing with high-dimensional data.

K-Nearest Neighbors (KNN):

Similarity to the labeled training examples is used by KNN, a non-parametric classification approach, to assign labels to fresh data points. Distances to the K nearest neighbors in the feature space are computed whenever a new data point is added. KNN is intuitive and useful for both binary and multi-class classification problems.

Ensemble Methods:

Ensemble methods combine multiple individual classifiers to make more accurate predictions. Popular ensemble methods include RFs, which combine DT's, and Gradient Boosting, which iteratively builds an ensemble of weak learners. Ensemble methods are often more robust and less prone to overfitting than individual models.

These examples of ML classification methods are just the tip of the iceberg. Data type, dataset size, computational resources, and task requirements all play a role in deciding which approach to use. It is common practice to try out different algorithms and pick the one that performs the task at hand the best.

1.6.1.2 CLUSTERING

Clustering is a method of ML used to organize data by identifying and grouping comparable pieces of information. Since it is an unsupervised learning technique, no labeled data or predetermined classes are necessary. Instead, it finds the hidden patterns inthe data and arranges them into manageable groups.

The primary objective of clustering is to enhance similarity within each cluster while minimizing similarity between them. In other words, there should be a higher degree of similarity between data points inside the same cluster than between clusters. Clustering algorithms employ various approaches to achieve this objective. Examples of popular clustering methods include the following:

K-Means Clustering:

K-means is often used as a clustering technique. The information is partitioned into k distinct groups, with k being a parameterized value chosen by the user. Theapproach iteratively reassigns data points to the nearest centroid (mean) depending on the new assignments. K-means seeks to reduce the average squared distance between data points and the chosen center.

Hierarchical Clustering:

By recursively merging or dividing clusters according to their similarity, hierarchical clustering creates a hierarchy of clusters. Agglomerative clustering and divisive clustering are the two most typical forms of hierarchical clustering. Agglomerative clustering treats each data point as its own cluster at the outset, and then joins the most similar clusters repeatedly until a termination condition is reached. Divisiveclustering, on the other hand, treats the entire dataset as a single cluster from the outset before subdividing it into smaller clusters at each subsequent iteration.

Gaussian Mixture Models (GMM):

GMM are probabilistic clustering techniques that rely on the assumption that the data was produced using a combination of Gaussian distributions. The method estimates the parameters (mean, covariance) of each cluster as a Gaussian component in an iterative expectation-maximization (EM) procedure. The likelihood that a data point belongs to each cluster is provided by GMM as a soft assignment of data points to clusters.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

This density-based cluster analysis tool is abbreviated DBSCAN. Data points with enough close neighbors and a close proximity to one another are grouped together. The system classifies data points according to their density connection as core, boundary, or noise points. DBSCAN is noise-resistant and capable of finding clusters of any shape.

Exploratory analysis methods like clustering are only as good as the data they're built on and the care with which their parameters are tweaked. Evaluation metrics such as silhouette score or cohesion-separation index can be used to assess the quality of clustering results.

1.6.1.3 REGRESSION

Predicting a continuous numerical value or continuous function from input data is the focus of the ML technique known as regression. By simulating the connection between the input features and the output variable, it facilitates continuous-range value estimate.

Regression seeks to discover a function that maps input variables to output variables, allowing for predictions of novel, unforeseen data. While the input features are known as independent variables or predictors, the predicted value in regression is frequently referred to as the dependent variable or the target variable.

Finding the optimum mathematical function that describes the relationship between the predictors and the outcome variable is the goal of regression algorithms. Common regression methods include polynomial regression, linear regression, Lasso regression, ridge regression, and support vector regression (SVR). These are only a few of the many ML regression techniques available.

Several factors, including data type, relationship complexity, need for interpretability, and computing constraints, should be taken into account while deciding on a regression technique. In order to discover the optimal model for a certain regression problem, it is usual practice to try out various methods and to fine-tune their parameters. Mean squared error, mean absolute error, and R-squared (coefficient of determination) are all useful ways to estimate a regression model's accuracy.

1.6.1.4 ASSOCIATION ANALYSIS

When applied to a dataset, the goal of association analysis, also known as association rule mining (Durairaj et al., 2013), is to identify potentially useful correlations or links between the individual records. It is often used in market basket analysis and recommendation systems to identify patterns and dependencies betweendifferent items.

The main objective of association analysis is to uncover frequent itemsets and generate association rules based on their occurrence. Items in a transaction are said to be part of an itemset, and an association rule defines the connections between itemsets. The rules often take the form of a "if-then" statement, where something on the left side of the statement suggests something on the right.

Association analysis provides valuable insights into the relationships and dependencies among items in a dataset. By discovering association rules and frequent itemsets, it enables businesses to make data-driven decisions, improve marketing strategies, and enhance user experiences. The well-known association analysis algorithm is the Apriori algorithm. Other algorithms are: FP-growth (Frequent Pattern-growth) and Eclat (Equivalence Class Transformation).

1.7 APPLICATION OF CLASSIFICATION METHODS IN THE HEALTHCARE SECTOR

Without a doubt, AI has increased the intelligence of computers. In order to mine enormous datasets and extract useful information, ML, a branch of AI, plays a crucial role. A machine's algorithm can save patients time and money by learning to recognize irregularities in the early stages of a disease through adequate training with a suitable train data set. Numerous research have examined the possibilities of ML and data mining in the healthcare business, and there is growing evidence that data created by Electronic Health Records (EHR) might provide insights for medical practitioners in terms of spotting irregularities for potential chronic diseases (Yin et.al, 2019).

Many different types of healthcare studies are making use of EHRs now (Uddinet al., 2019): analyses of healthcare utilization, measurements of hospital networks' efficiency, investigations of treatment patterns and costs, creation of risk prediction models for specific diseases, monitoring of long-term conditions, and comparisons of disease prevalence and treatment effectiveness. Disease risk prediction models were the focus of a lot of studies, particularly for chronic conditions like tuberculosis, diabetes, cancer, and cardiovascular disease. KNN, Neural Networks, Bayesian classification, DT, LR, Support Vector Machine (SVM) and Genetic Algorithm (GA) are only few of the ML methods utilized in cardiac diagnosis. These algorithms rely their models on patient-

specific training data that has been labeled. Patients in the test set are divided into subsets based on factors such as risk.

1.8 FEATURE SELECTION

Several studies (Jain, D., & Singh, V., 2018) The term "feature selection" is used to describe the process of selecting relevant information from a dataset. The effectiveness of a classifier is profoundly affected by the features chosen to train it. If you use inappropriate features, the classifier will become confused and produce inaccurate results. The difficulty can be solved by picking the right characteristics from the dataset, which has increased the classifier's precision and efficiency.

Feature extraction and feature selection are two separate processes. Theimportance of a limited number of data characteristics is closely tied to the featureselection process, while the feature extraction function creates novel features from raw features (Khan, S. et al. 2019). This technique eliminates superfluous information from raw datasets by picking a subset of characteristics from raw features. Selecting features or attributes is the same thing. Since the dataset is shrunk, the learning is more precise andthe outcome is clearer to the observer. There are two main types of search algorithms usedto find and delete irrelevant features: forward selection and reverse elimination.

Feature selection strategies are broadly categorized as filter methods, (Yildirim , 2015), wrapper methods, (Lee et al. 2018) and embedded methods, (Kamkar et al , 2015). Always feature selection algorithms use any of the three feature selection strategies.

1.8.1 Filter Method

The basic concept is to reduce a large feature set to a more manageable one by employing filters such as Bayesian conditional density filtering, kalman filtering, collaborative filtering, low variance filtering, highly correlated filters, Principal Component Analysis (PCA), ReliefF, Correlation-based Feature Selection, Fast Correlated Based Filter, and INTERACT. Variables are chosen using filtering procedures that ignore the model altogether. In cases where feature selection is determined solely by general considerations, like correlation with the target variable, these methods are straightforward and employ statistical procedures. Filtering techniques get rid of the factors that don't interest many people. The model categorization will incorporate additional variables, and

regression will be used to categorize or predict data based on predetermined characteristics. The complete procedure for filter method feature selection is depicted in Figure 1.2 below.

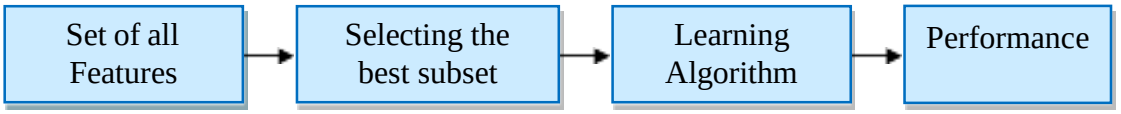

Figure 1.2 Process of Filter Method

Advantages of filtering methods

Some advantages of feature-selection-based filtering techniques are as follows:

- It is scalable
- It is self-regulating
- Feature-based filtering methods work faster than other methods
- It has better computational complexity than the wrapper methods based feature selection

Disadvantages of filtering methods

Some disadvantages of filtering methods based on feature selection are given below:

- Ignoring interaction with the classifier
- Ignore feature dependency

1.8.2 Wrapper Methods

The optimal features are easier to spot with wrapping approaches than the right features in the dataset. Wrapping methods use a subset of features and train a model using that subset of features and they are processed using a heuristic learning algorithm. Non-essential features in the subset are removed by training set backward elimination. The feature selection algorithm based on wrapper methods may have an effect on the classification algorithm. Using cross-validation helps prevent the over-fitting problem. Wrapper approaches are more expensive and time-consuming to choose acceptable features from the dataset, but they produce more accurate results than filtering methods. There are two major categories of wrapping techniques:

heuristic search algorithms and sequential selection algorithms. Method stages for a wrapper are shown in Figure 1.3.

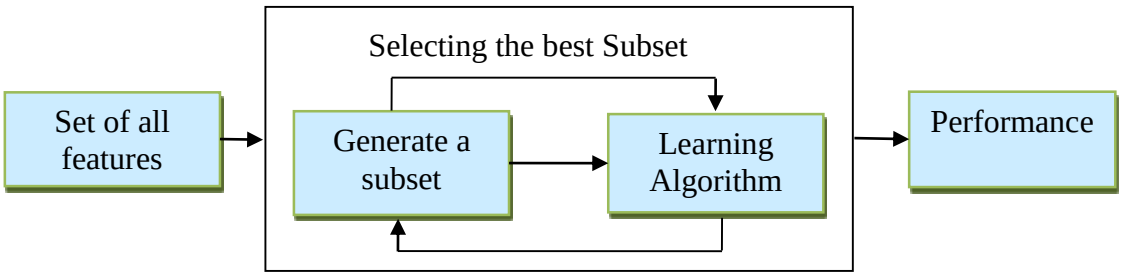

Figure 1.3 Process of wrapper methods

i) Heuristic Search Algorithms

Heuristic search algorithms are problem-solving techniques that aim to find efficient solutions by utilizing heuristic information or rules of thumb. These algorithms make informed decisions based on estimated values rather than exhaustive exploration of the search space. Greedy search, A* search, Best-first search, Hill climbing, Simulated Annealing, and the GA are all examples of popular heuristic search methods.

Examples of heuristic search algorithms that take their inspiration from the natural world include the Whale Optimization Algorithm (WOA) (Seyedali Mirjalili et al., 2016), the Ant Colony Optimization Algorithm (ACO) (Manonmani & Balakrishnan, 2020), the GA (Anbarasi et al., 2010), the Particle Swarm Optimization Algorithm (PSO) (Wong et al., 2012).

The theory of natural selection proposed by Darwin is the basis for GA. Exploration and optimization problems can both have approximate or exact solutions, and GA is a search strategy used to find them. The ACO is based on the fact that ants find extremely direct routes to their meals (Kabir et al., 2013). The random phenomenon of ants is not considered during the construction of subsets. PSO is a swarm intelligence- based optimization strategy that seeks a solution to the issue of optimization in the search field (Muthukaruppan & Er, 2012).

Heuristic search algorithms provide efficient strategies for solving complex problems by guiding the search based on estimated values. Choice of algorithm depends

on problem kind, domain knowledge, computing resources, and the trade-off between solution quality and speed.

ii) Sequential Selection Algorithms

Step-by-step, sequential subset selection algorithms typically evaluate the effect of each feature on the model's performance before deciding whether to keep or discard it. Here are two typical methods for selecting subsets in sequence:

Forward Selection: In forward selection, the algorithm begins with no previously chosen features and gradually fills them in through a series of iterations. After each iteration, the algorithm takes a look at how well the model is doing with the current set of features and picks the one that helps it out the most. Once a stopping requirement is met, the selected feature is included to the subset.

Backward Elimination: The algorithm in backward elimination begins with a full complement of features before gradually eliminating them one by one. The program iteratively removes features from the model and determines which feature removal has the smallest negative impact on model performance. A stop condition is reached, at which point the selected feature is deleted from the subset.

Both forward selection and backward elimination can be guided by various criteria to evaluate the impact of adding or removing features, such as statistical significance, information gain (IG), or a ML performance metric (e.g., accuracy, area under the curve, etc.). The problem at hand and the characteristics of the data dictate which criteria are applied.

Reducing the dimensionality of the feature space by sequential subset selection helps with model interpretability, overfitting, and maybe generalization performance. To be clear, getting the best subset of features is not a given when using sequential subset selection techniques. They rely on a step-wise search process that may lead to suboptimal solutions

Benefits of Wrapper Methods

The following are the benefits of wrapper methods for based feature selection:

- Communicates with the classifier
- Simple

- Dependence on templates
- More susceptible to local optimization

Drawbacks of Wrapper Methods

Drawbacks of wrapper methods based feature selection options are listed below:

- Computationally intensive
- Risk of over-fitting
- Classifier dependent selection

1.8.3 Embedded Methods

Embedded approaches incorporate feature selection into the algorithm itself and optimize it for use during model training iterations. This strategy combines wrapper and filter approaches and is hence sometimes referred to as the hybrid model. The computational overhead associated with reclassifying the various subsets done by wrappertechniques is diminished when embedded methods are used. Embedded approaches are depicted in the next Figure 1.4.

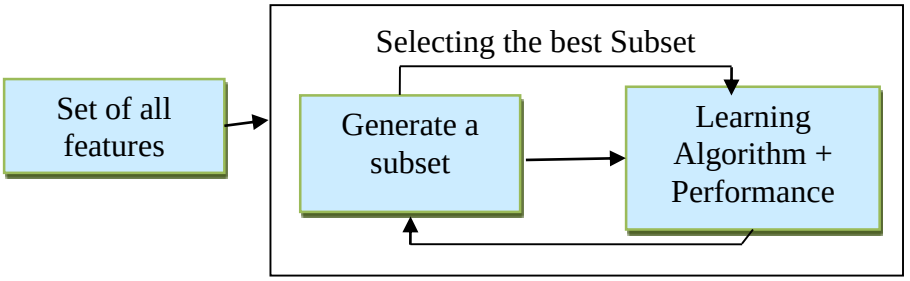

Figure 1.4 Process of Embedded methods

Advantages of Embedded Methods

The following are some of the advantages of using embedded methods for feature selection:

- Models feature dependence
- Improves computational complexity
- Interacts with classifier

Disadvantages of Embedded Methods

The embedded methods are dependent on classifier based feature selection

1.9 ENSEMBLE CLASSIFICATION

ML techniques known as "ensemble classifiers" pool the results of numerous models to get more reliable results. They take advantage of the fact that employingmultiple models together can often outperform using just one. As a result of their success in increasing prediction accuracy and deftly navigating complex challenges, ensemble classifiers have found widespread use across a variety of fields. Figure 1.5 depicts the processes involved in an ensemble categorization.

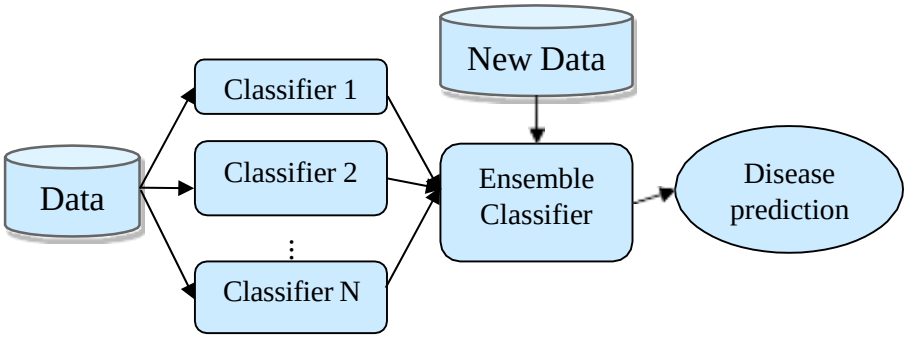

Figure 1.5 Process of Ensemble Classification

The homogeneous ensemble is one in which each classifier is of the same type. Heterogeneous describes everything else.

Bagging: Combining multiple models using bootstrap samples from the training data is known as Bagging (Bootstrap Aggregating), a broad ensemble method. Independently trained models then vote or average their results to arrive at a consensus. Classifiers like DT's (RFs) and SVM are only two examples of base classifiers that might benefit from Bagging (bagging SVMs).

Gradient Boosting: Together, gradient-boosted weak-learners (usually DT's) improve performance. Misclassified examples are prioritized during training, and each successive model is taught to correct the mistakes of the previous models. XGBoost and CatBoost (CatB) are two well-known gradient boosting techniques.

Majority-Vote Ensemble:

The predictions of many classifiers are combined in a majority-vote ensemble. Each individual classifier in an ensemble produces a prediction, and then the ensemble as a whole votes on which prediction to use.

Stacking:

Stacking is a type of ensemble learning that generates predictions by using a hierarchical combination of multiple independent classifiers or models. The stacking method differs from simple voting in that it relies on the training of a meta-model that learns to integrate the predictions of the separate models. Stacking can be a powerful technique for improving predictive performance, especially when the individual models in the ensemble have complementary strengths or expertise. However, it is more computationally expensive and requires additional data splitting and model training compared to simple majority voting.

One apparent argument against employing an ensemble classifier is that it is more labor-intensive to generate N-classifiers than a single classifier, and that the additional effort should only be warranted if the performance of the ensemble classifier is noticeably better than that of the single classifier. Intuitively, it appears reasonable to expect that N classifiers working together have the potential to produce improved prediction accuracy, albeit there is no guarantee that this will always be the case with the test data. The production of the classifiers and the combining of their predictions may have an effect in practice.

1.10 OVERVIEW OF THE MACHINE LEARNING PROCESS

An outline of the typical ML process is as follows:

Problem Definition: The problem should be clearly stated. this involves understanding the goal, the data at hand, and the desired results.

Data Collection: Collect information that is indicative of the issue at hand. This may entail gathering information from a variety of sources, including databases, APIs, and web scraping. Having sufficient and high-quality data is essential for model training.

Data Preprocessing: Clean up and prepare the acquired data for preprocessing so that ML algorithms can utilize it. In this stage, missing data are handled, duplicates areeliminated, features are normalized or scaled, and categorical variables are encoded.

Feature Engineering: Create new characteristics or modify existing ones with the goal of boosting the prediction power of the data. Techniques like feature scaling, dimensionality reduction (DR), or the development of interaction terms may be used for this.

Model Selection: Choose the appropriate ML algorithm or model that best suits the problem and data. Classification, regression, grouping, and recommendation are only a few examples of the many possible types of problems.

Model Training: Make a data training set and a validation set. Optimize the model's parameters or weights by applying an appropriate algorithm to the training data. The validation set is used to check the results of the model's training to prevent overfitting.

Model Testing and Evaluation: Put the model through its paces with some sandbox data it hasn't seen before. Depending on the nature of the problem, you should use assessment measures like accuracy, precision, recall, or MSE to assess the model's performance. This process is useful for checking the model's efficiency and finding any flaws or weak spots.

Model Deployment: Deploy the model to a production setting, where it can be used to make predictions or address the underlying problem, once its performance has been validated.

The ML process is iterative, and revisiting and refining previous steps based on the insights gained may be needed throughout the process. Also, ethics, fairness, and bias considerations should be taken into account at each stage to ensure responsible AI development.

1.11 RESEARCH MOTIVATION

Research helps put questions to rest, fills in knowledge gaps, and alters the way healthcare providers do their jobs. The following factors are motivating this study:

- Identify risk of illnesses at an early stage
- Decrease the number of people who would develop heart illness
- Provide a support system to medical practitioners in decision making

Depending on the economic system and GDP of the country, doctors are scarce and are not accessible everywhere (Raffaele et al., 2021). This provides inspiration for academicians and researchers to develop a clinical decision-making support system that

might aid in early prediction of the illness. Building such systems would expand healthcare access in low-income countries at a reasonable cost.

ML models when used for heart disease prediction have the potential to bridge the gap between research and clinical practice. By translating research findings into practical tools, healthcare professionals can utilize these models as decision support systems. ML-based risk prediction models can help clinicians identify high-risk patients, optimize treatment plans, and allocate healthcare resources effectively, thereby improving patient care and also outcomes.

1.12 PROBLEM STATEMENT

An effective technique for extending a patient's life is early diagnosis and prognosis. The dearth of resources and the scarcity of medical practitioners in developing nations, where a substantial fraction of the humans dwell, is a serious problem that creates a bottleneck in this respect and causes heart disease to be diagnosed only at an advanced stage. Moreover, there is a scarcity of medical experts in rural areas to analyze clinicaltests and provide expert opinions on the risk of heart disease. Most people approach medical practitioners after the onset of heart disease which results in low chances of recovery from the disease.

A system that predicts the risk of heart disease and advises seeking a medical expert’s opinion would address this problem. In this regard, numerous studies have been conducted to use machine learning techniques to predict cardiac disease. However, there is potential to explore improvements in the space of feature engineering, combining strengths of multiple models, etc., to make a new approach more robust. Hence, research on developing ensemble ML models focusing on improved performance is highly sought after. The problem statement is “to develop performance enhanced heart diseases risk prediction system to support clinical decision-making”.

1.13 AIM OF THE RESEARCH

To devise novel Machine Learning techniques towards developing an enhanced heart disease prediction system that supports clinical decision-making

1.14 OBJECTIVES OF THE RESEARCH

- To devise novel feature selection techniques to select relevant features for identifying the appropriate risk parameters for the prediction of heart disease with less computational cost
- To develop an ensemble classifier with high prediction accuracy by selecting an appropriate combination of diverse base learners
- To improve the ensemble model by optimizing the selection of classifiers in the ensemble

1.15 CONTRIBUTIONS OF THE RESEARCH

- In the first stage of the research effort, a wrapper feature selection method, ModifiedBoostARoota, is developed, which uses CatBoost as the base model and a novel feature elimination approach to efficiently identify the risk parameters of heart disease.
- Feature selection by Feature Importance Scores of Gradient Boosting Algorithms is proposed with a significant reduction in the search space of the sets of features
- A novel Super-Learner Ensemble Model (SLEM) with an appropriate combination of diverse classifiers, is proposed in the second stage of the research work, to accurately forecast cardiac illness
- An Optimized Super Learner Ensemble Model (OSLEM), using the Whale Optimization Algorithm and pairwise divergence measure to select an optimal set of base learners, is devised in the third stage of the research work, to improve the efficiency of the classification model.

1.16 ORGANIZATION OF THE THESIS

The remaining chapters of the thesis is organized as follows:

Chapter 2	Specifically, it surveys previous research on disease prediction algorithms, focusing on those that used feature selection and classification techniques
Chapter 3	Presents the research methodology and datasets used in this work, and describes the performance indicators.
Chapter 4	Describes feature selection by Feature Importance Scores of Gradient Boosting Algorithms and proposing the reduced search space of sets of features.
Chapter 5	Using the suggested ModifiedBoostARoota algorithm, a method is presented for identifying features that can help predict cardiovascular disease.
Chapter 6	Describes the study's secondary goal, a novel Super Learner Ensemble Model comprised of classifiers that have never been used before but produce high prediction accuracy.
Chapter 7	Provides an OSLEM for illness prediction by utilizing the Whale Optimization Algorithm and diversity measures as the study's third goal.
Chapter 8	Analyzes the suggested model's findings and evaluates them against similar studies.
Chapter 9	Presents conclusion and future work.

1.16 CHAPTER SUMMARY

Various forms of cardiac disease, potential causes, and current therapies are discussed in this chapter. Then, the ML process, types, classification, regression, and association analysis are concisely explained with elaboration on feature selection types. The problem statement, motivation for research, objectives, and research contributions are precisely mentioned in this chapter. The next chapter provides a systematic breakdown, by stages, of the relevant work on heart disorders utilizing various ML techniques.