Journal of Advances in Science and Technology
Vol. 21, Issue No. 1, March-2024, ISSN 2230-9659

DOI: https://doi.org/10.29070/jmr92f29

A Review on Early Detection and Classification of Heart Disease using Machine Learning Techniques

Shailendra Chaurasia [1] * , Dr. Megha Kamble [2]

1. Research Scholar, Department of Computer Science Engineering, LNCT University, Bhopal, Madhya Pradesh, India
chaurasia.shailendra@gmail.com ,
2. Professor, Department of Computer Science Engineering, LNCT University, Bhopal, Madhya Pradesh, India

Abstract: The human heart is the body's second most important organ after the brain, which is given more attention. It supplies and circulates blood throughout the body's organs. Predicting the incidence of cardiac disorders is an important task in the medical domain. Data analytics is helpful for making predictions using more information, and it aids medical centers in making predictions about different diseases. They keep a massive quantity of patient data every month. Predicting the incidence of future diseases may be aided by the recorded data. Predicting an occurrence of cardiovascular disease (CVD) is one use of data mining and ML. Despite being the most prevalent cause of mortality in the contemporary world, early detection of heart disease is notoriously challenging. Several data science challenges may be solved with the use of ML, which incorporates artificial intelligence. One popular use of machine learning is to make predictions using the data that already exists. In addition to summarizing the prior work, this article delves into the current algorithm. This study offers a comprehensive literature review of methods for predicting risk of heart disease, as well as an overview of the healthcare business in relation to heart disease, various diagnostic methodologies, kinds, hazards, and machine learning techniques.

Keywords: Heart Disease, Machine Learning, Decision Tree Algorithm, Naïve Bayes algorithm, Artificial Neural Network (ANN), Logistic Regression , RF and Ensmble learning

X

INTRODUCTION

The healthcare industry, alternatively referred to as the medical industry or health economy, encompasses various sectors of the economy that deliver palliative, preventative, curative, and rehabilitative care to individuals. It includes making and selling products and services that help people stay healthy or go back to normal after experiencing health problems. To meet the health needs of individuals and communities, the contemporary healthcare industry draws from three main pillars: products, finance, and services; it is further subdivided into numerous sectors and groups; and is supported by interprofessional teams comprised of trained paraprofessionals and professionals. Among the world's most important and rapidly expanding industries is healthcare. In the vast majority of industrialized nations, health care spending represents over 10 percent of GDP. In 2019, health care expenditures reached $3.8 trillion, or $11,580 per person. As a percentage of Gross Domestic Product, health expenditures represented 17.7 percent. From a few hundred dollars per capita in 1970s to an average of $4,000 per year in current buying power parties in power, pharmaceutical and health care expenditures in OECD countries have progressively increased.

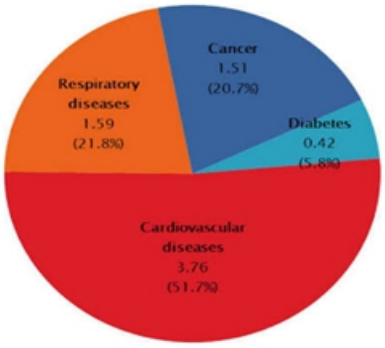

Figure 1: Global History of Death

Diagnostic procedures are utilized by doctors to clarify whether or not a cardiac disease is present. Usually, these tests are expressed using different statistical metrics. Figure 1 shows the anatomic structure of heart, clinically called cardiac, that has derived from Latin word. This organ is composed of four compartments of soft tissue muscle that are divided by pairs of blood arteries. The respective divisions are referred to as the atrium and ventricle. The ventricles of organ pump blood out of atrium, which collects blood in a group. The passage of oxygen gives the organism energy.

The healthcare sector is presently reliant on data collection and processing due to the substantial volume of data (big data) produced by various domains and sources (e.g., streaming machines, advanced healthcare systems, high throughput instruments, sensor networks, the internet of things, and mobile applications) [1]. Population growth and ageing contribute to the increase in cardiovascular disease-related fatalities in India, which is consistent with global trends. However, unlike in several other countries, India has not experienced the age-specific decline in mortality rates that is warranted. Cardiovascular disease is noted for being among the leading ten causes of death globally, including second in Japan in terms of mortality. Advancements in computer technology and a mounting fascination with artificial intelligence have contributed to a surge in research and development expenditures. Numerous preventive sensors and devices, including PCG and EGM, are developed due to the inability of the rudimentary ECG procedure to detect the majority of cardiac diseases [2]. A noninvasive medical decision support system, constructed by a group of researchers, utilises ML predictive models to address the intricacies associated with invasive methods of diagnosing cardiac disease. A decline in the mortality rate from cardiac disease has been observed as a result of these expert medical decision systems based on machine learning [3]. The repositories for cardiovascular diseases are located at the University of California, Irvine (UCI). Since its inception in 1988, the Cleveland dataset has been utilised by numerous researchers in the field of cardiac disease prediction. Additionally, it has been applied to machine learning. There are 303 occurrences of the record, each containing 14 attributes, of which 13 represent the independent variable [4]. This article investigates the predictive capabilities of a variety of machine learning algorithms for cardiovascular disease, including RF, SVM, Neural Network, KNN, DT, NB, BT, and LR. A principal contribution of this article:

- An overview of the conventional methods used to diagnose CAD.
- For the purpose of gaining a better understanding, the classification of the many forms of cardiovascular diseases (CVDs).

- An examination of the risk factors that are related with CVD, like high blood pressure, high cholesterol, and lifestyle options.
- Supervised, unsupervised, and ensemble machine learning approaches are all covered in detail when it comes to heart disease prediction.
- An extensive literature review was conducted, covering a wide range of current research and methodologies in the field of data mining and ML algorithms as they pertain to the prediction of cardiac disease.

DISEASE PREDICTION

"A goal of "Disease Prediction Using ML" is to correctly diagnose a patient based on their demographic data and symptom profile. If this initiative and other necessary procedures are successful in achieving early disease prediction, the disease can be treated, and this prediction system can be very valuable to health sector in general. The ultimate goal of this Disease prediction is to provide forecasts for the many and frequently occurring diseases which, if left untreated and occasionally ignored, can become fatal and create a great deal of trouble for the patient as well as their family members. Symptoms will be utilised to determine the most probable disease using this method. the health sector, which is an enormous industry with a substantial amount of unfinished business, and it is devoid of information and expertise. Thus, by implementing these strategies, algorithms, and methods, they have successfully concluded a project that has the potential to aid those in need.

Those who are now afflicted with disease must see a physician, which is both expensive and time-consuming. Because the sickness cannot be defined, it can be challenging for the user to locate hospitals and doctors when they are far away. Therefore, if the following technique can be performed utilizing automated software, which saves time and money, it would be advantageous for patients and make process more efficient. Data mining techniques are utilized by other HD prediction systems to assess a patient's risk. The suggested system uses data mining approaches to discover Chronic diseases in their earliest stages. The technique of programming computers to enhance their output depends on examples, or past data is known as ML. ML is a study of intelligent computer systems which may improve themselves via exposure to new information and practice. ML algorithms consist of a testing and training phase. Disease prediction based on patient symptoms and health records ML has long been a challenge.

A. Traditional Diagnostic Approaches

Common traditional approaches of CAD diagnosis include the following:

- **Medical History Assessment:** While gathering information about the patient's health history, the trained doctor will pay close attention to the patient's symptoms, risk factors (like obesity, high blood pressure, smoking, and family history of coronary artery disease), and past illnesses. By doing so, the patient's history may be better understood, and the risk of CAD can be more accurately assessed.
- **Physical Examination:** A physical examination entails the healthcare provider obtaining the patient's vital signs, including blood pressure and heart rate measurement, and evaluating heart sounds with a

stethoscope. A weak pulse or unusual cardiac sounds are physical symptoms that could lead to CAD suspicions.

· **Electrocardiogram (ECG/EKG):** The degree of electrical activity inside the heart may be evaluated with the use of electrocardiograms. As a result, it's useful for diagnosing arrhythmias, irregular heartbeats, myocardial ischemia, and prior heart attacks. To diagnose CAD, nevertheless, an ECG may not be enough.

· **Stress Tests**: The heart's response to an applied amount of physical strain is measured in a stress test An exercise stress test or a pharmaceutical stress test may do this as well. As the patient is under the effect of the anaesthesia, heart rate, blood pressure, breathing rate and electrocardiogram (ECG) are all being kept in check for the full duration of the stay. ECG abnormalities that stress induces can be an indication of lower blood flow to the heart muscles, which is a sign of CAD.

· **Coronary Angiography:** A contrast material that generates x-ray images of coronary arteries is injected into the coronary vessels in the context of coronary angiography. It helps to be able to see where the arteries are narrowed or blocked more distinctly. If it is an invasive procedure like an angioplasty or a stenting, it serves as a test for determining CAD.

· **Cardiac Imaging:** The equipment of and function of the heart can be assessed by a number of imaging modalities. The above mentioned techniques are the Cardiovascular MRI, PET, echocardiography (ultrasound based imaging), and SPECT (single-photon emission computed tomography). Heart function, blood flow, and the presence of anomalies or scar tissue suggestive of CAD may be shown by these imaging procedures.

· **Blood Tests:** To detect several biomarkers linked to CAD, blood tests may be conducted. As an example, there are inflammatory indicators such as troponin, which indicates cardiac muscle injury, C-reactive protein (CRP), and lipid profiles, which include cholesterol levels. Someone may be at a higher risk of CAD or continuing cardiac episodes if certain biomarkers are elevated.

HEART DISEASE LIKE CARDIOVASCULAR DISEASE (CVD)

Diseases of the heart are called heart diseases. Heart disease, or HD, and cardiovascular disease, or CVD, are synonyms. HD usually describes conditions like narrowed or blocked blood arteries, which may lead to stroke, angina (chest discomfort from decreased blood flow to the heart), or heart attacks. Forms of HD also include other cardiac disorders, such as those that harm the heart's valves, muscles, or rhythm. Depending on the type of HD you have, there are different symptoms. There may be differences in how men and women perceive HD. Women are more prone than men to have various chest discomforts as nausea, breathing difficulties and excessive fatigue. Men are more likely to experience chest pain altogether. When blood arteries in your arms or legs constrict, you may experience discomfort, numbness, weakness, or coldness [5]. Additional symptoms include discomfort in the back, neck, oesophagus, upper abdomen, and mouth. Valvular HDs, weakened heart muscles (dilated cardiomyopathy), irregular heart rhythms (arrhythmias), or weakened heart muscles (dilated cardiomyopathy) are all outcomes of HDs. A diagnosis of CVD is typically attained subsequent to the occurrence of a myocardial infarction, angina, or stroke. It's important to be aware of HD symptoms and to discuss them with your doctor. Sometimes CVD may be identified in its early stages, when treatment is easier. It is important to get medical help right away

if you have dyspnea, chest discomfort, or fainting. Multiple forms of HD may be avoided or treated with a healthy lifestyle.

Of all the human organs, the heart is very crucial. All throughout the body, it controls the circulation of blood. Any cardiac abnormality may cause pain in other body areas. HD refers to any condition that interferes with the proper functioning of heart. In the modern world, HD is one of the main causes of most fatalities. A high-fat diet, together with other risk factors like alcohol drinking, smoking, and an unhealthy lifestyle, can increase the risk of developing HD [6]. The World Health Organisation (WHO) estimates that HD kills over 10 million people worldwide each year. Heart-related disorders can only be prevented by leading a healthy lifestyle and being detected early. Figure 2 depicts the 2017 HD mortality rate structure.

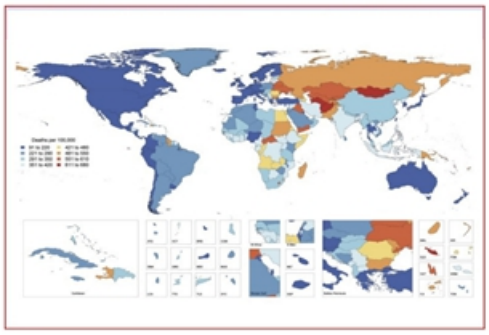

Figure 2: Global Map Standardized Death Rate of CVD in 2017

HD refers to a collection of disorders affecting heart, valves, muscles, arteries or internal electrical circuits responsible for contraction of a muscle. Heart disease is a major cause of mortality in several nations, including Canada, Australia, the United States, India, and the United Kingdom, according to the Centres for Disease Control and Prevention. CVDs account for roughly 31 percent (17.9 million) of all fatalities annually and are the major source of clinical (i.e., disability and death), health, and economic burden worldwide. In the USA, HD is responsible for one death out of every four. HD is prevalent among both women and men in majority of nations around the globe. Therefore, individuals should evaluate CVD risk factors. Although genetics play a part, certain lifestyle factors greatly influence HD. Radiation treatment for family history, smoking, gender, age, some chemotherapeutic medications and, malnutrition, cancer, high blood pressure, diabetes, obesity, high blood cholesterol levels, physical mobility, stress, & poor hygiene are among the established risk factors for CVD [7]. These are only a few of the many things that might raise a person's chances of acquiring CVD. Many different kinds of cardiac disease may manifest itself, including;

- **Coronary artery:** Coronary artery disease is caused by the occlusion of coronary arteries.
- Arrhythmia: An arrhythmia is a heart rhythm abnormality.
- **Heart infection:** Viruses and bacteria can cause heart infections.
- **Heart failure:** When the heart is unable to pump blood effectively, a condition called chronic heart failure develops.
- **Vascular disease:** When heart receives less blood, vascular disease develops.

· **Stroke:** Alterations to the blood supply lead to brain injury.

A. Types of Heart Disease

There are numerous classes of CVD. The many forms of cardiac disease based on clinical circumstances are depicted in Figure 3. These groups are broadly classified as heart arrhythmia, heart failure, myocardial infarction, cardiomyopathy, atrial fibrillation and angina pectoris based on their clinical evidence. Numerous characteristics of cardiac disease alter a structure or function of heart.

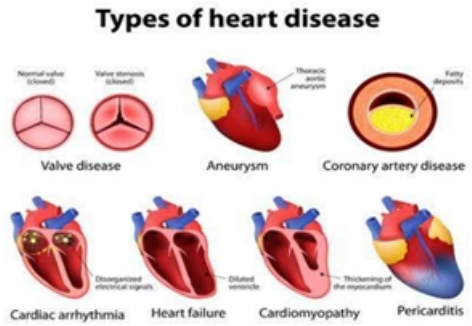

Figure 3: Types of Heart Disease

A classification of HDs focuses primarily on different clinical blood vessel dysfunctions, the presence of fatty substances in blood vessels that affect blood flow to heart, and numerous pathologies metrics like platelet counts of blood vessels, lipid profile dysfunctions, and white blood cell dysfunction that are a major contributor to HD.

1) Coronary Artery Disease

It is necessary to get medical attention for heart issues such coronary artery disease. A sufficient supply of oxygen and nutrients cannot be adequately supplied to the heart muscle by the coronary arteries, the primary blood vessels that provide blood to the heart. Deposition of cholesterol and inflammation inside the coronary arteries are the primary causes of coronary artery disease. If the heart isn't getting enough oxygen-rich blood, signs and symptoms of coronary artery disease can begin to manifest. Coronary artery disease is characterised by chest pain (angina), dyspnea, and reduced blood flow to the heart.

2) Acute Myocardial Infarction

A cardiac arrest is clinically known as an acute myocardial infarction. A cardiac arrest is a situation in which fatty substances in blood value impact the rate of blood flow, causing artery injury. If your arteries become clogged, your body may not get enough oxygenated blood, and your organs will stop working properly. Figure 1.6 depicts a type of cardiac arrest brought on by extreme pressure.

3) Rheumatic Heart Disease

Rheumatic HD is a disorder in which rheumatic fever produced by streptococcal infection damages the heart valves. Rheumatic fever causes inflammation throughout the body, which can affect the joints, heart, skin and brain. Acute rheumatic fever can affect anyone. However, it typically affects kids aged 5 to 15

years. The resulting rheumatic HD might persist for a lifetime. Every year, at least 8 out of every thousand newborns in UK are born with a cardiac abnormality.

4) Peripheral Arterial Disease

In peripheral arterial disease, the blood-supplying arteries to legs become constricted or entirely obstructed. Typically, the constriction of the artery occurs in the upper leg. The condition is characterised by a progressive accumulation of fatty material in the arterial walls (atherosclerosis). Atherosclerosis can also lead to the formation of a blood clot (or thrombus), which can totally block off the artery.

5) Chest Pain (Angina)

Angina is the medical term for chest pressure. It is frequently the case that individuals require emergency medical care. If they feel this kind of pain, patients need to be treated with ventilators very away. The lack of blood flow puts stress on the blood vessel walls that in turn puts stress on the blood vessels, which causes chest pain.

6) Stroke

A stroke is a severe neurological injury in which blood flow to a portion of brain is obstructed or ruptured (hemorrhage). The area of the brain supplied by a blocked or burst artery is no longer capable of receiving the oxygen carried by blood; as a result, brain cells suffer damage or pass away (become necrotic), reducing that area of the brain's functionality. To help you remember the main symptoms of a TIA or stroke, the acronym FAST stands for:.

· **Arms –** Due to arm numbness or weakness in one arm, the individual may be unable to raise and maintain both arms at the same time.

· **Face –** They may have a sagging face, be unable to smile, or have lost movement in one mouth or eye.

· **Time –** Dial 999 immediately if you observe any of the following symptoms

· **Speech –** Slurred or garbled speech, complete inability to speak, or difficulty comprehending what you say are all possible symptoms.

7) Congenital Heart Disease

All sorts of structural and functional defects of the heart can be attributed to aberrant or disordered heart development prior to birth, and this is what makes the term "congenital HD" so useful. Some conditions, like coarctation of the aorta, may not show symptoms for decades, while others, like a minor VSD (ventricular septal defect), may never cause any issues and are nevertheless compatible with regular physical activity as well as a normal lifespan. There are various kinds of congenital heart abnormalities, such as Source Reliable:

· **Atypical heart valves:** It is possible for a valve to not open properly or to spill blood.

· **Atresia:** One of the heart valves is missing.

· **Septal defects:** A hole has formed in heart's structure, most likely between upper chambers and lower

chambers.

8) Valvular Heart Disease

Symptoms of heart failure from valvular HD are similar to those from other potential causes of heart failure. Heart valve disease is a form of valvular HD. There are a variety of symptoms associated with cardiac valve illness, and they vary depending on which valve is malfunctioning.

· Fainting (syncope)

· Chest pain

· Irregular heartbeat

· Fatigue

· Swollen feet or ankles

· Shortness of breath

B. Risk Factors of Heart Disease

In some instances, a hereditary cause exists. Furthermore, some lifestyle choices and medical disorders can also contribute to an increased risk. These consist of the following:

1) High Blood Pressure

The arterial lining is damaged by high blood pressure, leaving the arteries more prone to plaque formation and narrowing, which increases the risk of HD and stroke. High blood pressure, defined as a reading of 130/80 mm Hg or above, affects almost half of the adult population in the United States, or around 116 million individual people. Only around half of the 86 million Americans who may benefit from taking medication to lower their high LDL cholesterol actually do so (55%).

2) High Cholesterol

In his work, describe that lipid profile fragments and characteristics are most significant components of healthy cell membranes as well as a healthy body. When blood cholesterol levels are high, HD is more likely to develop. The development of atherosclerosis is triggered by consuming high-cholesterol fatty foods and beverages. Lipid profiles can be split into two categories. Both low- and high-level lipid profiles exist.

3) Dietary Choices

Diets high in saturated fat, salt, trans fats and low in fruits, vegetables, and seafood have been related to an increased risk of CVD, while it is unclear whether or not these relationships prove causation. Low fruit and vegetable consumption is blamed for about 1.7 million deaths yearly, according to WHO. Consuming high-energy foods on a regular basis, especially processed foods with added sugar and fat encourages obesity and may raise your CVD risk.

4) Age

The greatest risk factor for developing CVD or HDs is age; the risk roughly triples for every decade past first.[8] Beginning in early adolescence, coronary fatty streaks can occur. According to estimates, 82 percent of those who die from coronary HD are aged 65 or older. After age 55, risk of having a stroke increases each decade.

5) Sleep

Insufficient quantity or quality of sleep has been shown to increase cardiovascular risk in both adolescents and adults. Infants are recommended to have at least 12 hours of sleep per day, while teenagers should get at least 9 hours, and adults should get at least 7 or 8 hours. One-third of American adults get less than 7 hours of sleep each night, and a small percentage of kids (2.2 percent in one research) got the recommended amount of sleep, with many of them getting poor-quality sleep.

6) Smoking

Cigarette smoking is widely believed to cause respiratory illnesses and cancer of the lungs. A big risk factor for HD is smoking, but you probably already knew that. In the United States, smoking is directly responsible for approximately 20 percent of deaths caused by HD. The likelihood of developing HD rises with number of cigarettes smoked. How long they've smoked is also relevant. If you smoke a pack of cigarettes every day, your risk of suffering a heart attack is double that of nonsmokers. The risks of CVD, blood clots, stroke and peripheral vascular disease, are considerably increased in women who take birth control tablets and smoke.

7) Diabetes Mellitus

[9] describe the molecular processes, recommended treatments, and epidemiological conditions associated with diabetes. Coordination of blood sugar flow results in the development of diabetes. Patients with diabetes who do not adequately manage their blood pressure, blood sugar, and cholesterol levels may develop CVD.

MACHINE LEARNING TECHNIQUES FOR HD

An important use of ML, a branch of AI, is the ability of computers to learn complicated tasks independently of human programming. The analogy that best describes ML is farming or gardening. Just as the gardener tends to the seeds, the nutrients nourish the data, and the plants grow from the programmes. ML is the study of making computers behave autonomously without being explicitly programmed. In past decade, Today, ML is so prevalent that you unknowingly employ it thousands of times every day. Numerous researchers concur that this is the greatest strategy to advance human-level artificial intelligence. ML is coming into its own, as it is increasingly acknowledged that it can play a crucial role in a vast array of crucial applications, including natural language processing, picture recognition, data mining, and expert systems.

Computer algorithms that enable computers to learn and enhance automatically based on experience are focus of ML research. Generally, it is considered a subfield of artificial intelligence. ML algorithms

enable systems to make judgments automatically and without assistance. Such decisions are made by discovering important patterns underlying complex data. There are a few main groups of ML algorithms, including unsupervised learning, supervised learning, and reinforcement learning, which are all distinguished by their respective learning approaches, the types of data they take in and produce, and the problems they are designed to address. There are some hybrid strategies and other prevalent methodologies which provide a natural generalization of ML problem types.

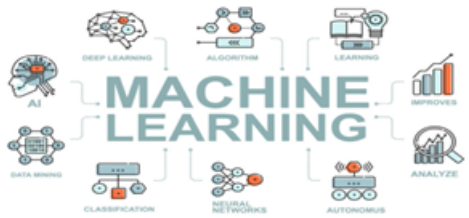

Figure 4: Architecture of Machine Learning

A. Different Types of Machine Learning

supervised learning, Unsupervised learning, Supervised learning, and Reinforcement learning [10], as shown in Fig. 5. In following, they briefly discuss every kind of learning method.

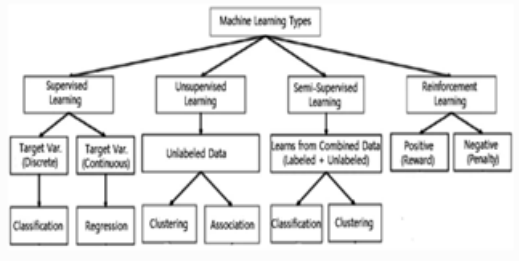

Figure 5: Types of Machine Learning

1) Supervised Learning

The term "supervised learning" is utilized to describe a process of training a model on a sample of data from a data source before applying that model to a test dataset that is also obtained from the sample [11]. There are 2 types of supervised learning: regression and classification. The model is trained on a labeled data set. It has data inputs and outputs. The classification and separation of data items into test and training sets. The purpose of training dataset is to teach our model, whereas testing dataset is to evaluate it on new data. Classification and regression are both examples of supervised learning. The division of datasets is an often-mentioned concept in supervised learning. Clinical researchers are more likely to use ML approaches (algorithmic models) that do not assume anything about the data and instead cross-verify outcomes [12].

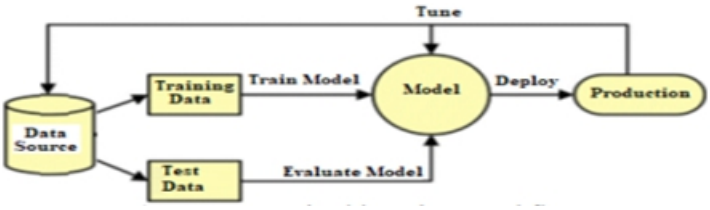

Figure 6: Supervised Learning Workflow

· **Classification model:** The process of finding a model that helps to separate data into discrete categorical groupings is known as classification. Data is first categorised using several labels according to input attributes, and then these labels are used to predict data. Categorical denotes that the output variable is a categorisation, such as spam or not spam, red or black, diabetes or not diabetic, etc.

· **Regression model:** A continuous (real-valued) output prediction is attempted. Predicting a house's value from a set of input variables is an example of a regression problem.

2) Unsupervised Learning

These training data items are neither categorised nor labeled. The objective is to uncover patterns within the data. It can readily anticipate hidden patterns for any novel dataset or training data, however during data exploration, it draws inferences from databases to explain hidden patterns[13]. Association analysis, Data clustering, and dimensionality reduction are the three primary uses of unsupervised learning, all of which are essential for evaluating and classifying data based on similarities, traits, and correlations. Therefore, in this section, they cover the basics of unsupervised learning techniques such as cluster analysis, association rules, and PCA.

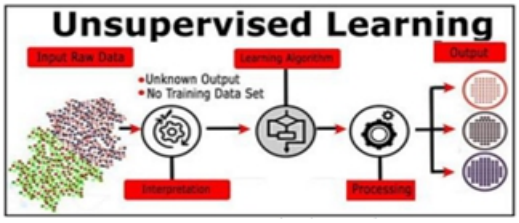

Figure 7: Unsupervised Learning Workflow

· **Association rule mining:** Product sales data, or any other relevant transactional data, can be mined for insights on whether or not there is a hidden relationship between buyers and products, or between products themselves, that could be used to boost revenue. Association Rule Mining is built around the extraction of these interconnections. They can extract relationships using SETM, Apriori, AIS, and FP growth algorithms.

· **Clustering:** Data without class or label information is suitable for clustering. They'd like to cluster the information so that comparable observations are together and for the spacing between clusters to be as large as possible. Simultaneously, the intra-cluster distance should be kept to a minimum. They can cluster the data of voters to identify their opinion of the government, or they can cluster products based on their characteristics and consumption. Market to specific income brackets or divide the population into subgroups and target them individually.

3) Reinforcement Learning

As a result of not using a labelled dataset and having uncorrelated outputs, the model learns from its own experiences. Here, a model learns from its surroundings to improve its performance and fix its shortcomings; testing and assessment determine the final outcome. Classification algorithms frequently

utilization supervised learning methods for determining the probability of HD occurrence.

Figure 8: Reinforcement Learning

4) Semi-Supervised Learning

The main difference between supervised and semi-supervised learning is that the former uses both labelled and unlabeled data. Since unlabeled data is easier and cheaper to obtain, it is combined with a smaller amount of labeled data to improve accuracy. Methods like classification, regression, and prediction may benefit from this kind of learning. A fully labelled training approach is not always feasible due to labelling costs; in such cases, semi-supervised learning might be helpful. Finding a person's face in a webcam image is one of the first applications of this.

Figure 9: Semi-Supervised Learning

B. Machine Learning Algorithm

The term "classifier" refers to a method in ML that can automatically arrange data into different "classes." An example of this is an email classifier, which sorts incoming messages according to predefined categories.

1) Decision Tree Algorithm

It has been found that the performance of linear regression and LR models deteriorates when the link among features and outcome is nonlinear or when features interact with one another. DTs are an easy way to handle these scenarios. For DTs to be effective, the data must be split numerous times, each time taking into account the potential gain in knowledge after the split (using some metric, such as information gain). Different subsets of the dataset are produced by the splitting technique. The very last subsets are known as leaf nodes. On these leaf nodes, the prediction is based on the mean result of the relevant training data. These studies have a commonality: they all use DT models to examine the relationship between variables and to classify participants into similar groups according to their observable characteristics. Decision algorithms with a similar structure work better with the DT since it accounts for the substantial interaction

between variables [14].

2) Naïve Bayes algorithm

In addition to the LR model, the Naive Bayes classifier was employed in this analysis. It's also a classification model for supervised learning, where the data is divided into categories based on the probabilities of certain factors. After computing the probability of each class, transaction is assigned to the class with the highest probability. Nave Bayes is a popular method for predicting classes for various sorts of datasets, including educational data mining [15] and medical data mining. This approach can also be used to classify various types of datasets, such as sentiment analysis and virus detection.

3) Artificial Neural Network (ANN) Algorithm

In an effort to replicate human brains, ANNs have consistently identified extensive applications for solving a vast array of nonlinear issues and have attracted 21 increasing research considerations. The most well-known characteristic of ANNs is its nonparametric, nonlinear, data-driven, and adaptive nature. ANNs do not necessitate knowledge of information's underlying statistical distributions. These factors include, but are not limited to, the selection of the appropriate network design, training technique, number of nodes in every layer, number of hidden layers, and activation functions. The selection of a network training algorithm could conceivably be regarded as the most crucial aspect of ANN modeling.

4) Logistic Regression Algorithm

The Supervised Learning approach includes LR, one of the most used ML algorithms. This approach may be used to predict the categorical dependent variable from a set of specified independent components. Predicting the value of a dependent variable that can be categorised into two or more categories is the objective of LR. Consequently, the outcome has to be a numerical value that falls into one of many categories. Instead of giving a precise Yes or No number, it gives probabilistic values ranging from 0 to 1, which may be anything from true to false, zero to one, etc. In all other respects, LR is very similar to its more well-known cousin, Linear Regression. Specifically, linear regression is utilized to address issues of regression, while LR is more suited to issues of classification.

5) Support Vector Machine

A SVM is a straightforward method that professionals can employ for classification or regression tasks. A hyperplane, which may be seen as a line dividing 2 groups of data, is what they use to do their work within a distribution. The method will choose the best possible hyperplane to use in the separation out of a set that contains many. The optimal hyperplane in the SVM model is the one that provides the largest margin of separation between the various classes [16].

6) Random Forest Algorithm

The supervised learning technique known as RF is used for classification and regression in the field of ML. It is a classifier that averages the outcomes of several DTs applied to diverse subsets of a dataset in order to enhance the projected accuracy of dataset. For enhancing the model's forecast accuracy without over-fitting, it takes the average of the results fitted to different DTs to different sub-samples data. The input

sample and the sub-sample both have the same size, but samples are produced through random replacement.

7) K-Nearest Neighbor

The k-Nearest-Neighbor classification algorithm, or kNN, is an ML technique that identifies the set of k objects in a training example that is geographically nearest to test object and then assigns a label based on the frequency with which a particular class appears nearby. This algorithm requires 3 significant components: [17]. "Euclidian Distance" is a commonly utilized proximity metric for kNN classification.

8) Ensemble Learning

This approach increases accuracy by combining numerous classifiers into a single model. There are three different kinds of Ensemble learning strategies. The first type is Bagging, which collects comparable classifiers through a voting process. The second type, boosting, is similar to bagging, except the new model is impacted by outcomes of prior model. Stacking is 3rd type which entails combining multiple ML classifiers into one model.

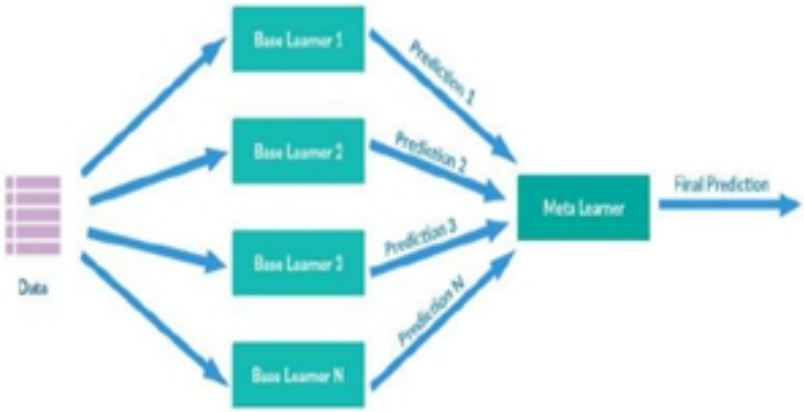

Figure 10: Example of Ensemble Learning

9) XG-Boost

The Gradient Boosted DTs algorithm has been implemented in XG-boost. The Gradient Boosted DTs algorithm has been implemented in XG-boost. This algorithm uses a sequential method to generate DTs. The weights have a significant impact on XG boost. All of the independent variables are given weights, and this information is fed into DT, which generates a forecast. When variables are included into second DT, they are given more weight if the tree made an inaccurate prediction. After that, a strong and precise model is constructed by merging all of these classifiers and predictors. It can rank data, do regression, classify it, and make predictions based on user input.

LITERATURE SURVEY

Medical institutions have conducted a great deal of research on illness prediction systems that use various data mining and machine learning methods.

In, Islam, Rafa and Kibria, (2020) PCA has been utilized to help with tagging. Final clustering was accomplished using a Combining k-means and the Hybrid Genetic Algorithm (HGA), but this was not the case for the preceding stages. It used a hybrid genetic strategy to cluster the data around this problem (HGA). The predictive accuracy of our suggested technique for early cardiac disease is 94.06%[18].

In, Sharma, Pal and Jaiswal, (2022) develop a reliable method and put it to use to generate a prediction-based strategy. Convolutional Neural Networks (CNNs) are explained, along with the deep learning technique that powers them and is particularly effective for processing images and classifying or interpreting them using CNN; the developed model achieves a 96.1% success rate in its predictions. When classifying medical photos connected to heart disease[19].

The analysis of the, Deepika and Balaji, (2022) proposed system is 94.28% more accurate than the best available alternatives. The proposed method achieved a 96% F1 score, 96% recall, and 96% precision in the analysis. Therefore, the proposed approach proved superior compared to other methods for predicting cardiovascular disease[20].

In, Reddy and Kanimozhi, (2022) to use a new, intelligent model to forecast cardiac disease and dynamic KNN (SVM). Analysis and Result: SVM accuracy is 67.21%, whereas Dynamic KNN accuracy is 84.44%. Analytically, Dynamic KNN and SVM differ significantly from one another. Conclusion: For novel heart disease prediction, dynamic KNN looks to perform noticeably better than SVM[21].

In, Kumari and Mehta, (2021) have been seven different ML algorithms employed to try and make forecast about heart disease, and ensemble methods. Compared to other algorithms, Linear Discriminate Analysis performs well; Compared to Logistic Regression, its accuracy is almost 80%, with a mean value of about 0.847 and a MAE of 0.185. The highest false identification rate is 0.076, while the lowest false positive rate overall is 0.33[22].

In, Yadav, Saini and Mittal, (2021) established by the assessment of the effectiveness of It all started with a few machine learning algorithms—KNN, SVM, RF, and NB—using the dataset to forecast heart diseases. An average accuracy of 87.78% is produced using the Naive Bayes model, which is the highest. Overall, the model works as expected. After some time, the innocent Base reaches the maximum average accuracy of 96%[23].

The purpose of this study, Miranda et al., (2021) aimed to create a data mining-based tool for early heart disease prediction. The experimental results showed that the CART application for early heart disease prediction had a high level of accuracy (86.33 percent), precision (88.00 percent), F Measure (84.62 percent), and recall (84.62 percent) [24].

This study, Junaid and Kumar, (2020) constructs a combined model based on data science principles and algorithms. Nave Bayes, ANN, SVM, and Hybrid Nave Bayes, ANN, and SVM have all had their precision, recall, and resemblance to the accurate classifier. The hybrid model is superior to its individual parts, with a specificity of 82.11% and a sensitivity of 91.47%[25].

In, Mehta and Varnagar, (2019) proposed a method for identifying stroke risk factors using data mining classification algorithms for the stroke dataset. An individual's risk of IHD. DT, KNN, LR, SVM, and NB

were applied to the Ischemic Stroke Dataset's five data mining techniques. The SVM method achieved the maximum precision (97.91%) in the results[26].

This study's primary objective is to, Ayeshmi M. and Peiris, (2022) unveil a mobile app that analyses the heart disease dataset at UCI using eight characteristics to generate a ML model for prediction. DT, KNN, and RF are ML algorithms used in this study. And the best accuracy level attained by machine learning models is 85%[27].

Primary inspiration for the work, Rajendran and Karthi, (2022) aims to suggest a brand-new ML pipeline for precise heart disease prediction. Experiment results show that IMV + OR pre-processing is far superior in pre-processing strategies for model evaluation. The suggested pipeline's ensemble model (LR + NB) outperformed state-of-the-art results in AUC (96.8%), Accuracy (92.7%), Specificity (91.5%), Precision (92.5%), and F1 Score (92.7) [28].

In, Riyaz, Butt and Zaman, (2022) This research made use of ensemble classifiers including AdaBoost, weighted average, bagging, gradient boosting, and majority voting. When tested on the provided dataset, the bagging ensemble classifier had the maximum prediction accuracy (85%) and was therefore named the best classifier[29].

In, Sharathchandra and Ram, (2022) proposed user-interactive dual disease prediction system. In this investigation, LR and SVM models predicted disease. Specifically, the proposed model can correctly predict cardiovascular disease in 85% of cases and diabetes in 78% of cases[30].

Numerous research works have investigated various approaches to the prediction of cardiac illness, like an use of CNNs for image classification, Principal Component Analysis (PCA) in conjunction with hybrid genetic algorithms, and ML algorithms like Dynamic KNN and SVM. Accuracy results using these methods range from 67.21% to 96.1%, with certain approaches showing better accuracy, recall, and specificity than others. Achieving accuracy of up to 96.8%, ensemble approaches like bagging and hybrid models have also been studied. Furthermore, attempts have been made to create user-interactive systems for dual illness prediction, with encouraging outcomes in the diagnosis of diabetes and cardiovascular disease. All things considered, these investigations show the many methods and developments in using machine learning to predict and diagnose cardiac disease.

Table 1: Related work for head disease prediction using machine learning techniques

Reference	Methodology	Results	Limitations	Future Work
[18]	PCA, Hybrid Genetic Algorithm	Predictive accuracy: 94.06%	Limited to early cardiac disease prediction	Enhance hybrid genetic strategy for better accuracy
[19]	Convolutional Neural Networks	Prediction success rate: 96.1%	Focuses on image classification	Explore CNNs for broader medical diagnostics

[20]	Data mining-based	F1-score: 96%, Recall: 96%, Precision: 96%	Limited comparison with alternatives	Investigate scalability and generalization of approach
[21]	Dynamic KNN, SVM	Dynamic KNN accuracy: 84.44%	Limited to heart disease prediction	Improve SVM performance or explore new methods
[22]	Machine learning algorithms	Accuracy: ~80%, Falses: 0.076 - 0.33	Limited ensemble methods exploration	Investigate ensemble methods for better performance
[23]	KNN, SVM, RF, NB	Naive Bayes accuracy: 96%	Limited to cardiac disease prediction	Explore ensemble methods for improved accuracy
[24]	CART method	Accuracy: 86.33%, Precision: 88.00%	Limited to early heart disease prediction	Investigate CART improvements for higher accuracy
[25]	Hybrid models (NB, ANN, SVM)	Specificity: 82.11%, Sensitivity: 91.47%	Limited individual model comparison	Optimize hybrid model parameters for better results
[26]	Data mining classification	SVM precision: 97.91%	Limited to stroke risk factors	Investigate SVM enhancements for broader use
[27]	DT, KNN, RF	Best accuracy: 85%	Limited to smartphone application	Refine ML algorithms for better prediction accuracy
[28]	IMV + OR preprocessing, Ensemble	AUC: 96.8%, Accuracy: 92.7%	Limited comparison with other pipelines	Explore advanced ensemble models and preprocessing

[29]	Ensemble classifiers	Bagging accuracy: 85.0%	Limited to ensemble methods	Investigate ensemble diversity for improved accuracy
[30]	LR, SVM	Cardiovascular disease prediction: 85%	Limited to dual disease prediction	Enhance model for broader disease prediction

CONCLUSION

A variety of data mining and ML methods have been compiled to predict a development of cardiovascular disease. Find out how well each algorithm predicts and implement the suggested system in the required region. To make algorithms work more accurately, use feature selection approaches that are more relevant. A patient diagnosed with a certain kind of heart disease has access to a variety of therapy options. On the right dataset, data mining may provide a wealth of information.

Finally, the literature review confirmed our suspicions that current predictive models for patients with heart disease are only partially effective; thus, they need more sophisticated models that combine different types of data to improve our ability to foretell when heart disease will begin to spread. An more sophisticated system will emerge from the database as its data input grows.

To make this prediction system more accurate and scalable, there are a lot of potential upgrades that might be investigated. The following study and work has to be done in the future because of time constraints. I am interested in experimenting with various discretization methods, voting strategies, decision tree kinds (namely, information gain and gain ratio), and multiple classifiers. Open to investigating various rules, including clustering methods, logistic regression, and the association rule.

References

1. A. Ed-Daoudy and K. Maalmi, “Real-time machine learning for early detection of heart disease using big data approach,” in 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems, WITS 2019, 2019. doi: 10.1109/WITS.2019.8723839.
2. M. A. Alim, S. Habib, Y. Farooq, and A. Rafay, “Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model,” in 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies: Idea to Innovation for Building the Knowledge Economy, iCoMET 2020, 2020. doi: 10.1109/iCoMET48670.2020.9074135.
3. A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,” Mob. Inf.

Syst., 2018, doi: 10.1155/2018/3860146.

4. P. Singh and I. S. Virk, "Heart Disease Prediction Using Machine Learning Techniques," 2023 Int. Conf. Artif. Intell. Smart Commun. AISC 2023, no. July, pp. 999–1005, 2023, doi: 10.1109/AISC56616.2023.10085584.
5. R. R. Sun, M. Liu, L. Lu, Y. Zheng, and P. Zhang, "Congenital Heart Disease: Causes, Diagnosis, Symptoms, and Treatments," Cell Biochem. Biophys., 2015, doi: 10.1007/s12013-015-0551-6.
6. T. Nagamani, S. Logeswari, and B. Gomathy, "Heart disease prediction using data mining with mapreduce algorithm," Int. J. Innov. Technol. Explor. Eng., 2019.
7. J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," in Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2016, 2016. doi: 10.1109/ICCPCT.2016.7530265.
8. J. A. Finegold, P. Asaria, and D. P. Francis, "Mortality from ischaemic heart disease by country, region, and age: Statistics from World Health Organisation and United Nations," Int. J. Cardiol., 2013, doi: 10.1016/j.ijcard.2012.10.046.
9. Suba, "Introduction To Heart Disease," pp. 1–15, 2015.
10. M. Mohammed, M. B. Khan, and E. B. M. Bashie, Machine learning: Algorithms and applications. 2016. doi: 10.1201/9781315371658.
11. R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," Int. J. Adv. Res. Artif. Intell., 2013, doi: 10.14569/ijarai.2013.020206.
12. S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," Stat. Surv., 2010, doi: 10.1214/09-SS054.
13. R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," Brief. Bioinform., 2017, doi: 10.1093/bib/bbx044.
14. A. Carmona-Bayonas et al., "Predicting serious complications in patients with cancer and pulmonary embolism using decision tree modelling: The EPIPHANY Index," Br. J. Cancer, 2017, doi: 10.1038/bjc.2017.48.
15. F. Razaque et al., "Using naïve bayes algorithm to students' bachelor academic performances analysis," in 4th IEEE International Conference on Engineering Technologies and Applied Sciences, ICETAS 2017, 2018. doi: 10.1109/ICETAS.2017.8277884.
16. F. W. Roush, "Applied linear regression," Math. Soc. Sci., 1982, doi: 10.1016/0165-4896(82)90010-5.
17. A. Cruz and P. Cortez, "Data Mining via Redes neuronais Artificiais e Máquinas de Vectores de Suporte," Rev. Estud. Politécnicos, 2009.
18. M. T. Islam, S. R. Rafa, and M. G. Kibria, "Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means," in ICCIT 2020 - 23rd International Conference on Computer and Information Technology, Proceedings, 2020. doi: 10.1109/ICCIT51783.2020.9392655.

19. A. Sharma, T. Pal, and V. Jaiswal, "Chapter 12 - Heart disease prediction using convolutional neural network," in Cardiovascular and Coronary Artery Imaging, A. S. El-Baz and J. S. Suri, Eds., Academic Press, 2022, pp. 245–272. doi: https://doi.org/10.1016/B978-0-12-822706-0.00012-3.

20. D. Deepika and N. Balaji, "Effective heart disease prediction using novel MLP-EBMDA approach," Biomed. Signal Process. Control, vol. 72, p. 103318, 2022, doi: https://doi.org/10.1016/j.bspc.2021.103318.

21. K. S. K. Reddy and K. V Kanimozhi, "Novel Intelligent Model for Heart Disease Prediction using Dynamic KNN (DKNN) with improved accuracy over SVM," in 2022 International Conference on Business Analytics for Technology and Security (ICBATS), 2022, pp. 1–5. doi: 10.1109/ICBATS54253.2022.9758996.

22. A. Kumari and A. K. Mehta, "A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms," in 2021 Asian Conference on Innovation in Technology, ASIANCON 2021, 2021. doi: 10.1109/ASIANCON51346.2021.9544544.

23. D. P. Yadav, P. Saini, and P. Mittal, "Feature Optimization Based Heart Disease Prediction using Machine Learning," in 2021 5th International Conference on Information Systems and Computer Networks, ISCON 2021, 2021. doi: 10.1109/ISCON52037.2021.9702410.

24. E. Miranda, M. Aryuni, C. Bernando, and A. Hartanto, "Application for Early Heart Disease Prediction Based on Data Mining Approach," in Proceedings - 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE 2021, 2021. doi: 10.1109/IC2IE53219.2021.9649419.

25. M. J. A. Junaid and R. Kumar, "Data Science and Its Application in Heart Disease Prediction," in Proceedings of International Conference on Intelligent Engineering and Management, ICIEM 2020, 2020. doi: 10.1109/ICIEM48762.2020.9160056.

26. D. B. Mehta and N. C. Varnagar, "Newfangled approach for early detection and prevention of ischemic heart disease using data mining," in Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019, 2019. doi: 10.1109/icoei.2019.8862544.

27. S. R. M. A. Ayeshmi M. and T. Peiris, "Heart Disease Stages Prediction using Machine Learning.," in 2022 8th International Conference on Big Data and Information Analytics (BigDIA), 2022, pp. 504–511. doi: 10.1109/BigDIA56350.2022.9874198.

28. R. Rajendran and A. Karthi, "Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers," Expert Syst. Appl., vol. 207, p. 117882, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117882.

29. L. Riyaz, M. A. Butt, and M. Zaman, "Ensemble Learning for Coronary Heart Disease Prediction," in 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022, pp. 1–9. doi: 10.1109/CONIT55038.2022.9848292.

30. D. Sharathchandra and M. R. Ram, "ML Based Interactive Disease Prediction Model," in 2022 IEEE

Delhi Section Conference (DELCON), 2022, pp. 1–5. doi: 10.1109/DELCON54057.2022.9752947.