Chapter 15
Prediction of Heart Disease Using Hybrid Machine Learning Technique

Nagaraj M. Lutimath, Chandra Mouli, B. K. Byre Gowda, and K. Sunitha

Abstract Researchers have paid close attention to the field of medicine. Several factors have been blamed for human early mortality by a sizable number of researchers. The relevant research has established that diseases are brought on by a variety of factors, one of which is heart-related illnesses. Numerous scholars suggested unconventional ways to prolong human life and aid medical professionals in the diagnosis, treatment, and management of the cardiac disease. Some practical techniques help the expert make a conclusion, yet every effective plan has limitations of its own. In data mining, support vector machines (SVMs) are an important classification technique. It is a method of supervised classification. It locates a hyperplane to classify the intended classes. A variety of heart-related illnesses make up heart disease. Vascular problems such as arrhythmia, weak myocardium, congenital heart disease, cardiovascular disease, and coronary artery disease are included in this category. A common form of heart disease is coronary artery disease. It causes a heart attack by decreasing the blood supply to the heart. Support vector machines are used in this study to assess the data set from the UCI machine learning repository made up of heart disease patients. Patients with cardiac disease are accurately classified, as expected. Python is used as the programming language for implementation.

Keywords Prediction · Hybrid · Heart disease · SVM · MAE · MSE · RMSE

N. M. Lutimath (✉)
Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Opp. Art of Living, Udayapura Road, Kanakapura, Bengaluru, Karnataka, India
e-mail: nagarajlutimath@gmail.com

C. Mouli
Department of Computer Science and Engineering, East Point College of Engineering and Technology, Jnana Prabha, Virgo Post, Bengaluru, Karnataka, India

B. K. B. Gowda
Department of Information Science and Engineering, Sir M. Visvesvaraya Institute of Technology, International Airport Road, Hunasamaranahalli, Yelahanka, Bengaluru, Karnataka, India

K. Sunitha
Department of Electronics and Communication, Government Women's Polytechnic, Holenarasipura, Karnataka, India

A. Rai et al. (eds.), *Paradigms of Smart and Intelligent Communication, 5G and Beyond*, Transactions on Computer Systems and Networks,
https://doi.org/10.1007/978-981-99-0109-8_15

15.1 Introduction

Heart disease has emerged as a serious health concern for many individuals due to its high mortality rate throughout the world. Detecting cardiovascular disorders including heart attacks, coronary artery diseases, etc. by routine clinical data analysis is a critical task; early detection of heart disease may save many lives. Making informed decisions and precise predictions is made possible by machine learning (ML). The application of machine learning techniques in the medical sector has advanced significantly. By more effectively identifying diseases at an earlier stage, ML can assist in lowering the rate of readmissions to hospitals and clinics. The ability to find and create novel treatments that have a great chance of aiding patients with complex illnesses has also advanced thanks to technology. Machine learning is used in various fields around the world. Healthcare is no exception. Machine learning is very important in determining if movement disorders, heart disease, and other medical conditions are present. With enough information up front, physicians can gain valuable insight and customize diagnosis and treatment strategies for each patient.

People can be impacted by diseases both physically and mentally since acquiring and coping with sickness can change a person's outlook on life. A condition that damages an organism's components but is unrelated to any recent exterior injuries. It is common knowledge that diseases are medical illnesses linked to particular symptoms and indicators. Arteria coronaria disease (blood flow obstruction), cerebrovascular illness, and lower respiratory infections are the three most lethal conditions that affect people. Heart diseases are the most unforeseen and unpredictable. Using machine learning techniques, we can anticipate cardiac disease. Since cardiac illness has a complex character, it requires cautious management. Failure to do so could harm the heart or result in premature death. To identify different types of metabolic syndromes, data mining and the perspective of medical research are employed. Heart disease prediction and data analysis both greatly benefit from data mining with classification.

The main cause of death in the world is heart disease. Heart disease is often recognized by symptoms such as shortness of breath, physical limitations, and swollen feet (HD). Sometimes there are no clinical experts available to treat the coronary condition, and exams take a long period. A specialist typically diagnoses HD after looking at the patient's clinical history and compiling a report on their physical examination. However, the results are frequently wrong.

The medical profession is still far from being able to treat patients suffering from various diseases. The abnormality of the heart, which cannot be seen with the naked eye and manifests itself right away when it reaches its limits, is one of the most deadly. No hospital can afford to have a patient die as a result of poor medical decisions. A suitable and affordable computer-based therapy and support system can be created to help people make wise decisions. The major goal of this study was to develop a model that could analyse historical data from the database of heart illness and deduce unknown information (patterns and correlations) connected to heart disease. It can

answer difficult heart disease detection questions and assist medical professionals in making wise clinical judgments that differ from the norm.

It is challenging to pinpoint cardiac disease due to the multiple risk factors that contribute to it, including diabetes, high blood pressure, excessive cholesterol, an irregular pulse rate, and many others. Finding out the severity of heart disease in humans has been tasked with a number of data mining and neural network methods. Several methods, including the K-Nearest Neighbor Algorithm, Decision Trees, Logistic Regression, and Support Vector Machine, are used to categorise the severity of the illness. Because the health of heart disorders is complex, the disease must be handled with care. Failure to do so could harm the heart or result in premature death. Medical science and data mining perspectives are applied to identify various types of metabolic illnesses.

Non-communicable diseases (NCDs), commonly known as chronic diseases, are diseases that cannot be transmitted from person to person. They often survive long and move slowly. They are the product of a confluence of behavioral, environmental, physiological, and genetic factors. Diabetes, malignancies, chronic respiratory disease, and cardiovascular disease are the four main categories of non-communicable diseases. Noncommunicable diseases (NCDs) are currently a public health concern. In fact, they are responsible for over 70% of deaths on Earth. According to the World Health Organization (WHO)'s first report on the status of noncommunicable diseases in 2010, of the 57 million deaths worldwide in 2008, 36 million (nearly 60%) were mainly caused by non-communicable diseases. Cardiovascular disease, cancer, diabetes, chronic pneumonia.

From 36 to 41 million people died in 2016, accounting for 70% of the global mortality rate. The leading cause of NCD mortality (17.9 million per year) is cardiovascular disease, which is then followed by cancer (9 million), respiratory conditions (3.9 million), and diabetes (1.6 million). The primary risk factors affecting non-communicable illnesses are a sedentary lifestyle, an imbalanced diet, cigarette use, and excessive alcohol intake. The 2020 Sustainable Development Program faces a significant challenge from non-communicable diseases. Reducing the risk factors for these diseases is a crucial control strategy. Preventive medicine holds that risk factors can help make an early diagnosis of patients or offer health advice.

This includes the static analysis and machine learning techniques such as K-Nearest Neighbors, Decision Trees, Logistic Regression, Neural Networks, and Support Vector Machines. These predictive analytics techniques are used to detect fraud, reduce risk, improve operations, simplify marketing, and more. Data mining (decision trees, neural networks, regression, classification, clustering, etc.), machine learning (supervised and unsupervised learning, etc.), and deep learning are used to predict NCDs, especially cardiovascular diseases. Various predictions are made.. Techniques (such as Autoencoder and Softmax) have been created.

The WHO estimates that 12 million people worldwide die prematurely from heart disease each year. Cardiovascular disease is responsible for over 50% of her deaths in the United States and elsewhere. In many countries it is one of the leading causes of death. It is considered the leading cause of death in adults. Heart disease can be classified as either coronary artery disease or cardiovascular disease. The term

"cardiovascular disease" refers to many conditions that affect the heart, blood vessels, and the body's circulatory and pumping systems. A variety of illnesses, disabilities, and deaths are caused by cardiovascular disease. One of the most important and demanding tasks in medicine is the diagnosis of diseases.

Cardiovascular disease is one of the most common causes of death worldwide and in developed countries. Hypertension, obesity, stress, diabetes, alcohol, high cholesterol, and smoking are other risk factors for cardiovascular disease that can be prevented and managed through healthy behavioral adaptations. However, other risk variables such as age, gender, and family history may not be controlled. Early detection of cardiovascular disease can reduce mortality because lack of awareness prevents people from knowing the cause of cardiovascular disease early. is taking place. The disease is usually discovered only in the late stages or after death.

In order to properly diagnose cardiac disease, machine learning is essential. Machine learning techniques include decision trees, neural networks, Naive Bayes classification, genetic algorithms, regression, and support vector machines, to name a few. The decision tree method is used to identify patterns that can be utilized to forecast cardiac disease. With the aid of the Cleveland data set, the C5.0 decision tree technique was completed. When compared to the other algorithms, it has an accuracy value of 85.33% (Ngom et al. 2020) (Maru et al. 2021). It was discovered to be superior to other data mining techniques. Data on the patient was entered via a graphical user interface, and a Weighted Association rule-based Classifier was used to determine whether or not the patient had heart disease. Results indicated that, in comparison to other Associative Classifiers presently in use, the Weighted Associative Classifier offered better accuracy. A classifier that uses probability is called Nave Bayes (Atallah and Al-Mousa 2019). To predict heart disease, medical factors like blood pressure, age, and sex were used. The implementation was done using MATLAB. By lowering the size of the tree, a prediction model that combines pre- and post-pruning of decision tree learning increased classification accuracy (Motarwar et al. 2020).

By utilising the most recent technologies, the health sector can be upgraded, extending the average population's life expectancy. Leading causes of death worldwide include cancer and heart disease. Cardiovascular disease-related deaths are rising alarmingly quickly each year. According to a World Health Organization report for 2016, heart attacks and strokes accounted for 85% of all cardiovascular disease-related deaths globally, accounting for 31% of all fatalities. In both industrialised and developing nations, the rising use of alcohol and tobacco directly raises the risk of heart disease. In industrialised nations like the United States, England, Canada, and New Zealand, the prevalence of obesity is rising, which increases the chance of developing heart-related issues.

When you consider how cardiovascular diseases affect the world's population, a machine-learning model for early identification is quite helpful. To address this growing, enormous problem, ongoing efforts are conducted employing a variety of technological advances. In order to address the ever-expanding health issues, various bioengineering approaches have been created in recent years. The improvement reduction rate is benefiting from the ongoing study in the field.

Regular Neural Networks (NNs) and Convolutional Neural Networks (CNNs) both have a wide range of applications and differ in terms of architecture. The heart disease diagnosis issues in this article were solved using both of the two machine learning models. We put the algorithms into practise, adjusted the settings, and ran a number of tests. We compare the two models' predictive abilities with various parameter values. For the diagnosis of cardiac illness, we used the Cleveland database, which was taken from the UCI learning dataset repository. According to the experimental findings, NNs typically outperform CNNs in terms of prediction accuracy.

Prediction analysis can also make use of additional data mining techniques like regression, neural networks, support vector machines, and genetic algorithms. Support vector machines with linear and sigmoid kernel functions are contrasted in this article. The UCI machine learning data set repository made the dataset accessible. The structure of this study is as follows:

The relevant research is discussed in Sect. 15.2, followed by a discussion of methodology in Sect. 15.3, examples of feature engineering in Sect. 15.4, examples of prediction analysis in Sect. 15.5 and a conclusion in Sect. 15.6.

15.2 Related Work

In order to anticipate the people who will have this disease, experts have focused their attention on heart disease. Usually, hidden patterns in the data set are extracted using a knowledge discovery technique. Data mining is an information extraction process used in knowledge discovery. It is crucial in the process of diagnosing the illness. Data mining techniques are used to categorize the data collection, including decision trees, neural networks, Naive Bayes classification, support vector machines, and genetic algorithms (Kohli and Arora 2018). Using appropriate medical data, decision tree C4.5 and Fast Decision trees were investigated (Basha et al. 2019). The UCI repository was utilized for medical data sets. Decision trees were accurate to a degree of 69.5%, and fast decision trees were accurate to a degree of 78.54%.

Analysis and forecasts of coronary artery heart disease were carried out using a data collection of 335 records representing the various 26 variables (Lin et al. 2020). The data set was pre-processed using the correlation concept. The characteristics were located and extracted using Particle Swarm Optimization (PSO). Fuzzy, decision tree, fuzzy regression, and neural network models were among the models. The data set was used with the neural network model. The accuracy percentage was found to be 77%. It was also used for regression modeling. The accuracy as a result was 83.5%. There were no significant changes in the other fuzzy and decision tree models.

The data set was then optimized using the pre-processing technique. We used K-means clustering, feature extraction and selection with PSO, and correlation. One of the methods, or a combination of them, was used to categorize the data set. The accuracy of the regression model's results was 88.4%. The data set was also subjected

to a hybrid model. The accuracy of classification techniques increased from 8.3 to 11.4%.

With the use of the Gini Index and support vector machines, a second study on the prediction of heart disease was completed (Gavhane et al. 2018). The classification of the data was then finished using the proper classification techniques. Sequential Minimal Optimization (SMO) and Naive Bayes probability classification were the techniques employed for classification. Artificial neural network models and SMO with bagging were also incorporated for analysis. SMO with bagging had an accuracy of 93.4%. Naive Bayes probability classification has a 75.51% accuracy rate. SMO accuracy was 94.08, whereas neural network models had an accuracy of 88.11. The 10-cross-fold validation procedure was used to complete the results' verification.

Using an appropriate medical data set, a Transaction Reduction Method (TRM) Apriori approach was used to diagnose heart illness (https://www.alivecor.com/how-it-works; https://doi.org/10.1007/s10462-01). The outcomes were contrasted with some of the traditional methods. The algorithm produced an accuracy of 93.75%. 92.09% accuracy was attained when SMO was used. 89.11% accuracy was attained when SVM was applied. The accuracy of the C4.5 decision tree was 83.85%, and the accuracy of the Naive Bayes probability classification was 80.15%.

There is no shortage of documentation about the symptoms experienced by patients having heart attacks. However, they are not being used to their full potential to help us predict comparable possibilities in otherwise healthy adults. For illustration: According to the Indian Heart Association, 25% of all heart strokes in Indians happen before the age of 40, and 50% of heart attacks happen before the age of 50. The risk of heart attacks is three times higher in urban areas than in rural ones (Gavhane et al. 2018).

To maintain heart health, many experts advise a balanced diet and moderate exercise. The following are the factors that were taken into account when creating the system for the study and have a high-risk percentage for heart disease:

Age, sex, blood pressure, heart rate, diabetes, and high cholesterol are the first six factors.

AliveKor It is available as a bracelet or a touchpad that connects to your cell phone over a wireless network. Through Bluetooth, the touchpad simulates the patient's ECG on his mobile device. As a result, all the relevant parameters, including blood pressure and heart rate, are readily available. On the wristband, however, the pulse function is shown on the dial through finger contact. Additionally, it may signal atrial fibrillation (https://www.alivecor.com/how-it-works). MyHeart, B. This system uses a variety of on-body sensors to wirelessly transmit physiological data to a PDA. The data is processed, and the analysis is used to provide the user with health suggestions (https://doi.org/10.1007/s10462-01). C. HealthGear HealthGear is a programme for keeping track of the most popular indicators, including physical and lab measurements (Boudi 2016). Fields consist of: Blood pressure, haemoglobin, WBC, RBC, and platelets are among the physical indicators, along with height, weight, and BMI—[Lipids]: Triglycerides, HDL, LDL, and VLDL—[Sugar] Fasting Glucose, HbA1C, and After Meals (Boudi 2016). Fitbit, D This sensor is used to monitor one's health and has functions for detecting heart rate, blood pressure, and calories burned. After

conducting this analysis, we came to the conclusion that Fitbit would be the most convenient and cost-effective way to gather data, while HealthGear would be used for all other metrics.

All of the methods discussed above deal with predictive analysis utilizing conventional techniques. When employing appropriate medical data sets, classification methods such as decision trees, Naive Bayes, support vector machines, or neural networks should be taken into account.

Purusothaman et al. (2015) introduced the common data mining classification approaches such as ANN, fuzzy logic, Neural Networks, Decision trees, data mining genetic Algorithm, and Nearest Neighbor method. In the paper, hybrid applied data mining techniques were suggested. The importance of big data analytics was emphasised in the work by Cheryl Ann Alexander et al. for diagnosing, treating, and predicting chronic diseases (Sharma and Rizvi 2017). The study put forth the concept of IoT and cloud computing technologies in the field of medicine.

Johnson-Coyle et al. (2012) developed an efficient data analysis model for the prediction of severe heart syndromes by utilising a variety of classification techniques. It is common for datasets to have noise features, which can suddenly damage good data. As a result, they aimed to reduce the noise by cleaning and pre-processing the dataset as well as by reducing its dimensionality. They discovered that neural networks may produce results with good accuracy.

Peripartum cardiomyopathy (PPCM) is a form of enlarged cardiomyopathy with an unknown cause, according to Leah Johnson-Coyle et al. Although the rate is low—less than 0.1% of pregnancies—disappointment and death rates are significant, ranging from 5 to 32%, and they occur in already healthy women in the final month of pregnancy and up to 5 months following delivery. While for some women, PPCM progresses to cardiovascular disappointment and even abrupt cardiac passing, for others, the clinical and echocardiography status improves and may return to normal. Clearly linked to the restoration of cardiac function is the guess of PPCM. Failure of the heart to return to its normal size is linked to increased mortality and gloom. As described in Sebastian et al. (2012), dilated cardiomyopathy is characterised by left ventricular enlargement that is associated with systolic brokenness. Right ventricular function can become impaired and diastolic dysfunction can occur. Affected individuals have the risk of both left and right ventricular failure. A fundamentally hereditary or explosive foundation underlies a sizable portion of DCM cases. Although distinct aspects of heart remodelling educate forecast and provide restorative advice, estimation of LV size and launch percentage remains essential to conclusion, risk classification, and treatment. Evaluation of myocardial fibrosis forecasts both the risk of sudden cardiovascular death and the likelihood of LV functional recovery, and may influence the patient's decision about the insertion of a cardioverter-defibrillator. Finding preclinical DCM could greatly save suffering and fatalities. Preclinical DCM detection could significantly save suffering and fatalities by enabling early evaluation of cardio-protective analyses.

Heart disease prediction as a method of medical diagnosis has been the subject of numerous investigations. First, a study using a neural network to analyse data from a self-applied questionnaire (SAQ) in order to create a system for predicting cardiac

disease has been proposed by R. W. Jones, M. Clarke, Z. Shen, and T. Alberti. The study highlights not just typical illness risk variables but also additional SAQ data. By comparing the results of the neural network with the "Dundee Rank Factor Score," which statistically correlates three risk factors (blood pressure, smoking, and blood cholesterol) with sex and age to assess the risk of developing heart disease, the work's validity was confirmed. A multi-layered feedforward neural network that was trained using the backpropagation algorithm was employed in the study. The neural network they employed included three layers: input, hidden, and output. By expanding the neural network's input quantities, the performance was enhanced to a Relative Operating Characteristic (ROC) area of 98%. The best categorization strategy for the intended system, according to Ankita Dewan and Meghna Sharma, who explored numerous techniques for constructing a heart disease prediction system. Additionally, they suggested employing genetic algorithms to overcome the limitation of local minima in backpropagation algorithms. The suggested methodology was designed to be used in the future with accuracy close to 100% or with few mistakes.

S. Y. Huang, A. H. Chen, C. H. Cheng, P. S. Hong, and E. J. Lin have planned and carried out a further investigation on the prediction of heart disease. The learning Vector Quantization Algorithm, one of the Artificial Neural Network learning techniques, was used to train the classification and prediction. Their approach consisted of three steps. The first step was to choose three of the 13 clinical features—age, cholesterol, kind of chest pain, exercise-induced angina, maximum heart rate, fasting blood sugar, number of vessels coloured, old peak, resting ECG, sex, slope, thal, and trestbps—that are more significant than the others. The second used a classification system based on artificial neural networks. Finally, a technique for predicting heart disease was created. The study's results showed a prediction accuracy rate of close to 80% (Motarwar et al. 2020). D. R. Patil and Jayshril S. Sonawane have developed a new Artificial Neural Network technique for heart disease prediction. The Vector Quantization Algorithm uses random order incremental training to train the employed network. The network in use has three levels: input, hidden, and output layers. In the input layer, there were 13 neurons, which is equivalent to the number of clinical data in a database of cardiac diseases. To achieve fewer mistakes and greater accuracy, the neurons of the buried layer could be modified. In the layer of output that indicates whether or not there is cardiac disease, there was only one neuron. Training with a variety of neurons and training epochs increased the system's performance. The outcome demonstrates that, when compared to other researchers, they had the best accuracy (85.55%), as claimed in the report (Kohli and Arora 2018). Another work by Majid Ghonji Feshki and Omid Sojoodi Shijani uses a feature selection and classification approach using a particular dataset to predict cardiac disease. There were three steps in the suggested method. The procedure of splitting the dataset into two subsets as sick and healthy people was the initial step. The complete characteristics were divided into 8192 subsets in the second step. The optimal subset with the highest accuracy was identified in the third stage by combining the PSO algorithm with the Feed Forward Backpropagation Algorithm, a classifier algorithm. Four classifier methods were employed in the methodology: C4.5, Multilayer Perceptron, Sequential Minimal Optimization, and Feed Forward Backpropagation. The most

effective approach was identified as a neural network with the PSO algorithm using feature selection and backpropagation. The study's results showed a 94.94% accuracy rate. The goal of a study by R. R. Manza, Shaikh Abdul Hannan, R. J. Ramteke, and A. V. Mane is to predict the diagnosis of cardiac disease using an artificial neural network as a classifier. There were 5 steps in the suggested process. Step 1 involved gathering information regarding prescription medications and heart disease patients. Step 2 involved converting heart disease symptoms and medications into binary form (0 or 1), where 1 denotes the presence of a symptom or medication. In phase 3, the Radial Basis Function was trained. Step 4's performance evaluation of the classifier made use of testing data. In step 5, the Radial Basis Function administered the patients' drugs. The used network has three layers: input, hidden, and output. Information processing is not the duty of the input layer. Distribution of the input vectors to the hidden layer is the only duty of the input layer. There were several Radial Basis Function units in the hidden layer. 97% accuracy was found in the study. It was mentioned that the presented method might be expanded using the generalised regression neural network.

Syed Umar Amin, Dr. Rizwan Beg, and Kavita Agarwal have presented a hybrid approach using genetic algorithms and artificial neural networks to forecast cardiac illness based on risk factors. The backpropagation algorithm was employed to train the neural network. The backpropagation algorithm's two primary flaws have been identified. The first issue is that finding globally optimal beginning weights is essentially impossible. The backpropagation algorithm's slow convergence rate is the second issue. This issue was resolved by utilising a genetic algorithm to enhance the performance of an artificial neural network by optimising its connection weights. 12 input nodes, 10 hidden nodes, and 2 output nodes made up the neural network that was utilised in this investigation. The results show that the training accuracy is 96.2% and the acquired validation accuracy is 89% (Gavhane et al. 2018). A study by Jayshril S. Sonawane and D. R. Patil aims to use artificial neural networks to forecast heart illness. A multilayer perceptron neural network is used by the system. The suggested system consisted of two steps. 13 clinical data were received as input during the first phase, and the backpropagation algorithm was then used to train the network. The input, hidden, and output layers made up the network. There were 13 neurons in the input layer, matching the number of clinical data from a database of cardiac illnesses. To achieve low error and high accuracy, the hidden layer's neurons can be altered. There was only one neuron in the output layer that indicated the presence or absence of cardiac disease. 98% is the accuracy rate that the study found. Another study by Usman Qamar, Saba Bashir, and M. Younus Javed aims to predict cardiac disease. The suggested approach employs a hybrid model that combines Decision Tree, Support Vector Machine, and Naive Bayes. These three classifiers succeeded in obtaining the majority voting scheme. The suggested strategy included two steps. The first one produced the results of every three classifiers. The second included adding the choices together to create a new model using a majority vote system. The findings demonstrate that the study's accuracy rate is significantly higher than that of the competition. The study's findings for predicting heart disease included 74% sensitivity, 82% accuracy, and 93% specificity (Karayõlan and Kõlõç 2017).

To enable effective heart attack prediction, Patil and Kumaraswamy (2009) suggested removing notable patterns from the dataset. The clustering method K-Means was used. The weight of each item was calculated using the MAFIA algorithm. Based on the calculated weights, patterns with values above the threshold were considered for prediction. Using a data set of 15 features and data mining techniques, including ANNs, time series, clustering rules, and association rules, Soni (2011) report on heart disease prediction. The paper recommended the adoption of genetic algorithms to improve accuracy while reducing the amount of data.

To create models for heart disease prediction, Ananey-Obiri and Sarku (2020) used ML approaches such as linear regression, decision trees (DT), and Gaussian Naive Bayes (GNB). The models were created using a k-fold cross-validation method, and their effectiveness was evaluated using receiver operator characteristic (ROC) curves. The analysis for this study employed the UCI dataset of heart disease patients. DT classifier model accuracy was 79.31%, GNB accuracy was 76%, and LR accuracy was 82.75%.

15.3 Methodology and Data Set Analysis

Data mining is the process of extracting knowledge from secret data sets. Multiple sources are used to gather multi-dimensional data, which is then pre-processed and formatted appropriately. Then, these data are subjected to data mining techniques for additional classification.

The goal of machine learning is to develop effective software applications that can automatically access and utilise data. Machine learning is a type of system learning process that gives a system the ability to operate automatically or by itself with the help of correct training, improving system performance and experience effectively and without the need for human intervention. Machine learning will enable training on data sets using efficient learning techniques. The rules and obligations that result from these algorithms will be based on conclusions drawn from the data. The system may build many system models while being trained using various datasets and the same learning technique.

A. The variance in the neighbours within a class determines the class of a certain data point using the K nearest neighbours (KNN) classifier. Test scores with neighbours ranging from 1 to 20 are utilised to calculate test scores.
B. Based on the class values that will be given to each data point, the decision tree classifier creates a tree. One to thirty-point increments is used to count the features.
C. Each tree in the random forest splits out into a class prediction, and the model is made up of a collection of distinct decision trees. The bagging method is used in the Random Forest approach to add more randomness and diversity to the feature space. In other words, it randomly samples elements of the predictor space rather than looking greedily for the best predictors to generate branches.

This increases the variety and lowers the variance of the trees at the expense of an equal or larger bias. This procedure, also known as "feature bagging," is what results in a more reliable model.

D. Naive Bayes a statistical classifier makes no assumptions about the relationship between attributes. A supervised algorithm is the Naive Bayes classifier. It is a simple classification method that uses Bayes theorem. Strong (Naive) independence is assumed. among qualities. A Bayes theorem formula to calculate the percentage The Predictors are not connected to one another neither relate to another nor at least one another. All the attributes alone participate to the in order to maximise it. It identifies conditional independence, which is independent of the values when compared to the values of other characteristics but assumes an attribute value on a specific class.

E. An input layer, a number of hidden layers, and an output layer make up a regular neural network (NN). An input array is transformed by being placed in the input layer, passing through several hidden layers, and then receiving the prediction result from the output layer. Each layer of a NN is composed of a collection of neurons, with every layer being fully connected to every neuron in the layer before it. Consider a neuron with n inputs. Each input Xi is multiplied by the corresponding weight Wi, the sum is computed, the activation function is run, and the output is the result.

   The human brain, which has remarkable processing power due to its network of interconnected neurons, serves as the model for artificial neural networks (ANN). ANNs are created utilising the perceptron, a type of fundamental processing unit. The single-layer perceptron algorithm handles issues that may be divided into linear segments. Multilayer Perceptron Neural Network can be used to solve issues that cannot be resolved linearly (MLP). Input, hidden, and output layers are among the many layers present in MLP. The multilayer perceptron neural network was used in the creation of the suggested heart disease prediction system. Three layers make up the planned ANN: the input layer, the hidden layer, and the output layer.

   13 neurons were intended to be in the input layer. It was decided that the number of neurons would match the number of attributes in the data set. • Three neurons were intended to be housed in Hidden Layer. This sum was chosen as the starting point. By comparing their performances and then choosing the best one, the number was modified by going up one at a time until it reached the number of input layer neurons. This strategy is based on one of the principles of machine learning, according to which the number of neurons in the hidden layer should be equal to the average of those in the input and output layers. • Two neurons were included in the Output Layer's architecture. The created NN is a classifier that is now operating in machine mode.

F. The architecture of convolutional neural networks (CNNs) differs from that of traditional neural networks (NNs) in a number of ways. In each layer, a CNN first arranges its neurons in three dimensions (width, height, depth). Second, not every neuron in a layer is linked to every other neuron in the layer above.

The classifier and the feature extractor are two more of the CNNs' component parts. To find the features in the feature extractor section, the network will do a series of convolutional (by convolution layers) and pooling (by pooling layers) operations. On the extracted features in the classifier section, the fully connected layers (FCNet) function as a classifier. It also assigns a probability to the input array in order to illustrate what it predicts.

G. Support vector machines (SVMs) are supervised learning techniques that, unlike the C4.5 algorithm, do not rely on decision trees to complete tasks. The likelihood of classification errors is reduced when support vector machines are used.

15.3.1 Experimental Procedures

Significant supervised classification techniques include SVM. The target classes are classified using a hyperplane. Identification of the hyperplane separating one class from the other classes is the process of classification. Although the SVM takes a very long time to train, it is quite accurate at predicting the target classes.

15.4 Feature Engineering

To study the UCI machine learning archive classification workflow dataset for the Cleveland epileptic recognition dataset is performed. The training dataset and testing dataset are the two sets created from the data set. On the training data, associated feature engineering is carried out, and the resulting model is applied to the test data to make predictions. The following is the job description for the next position: Task Description: "Predict the value of patients with heart disease" by analysing the number of patients with heart disease consisting of 303 records. The attributes f_ca, f_thal, f_oldpeak, f_thalac and f_cp. attributes are selected based on the values that correlate with the target attribute num.

Data attributes are,

f_age—age data base feature is given in years.

f_sex—features in the Gender database are classified with a male value of 1 and a female value of 0.

f_cp—the scores for angina, atypical angina, nonanginal pain, and asymptomatic pain in the chest pain database are 1, 2, 3, and 4, respectively.

f_trestbps—this database characteristic attributes the resting blood pressure (BP) at the time the subject was admitted to the hospital, expressed in mmHg.

f_chol—this database characteristic is serum cholesterol expressed in mg/dl.

f_fbs—this database characteristic is the Fasting Blood Glucose >120 mg/dL attribute, which is digitized as 1, 0 for true and false.

f_restecg—this database feature is resting ECG results expressed as 0.1 values for normal and ST T-T wave abnormalities (T wave inversion and/or ST elevation or depression >0.05 mV) standard.

f_thalac—this database function applies to the patient's maximum heart rate.

f_exang—this database attribute pertains to exercise-induced angina and is numerically 1 and 0 for categorical values Yes and No.

f_oldpeak—this database feature relates to exercise-induced ST depression versus rest.

f_slope—this database characterization of ST segment slopes during peak exercise expressed in terms of ascending, flattening, and descending slopes with values of 1, 2, and 3, respectively.

f_ca—this database function counts the number of major vessels ranging from (0 to 3) by fluoroscopic staining.

f_thal—this database feature applies to the types of cardiac defects, with values of 3 for normal, 6 for fixed defects, and 7 for reversible defects.

f_num—this database feature is used to predict patients suffering from heart disease.

The 303 input tuples are divided into 91 tuples for the test data set and 212 tuples for the training data set. Equations 15.1 and 15.4 are used to create the training dataset, which is then performed in Python.

$$\text{X\_train} = \text{df.iloc}[\,0 : 212\,,\ :\,] \tag{15.1}$$

$$\text{y\_train} = \text{np.array}(\text{X\_train.iloc}[:, -1]) \tag{15.2}$$

$$\text{X\_test} = \text{df.iloc}[\ \text{int\_a:303}\ ,\ :\ ] \tag{15.3}$$

$$\text{y\_test} = \text{np.array}(\text{X\_test.iloc}[:, -1]) \tag{15.4}$$

$$\text{cls} = \text{svm.SVC}(\text{kernel} = \text{`rbf'}, \text{gamma} = \text{`auto'}) \tag{15.5}$$

$$\text{cls.fit}(\text{X\_train},\ \text{y\_train}) \tag{15.6}$$

The training set derived from Eqs. 15.2 and 15.3 is the variables used in the svm function in Eq. 15.4. Kernels used for splitting include linear, polynomial, sigmoidal, and radial values. Equation 15.4 above uses a kernel with linear values. Kernels with sigmoids are also used for additional analysis. Equations 15.5 and 15.6 are used to fit the training set.

15.4.1 Performance Analysis

Mean Absolute Error (MAE), Sum Squared Error (SSE), and Mean Squared Error are a few essential metrics used to evaluate a dataset's performance. The average absolute difference between an instance's actual value and expected value is referred to as "MAE." By adding the squares of the dataset's actual instance values minus its projected instance values, the SSE is determined. The MSE of a dataset is calculated as the average of the squares of the actual instance values less the projected values (Japp, 2016).

15.5 Predictive Analysis

Preprocessing the data and evaluating the missing attributes using the average of the attributes is done before analyzing the predictions. We then calculated the power scores MAE, SSE, and MSE for the entire heart disease data set as shown in Table 15.1 The global dataset considered 80, 70, and 60% of the training dataset. We found lower values for MAE, MSE, and RMSE for the 70% test data set compared to 80% and 60%. Therefore, we treat 70% of the training dataset as a shared dataset. From Table 15.2, it can be seen that MAE, MSE and RMSE are minimum when the f_ca values are greater than the mean f_ca value. As a result, when the value of f_ca values greater than the mean f_ca value a higher degree of accuracy. Table 15.3 shows that MAE, MSE and RMSE are lower when f_thal values lesser than mean or equal to mean f_thal values. Therefore, the f_thal values lesser than or equal to mean f_thal value is better at predicting. Now Table 15.4 shows that MAE, MSE and RMSE are minimized when f_oldpeak values lesser than or equal to mean f_oldpeak values, thus f_oldpeak values lesser than or equal to mean property gives better predictions when the f_oldpeak values greater than the range. Now, looking at Table 15.5, it can be seen that MAE, MSE and RMSE are low when f_thalac values greater than mean values. Thus f_thalac values greater than mean values are better in prediction accuracy. Table 15.6 shows that MAE, MSE and RMSE are equal for both the cases. Thus f_cp prediction for a given range is the same. Analysing the Tables 15.2, 15.3, 15.4, 15.5 and 15.6, we find that MAE, MSE and RMSE lowest value, f_thal values lesser than or equal to the mean f_thal value has the least value. Thus, this is considered as the attribute property for prediction in the heart disease data set.

Table 15.1 MAE, SSE and RMSE for overall test dataset

Error type	60% training dataset	70% training dataset	80% training dataset
MAE	0.9426	0.9120	1.0
MSE	2.254	2.120	2.311
RMSE	1.501	1.456	1.52

Table 15.2 MAE, SSE and RMSE for f_ca attribute

Error type	f_ca values lesser than or equal to mean f_ca	f_ca values greater than mean f_ca
MAE	0.9426	0.9120
MSE	2.254	2.120
RMSE	1.501	1.456

Table 15.3 MAE, SSE and RMSE for f_thal attribute

Error type	f_thal values lesser than or equal to mean f_thal	f_thal values greater than mean f_thal
MAE	0.5344	1.575
MSE	1.224	3.696
RMSE	1.106	1.922

Table 15.4 MAE, SSE and RMSE for f_oldpeak attribute

Error type	f_oldpeak values lesser than or equal to mean f_oldpeak	f_oldpeak values greater than mean f_oldpeak
MAE	0.6065	1.533
MSE	1.262	3.866
RMSE	1.123	1.966

Table 15.5 MAE, SSE and RMSE for f_thalac attribute

Error type	f_thalac values lesser than or equal to mean f_thalac	f_thalac values greater than mean f_thalac
MAE	1.355	0.4782
MSE	3.266	1.0
RMSE	1.8073	1.0

Table 15.6 MAE, SSE and RMSE for f_cp attribute

Error type	f_cp values lesser than or equal to mean f_cp	f_cp values greater than mean f_cp
MAE	0.9120	0.9120
MSE	2.120	2.120
RMSE	1.456	1.456

15.6 Conclusion

A major challenge for machine learning has been to effectively classify medical datasets. When associations and patterns are rapidly extracted from these complex health datasets, the diagnosis, prediction, and accuracy of cardiovascular disease outcomes can be improved. Processing cardiac information with raw healthcare data enables long-term lifesaving and early detection of heart disease abnormalities. Machine learning techniques were applied in this study to process the raw data and provide fresh and unique insights into heart disease. Predicting heart disease is difficult and important in the medical industry. However, early detection of the disease and taking preventive measures as soon as possible can significantly reduce mortality. More complicated models and model combinations are required to increase the accuracy of detecting the early start of heart illnesses. In this paper the hybrid support vector machine is used for a prediction made for the heart disease taking the UCI machine learning Cleveland data set repository. The attributes f_ca, f_thal, f_oldpeak, f_thalac and f_cp attributes are selected based on the values that correlate with the target attribute num. MAE, MSE and RMSE are calculated considering the data set attributes. We find that f_thal values lesser than or equal to mean f_thal value has the least value has a higher chance of prediction than the other attributes in a given range. In future other prediction methods such as artificial neural networks, genetic algorithms, decision tress and deep learning procedures will be utilized for prediction.

References

AliveKor [Online]. https://www.alivecor.com/how-it-works

Ananey-Obiri D, Sarku E (2020) Predicting the presence of heart diseases using comparative data mining and machine learning algorithms. Int J Comput Appl 975

Atallah R, Al-Mousa A (2019) Heart disease detection using machine learning majority voting ensemble method. In: 2019 2nd international conference on new trends in computing sciences (ICTCS), Amman, Jordan, IEEE, pp 1–6

Basha N, Ashok Kumar PS, Gopal Krishna C, Venkatesh P (2019) Early detection of heart syndrome using machine learning technique. In: 2019 4th international conference on electrical, electronics, communication, computer technologies and optimization techniques (ICEECCOT), Mysuru, India, 11th June, pp 1–5

Boudi FB (2016) Risk factors for coronary artery disease [Online]. https://emedicine.medscape.com/article/164163-overview

Gavhane A, Pandya I, Kokkula G, Devadkar K (2018) Prediction of heart disease using machine learning

Japp AG, Gulati A, Cook SA, Cowie MR, Prasad SK (2016) The diagnosis and evaluation of dilated cardiomyopathy. J Am Coll Cardiol. American College of Cardiology Foundation Published by Elsevier

Johnson-Coyle L, Jensen L, Sobey A (2012) Peripartum cardiomyopathy: review and practice guidelines. Am J Crit Care 21(2)

Karayõlan T, Kõlõç Ö (2017) Prediction of heart disease using neural network

Kohli PS, Arora S (2018) Application of machine learning in disease prediction. In: 2018 4th international conference on computing communication and automation (ICCCA, Greater Noida, India, 29th July, pp 1–4

Lin C-H, Yang P-K, Lin Y-C, Fu P-K (2020) On machine learning models for heart disease diagnosis. In: 2nd IEEE Eurasia conference on biomedical engineering, healthcare and sustainability 2020, Tainan, Taiwan, pp 158–161

Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Online, 25 March 2017. https://doi.org/10.1007/s10462-01

Maru A, Sharma AK, Patel M (2021) Hybrid machine learning classification technique for improve accuracy of heart. In: Proceedings of the sixth international conference on inventive computation technologies [ICICT 2021] IEEE Xplore Part Number: CFP21F70-ART. ISBN: 978-1-7281-8501-9, pp 1107–1110

Motarwar P, Duraphe A, Suganya G, Premalatha M (2020) Cognitive approach for heart disease prediction using machine learning. In: 2020 international conference on emerging trends in information technology and engineering (ic-ETITE), 27 April, Vellore, India, pp 1–5

Ngom F, Fall I, Camara MS, Alassane BAH (2020) A study on predicting and diagnosing noncommunicable diseases: case of cardiovascular diseases

Patil SB, Kumaraswamy YS (2009) Extraction of significant patterns from heart disease warehouses for heart attack prediction. IJCSNS 9(2):228–235

Purusothaman G, Krishnakumari P (2015) A survey of data mining techniques on risk prediction: heart disease. Indian J Sci Technol 8(12). Alexander CA, Wang L (2017) Big data analytics in heart attack prediction. J Nurs Care 6(393):1168–2167

Sebastian VB, Unnikrishnan A, Balakrishnan K (2012) Grey level co-occurrence matrices: generalisation and some new features. Int J Comput Sci Eng Inf Technol (IJCSEIT) 2(2):151–157

Sharma H, Rizvi MA (2017) Prediction of heart disease using machine learning algorithms: a survey

Soni J et al (2011) Predictive data mining for medical diagnosis: An overview of heart disease prediction. Int J Comput Appl 17(8):43–48