Computer-aided Design and Diagnosis Methods for Biomedical Applications

Edited by

Varun Bajaj and G.R. Sinha

Contents

9 Medical Image Classification Algorithm Based on Weight Initialization-Sliding Window Fusion Convolutional Neural Network

Ankit Kumar, Pankaj Dadheech, S. R. Dogiwal, Sandeep Kumar, and Rajani Kumari

CONTENTS

9.1   INTRODUCTION

As stated by the Global Healthiness Society, nearly 17.9 million individuals worldwide die each year, and the causes of death include many diseases associated with heart disease, which are caused by heart attacks and strokes [1]. There may be many genetic facets throughout which four generations last with a solitary form of cardiac ailment. Individually the leading ordinary causes of heart ailments are elevated blood pressure, fasting blood sugar, diabetes, cholesterol, BMI, and heart rate. People with diabetes have an increased probability of developing several different types of health issues, including heart disease. Over time diabetes increases the blood sugar, and elevated blood sugar harms the blood vessels of the heart and other organs, causing other health problems. This means that diabetes doubles the rate of heart disease. Unhealthy diet, lack of sleep, physical inactivity, depression, obesity, or family history could be factors causing a heart ailment. Monitor blood load, maintain cholesterol and triglyceride levels in control, maintain a balanced lifestyle, work out frequently, maintain cholesterol as well as triglyceride levels in management, and maintain fear is the protection of ticker disease. With the popularity of machine learning in the medical industry, we are able to help detect and prevent heart diseases. To enhance the quality of patients' lives, extracting useful information from medical data reports helps with better decision-making in data mining. An individual human being dies every 35 seconds owing to a heart ailment. They are affected by heart diseases in young people aged 20–30 years. Diagnosing heart ailments is an extremely important job in the field of healthcare.

Removal of evidence sets out the methodology and capacity to modify details on medical facts that are keen on the handy management method [2]. The classification consists of 13 medicinal specifications such as sexual characteristics, blood pressure, cholesterol, angina, heart rate, and obesity.

The heart is a fibrous body part that drives the blood around the body in sequence to deliver nutrients through the blood vessels, in addition to getting rid of the metabolic waste of the body [3]. Throughout a regular pulsation, blood circulates from tissues plus lungs keen on atria after that into ventricles. The wall within the heart vociferates the inter-ventricular septum to keep the blood on the left and right. Two spouts between the atria and ventricles keep blood from beginning on the atria towards the back. The tricuspid regulator opens hooked on the right ventricles in addition to the bicuspid regulator which opens hooked on the left ventricle. Muscular pain material in situ vociferate Chordae tedious grip valves in a strong ventricle contact. Blood going through the ventricles pass through the pulmonary valves in the right ventricles then the pulmonary trunk in addition to the aortic regulator linking the left ventricle as well as the aorta.

The heart power vociferates myocardium is differing in a distinctive outline in series drive blood well. The surface of the myocardium enfolds the region of the inferior section of the heart. They twist as well as pass in diverse directions to push blood throughout the heart. Leader cells produce the voltaic signal of the heart. Myocyte agreement vociferates the heart power like a cluster [4]. On the left as well as the right part, which functions jointly as a double pump, the heart is alienated keenly. The blood moves through the left ventricles, which contract and push blood out of the heart through the aorta to nourish the tissues. The primary branch of the aorta is the coronary artery which provides heart strength with oxygen and supplements. At the peak of the aorta, the artery penetrates to send the blood to head as well as arms and provide blood to the rest of the body. Heartbeats range from 60–100 BPM, i.e., 5 Quarts per minute.

9.1.1 HEART AILMENT

The heart needs strength to pump the blood; arteries provide blood to the heart muscle, in addition to the valves, which certify that the blood inside the heart is pumped in the correct route. Problems can occur in several of these areas.

9.1.1.1 Coronary Heart Ailment

There is an amount of work to manipulate the output of the heart, many of which are tapering or obstruction of the coronary arteries because of coronary artery disease, usually caused by atherosclerosis. Atherosclerosis (occasionally vociferate the "hardening" or else "pressing" of the arteries) produces cholesterol along with pinged deposits (called plaques) which lies along the internal ramparts of the arteries. This plaque may confine the bloodstream to the heart strength via closing the artery or with irregular arterial tone and role. With sufficient blood delivery, the heart is filled with oxygen along with the essential nutrients that are required for it to function correctly. This can be the reason for the chest pain called angina. If the blood delivers to a part of the heart that is completely blocked, or if the energy needed by the heart exceeds its blood supply, a heart attack can occur.

9.1.1.2 Peripheral Artery Diseases

Peripheral artery ailments are common circulatory problems in which tapering arteries lessen the stream of blood to the organs. When you have peripheral artery disease (PAD), your extremities—typically your feet—don't obtain enough blood to keep up with the demand.

9.1.1.3 Innate Heart Ailment

About 1% of surviving births show a deficiency in the heart structure that occurs during fetal development consisting of simple "holes" or narrowed heart valves in the heart. Faults may be in the heart ramparts, heart valves, and arteries, as well as the veins close to the heart [5]. These are capable of hindering the usual flow of blood through the heart. Blood flow may be slow, misdirected, or misplaced, or completely blocked.

9.1.1.4 Cardiomyopathy Diseases

Cardiomyopathy is known as a muscle ailment. With cardiomyopathy, the heart has difficulty pumping the blood. For this reason, blood can pool in the lungs and parts of the body. This manifests as swelling in the abdomen, feet, and legs. Cardiomyopathy can be acquired as a result of another disease or inherited [6].

9.1.1.5 Other Cardiovascular Diseases

- Heart malfunction: Heart malfunction, occasionally identified as heart failure, happens when the heart is unable to pump the blood. Certain conditions in your heart, such as narrow arteries (coronary artery ailment) or high blood pressure, make it extremely feeble [7].
- Arrhythmia: A cluster of cells in the heart vociferates a cardiac transmission classification, which uses electric pulses to manage the speediness as well as the pulse of each beat. A cardiac arrhythmia occurs when electrical impulses in the heart do not function properly [8]. For example, several individuals experience an abnormal heartbeat, which may possibly feel similar to a racing heart.
- Heart valve problem: In heart valve disease, one or more valves in the heart do not function properly. There are four hydrants in the heart that maintain the blood flow in. In some cases, one or more valves do not open or close appropriately. This can cause blood flow to your body to be obstructed through your heart [9].

9.1.2 Fact Removal

Fact removal is a procedure used for information discovery or mining of a large number of facts known as a KDD (knowledge discovery from data) [10]. Fact removal is used in healthcare and business, analyzing a huge set of facts to identify helpful and stable patterns. In the medical trade, in particular, fact removal can be used to reduce costs and increase efficiencies, improve the quality of a patient's life, and perhaps most importantly save the lives of new patients. Processes such as fact removal involve many functions, such as forecast, clustering, and time-series study.

The medical trade has many sources of data, such as electronic health check records, managerial information, and erstwhile benchmarking results. Nowadays fact removal is used to provide treatments at an effective cost, in order to help predictors use various diseases with clinical forecasts and recommendations from physicians in different decision-making [11, 12].

9.1.2.1 Classification Algorithm

This section presents an introduction to the classification algorithm used in our comparative study. As we stated in the introduction, the reason for this study is to scrutinize the stability and presentation of categorization algorithms for predicting patient status using previously known patient data along with cardiovascular statistics [13]. The classification algorithm employed was a decision tree, naïve Bayes, random forest, logistic regression, support vector machine (SVM), and *k*-NN.

Decision Tree: A hierarchy of decisions is seen as individual approaches intended for a creator of categorization. This is equivalent to a flowchart where the interior join describes the position of a feature, every branch acts for the resulting position with every leaf node representing a group label, and then all the attributes are decided after the calculation is performed. The passage as of source to leaf represents the categorization rule. In the medical field, the decision tree determines the sequence of the characteristics. First, it produces a set of solved cases; then the entire set is then divided into training sets and test sets. The decision tree is used to find the accuracy of the obtained solution.

Naïve Bayes: naïve Bayes is a prospect classifier based on Bayes' theorem, applying tough autonomy assumptions between the features. It is easy to construct a naïve Bayesian model without complex recurrence parameter estimates, which makes it useful in the medical field for the analysis of cardiac patients. In spite of its ease, the naïve Bayesian classifier frequently performs well, and it is extensively used as it frequently outperforms highly complicated categorization methods [14].

Random Forest: Heart ailments are the leading cause of early death worldwide. Numerous fact removal methods were used by researchers to assist physical condition professionals to forecast heart ailments. A random forest is a group learning algorithm for medical applications. It is the same as a decision tree, although the algorithm builds a forest of decision trees by way of the location of randomly selected features. Computing has the advantage of improving efficiency while improving forecast accuracy without a significant increase in computing costs. It can also predict thousands of other variables.

Support Vector Machine: The SVM projected by Vapnik as well as Cortes has been fruitfully useful to sex categorization exertion for a lot of researchers. A SVM classifier is a linear arrangement where different hyperplanes are selected to diminish the normal categorization mistakes of the hidden analysis pattern. A SVM consists of a robust arrangement so as to recognize two modules. In training the SVM classifies the picture analyzed into a group that has the greatest space at the closest point. The SVM training algorithm creates a replica that predicts whether the experiment figure falls into one group or a different group. A SVM requires an enormous quantity of preparation facts to choose a single range and computational price, while we still regulate ourselves in order to recognize a common currency. A SVM is a knowledge algorithm meant for categorization. It tries to discover the best stripping hyperplanes, so the normal categorization error for the hidden pattern is reduced. Inputs meant for linearly non-separable facts are mapped onto a high-dimensional attribute space where they are unconnected to hyperplanes. This outcrop in a high-dimensional attribute space can be improved by using a kernel. Further, a set of training samples in addition to an equivalent decision value $-1, 1$ SVM aim to discover the most distinct separation hyperplanes given by $WTX + B$ so as to maximize the space between the two modules [15].

Logistic Regression: Logistic regression is essentially used to categorize low dimensional facts with non-dimensional limits. In addition to providing dissimilarity in the proportion of the dependent relative variable, and the position of a person varies according to its significance. So, the major purpose of logistic regression is

to decide the outcome of every variable accurately. In addition, logistic regression is known as a logistic model that gives two categories of a definite component, such as beam or gloomy, slim or fit.

k-nearest neighbor: A k-NN is a basic calculation that stores all cases and orders new cases dependent on a likeness measure.

Classification methods are the most extensively used algorithms in the healthcare field because they help to predict the status of patients by classifying their records and locating the class that matches. Classification is referred to as a supervised learning technique that requires initial classifying of the data into classes or labels. These data are then entered into a classification algorithm in order for it to learn. In particular, relationships between attributes need to be explored by an algorithm to predict the outcome. In this progression, the arrangement calculation makes a classifier from a preparation set made up of dataset tuples and their comparable class marks. Each of the tuples that comprise a preparation set is alluded to as a classification or class. When a new case arrives the developed classification algorithm is used to classify it into one of the predefined classes. The term specifies "how to have a good algorithm" which is called the prediction accuracy. For example, trained medical databases will provide much more relevant information of a previously recorded patient, whether or not patients have heart disease [16].

In this work, this administered AI idea was used for making the predictions. The correlation of particular highlights utilizing diverse AI methods, like support vector machine, decision tree, random forest, naïve Bayes, and k-nearest-neighbor classifiers, were utilized for anticipating a coronary illness or not [17]. The examination was done with a few degrees of cross-approval, and a few rate split assessment strategies separately. The Cleveland and Statlog coronary illness dataset from the UCI AI archive was utilized for coronary illness predictions. The predictions were completed by an older model using arrangement calculations. Guileless mathematician used chance to predict coronary disease, SVM was used for order and relapse operations, random forest operates with a modified decision tree, k-NN was used to locate nearest neighbours. Such estimates suggest an alternative precision. In order to acquire greater accuracy, we sought to tune our methods, which would be useful for increasingly accurate forecasts.

This research addressed the challenge of improving predictive models for heart disease patients and providing timely feedback. A significant problem facing medical management is the provision of precious amenities at a sensible cost. Poor decisions concerning the patient can lead to unwanted as well as permanent consequences. Similarly, medical amenities should diminish health testing costs. A fact removal replica can achieve these results using real-world data furthermore facilitating correct decision creation [18].

The Cleveland and Statlog heart disease datasets from the UCI machine learning repository were exploited for the creation of heart ailment predictions in research work. The major purpose of this learning was to estimate whether a patient was affected by a heart ailment or was not using diverse machine learning algorithms but relying on eligible datasets. We discovered associations among dissimilar attributes, predictions, and drew performance comparisons as well. We forecast heart danger

and possibilities for potential situations that assist in making informed choices after receiving an apparent thought about our intended evidence removal procedures and evaluating implications. Recently, the health industry released data on the predictions of health worldwide. When the medical data is huge, an machine learning algorithm is used for analysis.

9.2 REVIEW OF LITERATURE

Machine learning techniques are used for finding new rules and patterns, which are extracted from huge quantities of information. Fact removal helps to predict disease. Analysis of disease consists of the performance of a number of tests on patients. However, the data mining technique reduces the number of tests in healthcare. With a reduced number of tests, the wastage of time and money reduces. The data mining of heart disease is important as it provides support to doctors by informing them which attribute or feature is important, which helps in diagnosing heart ailment more efficiently. Many fact removal methods are used in the judgment of heart disease, including naïve Bayes, decision trees, neural networks, bagging algorithms, as well as support vector machines for viewing the dissimilar levels of accuracy.

Otoom et al. [2015] constructed an intelligent classifier for predicting heart problems by means of clinical data and machine learning algorithms, entered in by the user or physician. This analytic part is included in the mobile phone function with concurrent monitoring that continually monitors and raises the alarm when an emergency occurs. Our results suggest that the proposed diagnostic component proved to be successful with cross-validation tests, with a classification performance accuracy of over 85%. In addition, the monitoring algorithm provided a 100% detection rate [2].

Amin et al. [2013] presented a method for the forecast of heart problems by means of the main risk factors. The method included two pieces of fact removal equipment: neural networks and genetic algorithms. The implemented mixture scheme used a worldwide optimization for the benefit of a hereditary algorithm to initialize the neural system load. Knowledge is faster, further constant, and correct than back proliferation [3]. This scheme was implemented in Matlab in addition to predicting heart disease risk with an accuracy of 89%.

Desai et al. [12] utilized the Cleveland dataset to evaluate the precision of characterization models for the prediction of coronary illness. An investigation of parametric and non-customary methodologies for characterizing coronary illness was introduced. Two arrangement models, the back-dissemination neural system (BPNN) and calculated relapse (LR) were utilized in the examination. The created grouping model will help specialists to make more compelling clinical choices. A ten-overlap cross-approval strategy was utilized to gauge the impartial estimation of these arrangement models [12].

Dangare et al. [2012] broke down the prescient frameworks for coronary illness by utilizing a more prominent number of information attributes. In this framework, 13 qualities, for example, sex, circulatory strain, and cholesterol, were utilized to appraise the likelihood of the patient getting a coronary illness. Up until this point,

13 highlights have been utilized in the forecast. Two additional highlights were added to this examination paper, for example, corpulence and smoking. Information mining order systems on the coronary illness database, to be specific choice trees, naïve Bayes, and neural networks were investigated. The presentation of these strategies was looked at based on precision [16]. As indicated by our outcomes, the exactness of neural systems, choice trees, and naïve Bayes were 100%, 99.62%, and 90.74%, respectively. Our investigation proposed that of these three grouping models the neural system predicts coronary illness with the most elevated precision.

Parthiban et al. [2007] provided another methodology dependent on a coactive neuro-fluffy infusion framework and CANFIS that was displayed for the prediction of coronary illness. The CANFIS model, along with neural net versatile abilities and fluffy judgment subjective methodologies, was then included with hereditary calculations to break down occurrences of the disease [18]. The introduction of the CANFIS model prompted an assessment of direct execution and characterization exactness, and the outcomes indicated that the anticipated CANFIS model had a noteworthy potential for foreseeing coronary illness.

McAlister et al. [2019] states that potential to score for coronary disease in Andhra Pradesh. The investigation on 14 informational indexes from UCI information was performed utilizing ten cross-crease approvals. They produced class affiliation rules utilizing an element submittal choice measure, for example, SU, Ig, and a hereditary inquiry to decide which qualities contribute more toward coronary illness. Analytic tests must be taken by a patient. They utilized a cooperative approach to improve the exactness of the arrangement [19].

Habib et al. [2019] proposed an improved multilayer perceptron calculation that partitioned the dataset into a few subsets. At that point, the MLP calculation was performed exclusively for every subset and the outcomes were acquired from various subgroups using the greater likelihood rule of the democratic combiner. Lastly, the affectability, explicitness, exactness, and execution of these methods were estimated through ROC [20].

Prakash et al. [2019] created a best-fit arrangement calculation that had a noteworthy exactness over-characterization for coronary illness predictions [21]. The information was processed using an immense component choice strategy to find the datasets. The reduced information was then given for characterization. In the examination, the crossbreed property determination technique, joining together CFS and a channel subset assessment, gave more exact results. They proposed a component determination strategy calculation which was a half and half technique consolidating the CFS and Bayes hypothesis.

Mythili et al [2013] proposed a rule-based replica to evaluate the correctness of applying regulations to the entity results of logistic regression on a rule vector mechanism, a decision tree, as well as the Cleveland heart ailment record to predict heart problems. A rule-based loom is the generally used method that combines the results of several models. Rule-based models such as C4.5 have also been implemented, but never in conjunction with any predictive models. Thus, this paper presents a unique replica meant for the comparative study of rule-based algorithms for combining SVMs, decision trees, and logistic regressions [22].

The information set used by Vembandasamy et al. [2015] was a medical data set composed by the foremost diabetes research institutes in Chennai and had records of about 500 patients. The medical dataset requirement provided brief, unmistakable definitions meant for diabetes-related items. The naïve Bayes algorithm was the fact removal method to identify patients with heart disease. This paper analyzed attributes in addition to predicting heart diseases, thereby revealing a cardiovascular disease prediction scheme and HDPS based on the data mining approach [23].

Dalvi et al. [2016] forecast heart ailments in the medical field by means of data knowledge. There has been a lot of research completed linked to this difficulty, but the accurateness of the predictions still needed to be enhanced. As a consequence, this analysis centered on facility variety methods and algorithms where numerous heart ailment datasets were used to show rapid minor progress inaccuracy as an instrument; decision tree, logistic regression, logistic regression, SVM, naive Bayes, and random forest; algorithms are used as methods of attribute preference in addition to results are improved by accurate output [24].

Gandhi et al. [2015] developed a framework for classifying medical data where orthogonal limited safety projection (OLPP) was used to diminish attribute measurement. Once an attribute was lacking, it resolved a prediction based on the classification. The database consisted of 76 features; though, every dispersed test was consigned to using only 14 of them. The accurateness standards of the three datasets were 92.08% for the Cleveland dataset, 92.85% for the Hungarian dataset, and 94.93% for the Switzerland dataset. The sensitivity values for the three datasets were 0.902% for the Cleveland dataset, 0.994% for the Hungarian dataset, and 0.5% for the Switzerland dataset. Typical values for the three datasets were 0.920% for the Cleveland dataset, 0.811% for the Hungarian dataset, as well as 0.99% for the Switzerland dataset [25].

Repaka et al.[2019] used data on humanity as seen by the Registrar General of India that heart disease is a significant cause of death in India, and about 33 percent of deaths in rural areas in Andhra Pradesh are attributed to coronary heart disease. Therefore there is a need to expand the classifications used for predicting heart ailments. In this paper, we recommend resourceful associative categorization algorithms by means of hereditary approaches for the prediction of heart diseases. The most important motivation for using hereditary algorithms in the search for a high-level forecast system was that the revealed system was extremely predictive, by means of high predictive accurateness and elevated concepts of importance. Results suggest that classifier regimens primarily aid the prognosis of a heart ailment, which assists hospitals in their decisions [26].

Tuli et al. [2020] used classification models to identify which selective features played an important role in the prediction of heart ailments using the Cleveland and Statlog Project heart ailment datasets. The accuracy of the random forest algorithm in both classification and feature selection models was observed to be 90–95% based on three different percentage splits. The eight and six selected features incur minimal feature requirements to create an improved performance model. Therefore, omitting eight or six selected features may not lead to better performance for the prediction model [27].

Yilmaz et al. [2013] used a least-squares support vector machine (LS-SVM) with a binary judgment hierarchy for categorizing cardiotoco graphs to determine fetal position [28]. The attributes of a toll-SVM were optimized with a particle swarm optimization. The sturdiness of the process was checked by consecutively creased cross-validation. The presentation of the process was evaluated with the stipulations of an all-purpose classification accurateness. Receivers were also presented to analyze and envisage the presentation of the procedure trait analysis with the cobweb symbol process. Tentative results revealed that the planned process achieved an outstanding categorization accurateness pace of 99.62%.

Fact removal has played a significant role in intelligent medical systems. Relationships of real causes as well as special effects of diseases as seen in patients and can effectively test consumers by built-in apps. Huge databases can be implemented using the facts entered by the extendibility of the software. In addition to secret information connections, the belongings of those associations are preserved between the broad medical dataset so that not to test are satisfactorily investigated. This study was discovered by searching for frequent items using a candidate generation. Medical databases may contain sets of simultaneously observed diseases to be minimized through the non-candidate approach. Information on risk factors can be linked by healthcare professionals to recognize high-risk heart ailment patients. Arithmetical examinations in addition to fact removal techniques assists health care professionals in the judgment of heart ailments.

9.3 PROPOSED METHODOLOGY

9.3.1 Dataset Structure and Description

The current research aims to improve the analysis of heart ailments by examining patient symptoms through fact removal categorization methods; to achieve this goal, a literature review to review data mining operations connected to the analysis of heart ailments was completed. After that, six classifiers (e.g., naïve Bayes, decision tree, support vector machine, k-nearest neighbor, logistic regression, and random forest) were selected to create the model with the maximum accuracy possible. We also explored precision scores, recall scores, F-scores, and false-negatives using a confusion matrix for every algorithm utilized [29]. Each technique had a different way of creating a classifier, which ensured that these techniques behaved differently and produced different results. To underline the practical feasibility of our approach, the selected classifier was implemented using machine learning, in all conducted experiments, to assess the performance of the algorithm where the cross-validation model was omitted, and where the original dataset differed—different training and testing data were divided into sets. In each validation cycle, an observation was conducted where the remaining observations served as training sets, and they were used to create the classifier model. This stride was repeated until all observations served as test observations. Accordingly, performance errors were calculated at each iteration, and the average performance was calculated at the end of the validation.

Two datasets were used in this study:

- The first was obtained from the Cleveland Clinic Foundation and contained 303 records, 297 complete and six with missing/unknown values. In this experiment, six missing values in the Cleveland dataset were cleaned and transformed by data preprocessing; only 297 instances were taken for this study. The dataset itself (297 instances) was divided as 80 % training data and 20% testing data [30].
- The second was the Statlog dataset that included 270 complete records. Originally, both of these datasets had 76 attributes, but they were preprocessed to produce 14 features in an attempt to reduce the number of variables; as a result, we compared these specific characteristics with other literature used. The dataset itself (270 instances) was divided as 80% training data and 20% testing data. The attribute "num" was the heart disease diagnosis attribute. It was classified as presence and absence. If it was present, the value of num would be low or medium, or high or very high. If it was absent, the value of num would be zero.

Figure 9.1 shows the flow chart of heart disease prediction. Our aim was a comparative study of two datasets of heart disease, which were the datasets of Cleveland and Statlog heart disease, to predict the presence or non-appearance of the disease or which, on the basis of fewer characteristics, affected the patient. The heart ailment datasets were obtained from the UCI machine learning library, which contained 14 attributes for prediction on the basis of a smaller number of attributes and with a faster performance for correct predictions.

Algorithm for Performance Evaluation

1) Load data and import libraries to calculate the number of records and missing values from the large infinite data.
2) Convert the large data into categorical valued data for feature selection and fix the missing values by applying data preprocessing.
3) Use the correlation matrix for depicting the relation between feature attribute targets and then find the strong and negative correlation between them.
4) Apply the visualization technique for data analysis of the feature selection of the relevant data using a standard scalar to fit and transform the data into 0 and 1 form, which is easy for predicting disease.
5) Split the dataset into an 80% training and 20% testing dataset to fit the parameters and assess the performance of the model, respectively.
6) Apply the machine learning algorithm on both datasets and compare the performance using the performance matrix for depicting which is a better algorithm.
7) After modeling and predicting with the machine learning classifier, check the accuracy score, precision, recall, and F-measure of both datasets.
8) Evaluate the results of both datasets and calculate the performance measure by plotting a graph of a different classifier.

FIGURE 9.1   Flow chart of heart disease prediction.

9) In our result, a random forest gave a higher accuracy, which was 99.80% in the Cleveland dataset and 87.03% in the Statlog dataset over logistic regression, which was second-lowest.

9.3.2 DATA COLLECTION

A heart disease dataset is suitable to extract the features necessary to create an intelligent decision support system. Generally, medical data is not easily released by many health centers due to their confidentiality agreements. Therefore, this research work used standard heart ailment datasets from the UCI machine learning repository, including the Cleveland and Statlog heart ailment datasets from the UCI dataset [31]. The heart ailment dataset can be easily downloaded from the machine

available at the UCI repository. We used 14 of them to obtain accurate results with a small number of attributes. The "goal" field refers to the occurrence of a heart ailment in the patient (target). It is numerically valued from 0 (no occurrence) to 4 (occurrence). The main advantage of using this dataset was that it could also be used by other manufacturers or health experts. Advice from other manufacturers may help improve the results of the proposed framework [32]. Table 9.1 describes the characteristics and their possible data types or values-selected in the heart disease dataset.

In this study, all the supposed attributes were transformed into numerical attributes.

9.3.3 Data Preprocessing

The collected data are usually weakly organized and have missing values and out-of-range values. Datasets that do not encounter these types of problems may produce incorrect results. The selected data were examined for sound, discrepancies, and omitted values. In addition to extracting data from the dataset for training data, hard missing values were replaced by means of allocation occurrence, most close to those calculated by the nearest neighbor algorithm. However, these values must be treated

TABLE 9.1
Attribute Description

Number	Feature Name	Feature Information
1	Age	Age of patients in existence
2	Sex	1 = male, 0 = female
3	Cp	Value 1: Typical Angina, Value 2: Atypical Angina, Value 3: Non-Angina Pain, Value 4:Asymptotamatic
4	Trestbps	Resting Blood Pressure in mm Hg in admittance to the sanatorium
5	Chol	Serum Cholesterol of patients calculated in mg/dl
6	Fbs	Fasting blood sugar of the patient. If >120 mg/dl Value 1 = accurate, Value 0 = fake
7	Restecg	Value 0 = Usual, Value 1 = having ST-T Wave irregularity, Value 2 = Showing likely or exact Left Ventricular Hypertrophy by Estes' criterion
8	Thalach	Maximum Heart pace Achieved of Patient
9	Exang	Value 1= positively, Value 0 = negatively
10	Oldpeak	ST Depression induced by exercise relative to rest
11	Slope	Value 1 = Up Sloping, Value 2 = Flat, Value 3 = Down Sloping
12	Ca	quantity of foremost vessel{0-3}
13	Thal	Value 1 = Usual, Value 2 = Permanent Defected, Value 3 = Reversible Defect
14	Target	Numerical value 0 to 4

prior to use, as they may predict failure classification or inaccurate disease allocation [33]. For the preprocessing of data, there are several steps, such as cleaning, normalization, and transformation. The result of data preprocessing is an important dataset with minimal characteristics.

The presence of redundant records is the main drawback of healthcare providers. The existence of redundant instances in the dataset can lead to inappropriate learning, often biased toward certain data, which could raise ethical issues. To achieve better accuracy, redundant records must be removed from the data. Thousands of records in healthcare data are added to a relational database table, and records that fail to match the layout can be removed from the record set. By average value, if any instance or attribute falls outside this range then those instances are removed from the dataset [34]. Irrelevant and missing data are also removed from the record set. During preprocessing, the entire dataset is taken as one input, and various data preprocessing approaches are applied to reduce invalid instances from the dataset [22].

The preprocessing process is done as follows:

- Thousands of records from the healthcare data [35] are merged into a relational database table and records that fail to match the layout are removed from the record set.
- Records [36] that go beyond the range of the mean value are also removed.
- The record with the missing value is replaced with the average of the entire column.
- Removing instances [23] that have more than one value, that is, removing the multi-level attributes (Figure 9.2).
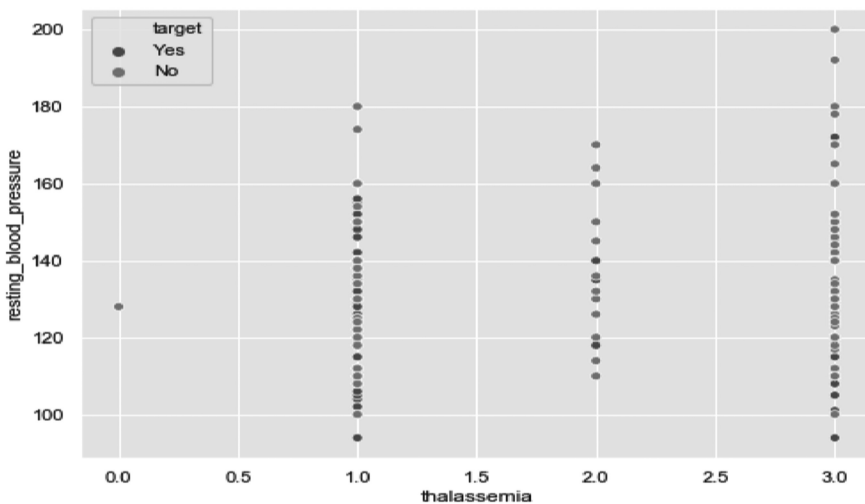

FIGURE 9.2   Thalessemia V/s resting blood pressure.

9.3.4 Correlation Matrix

Correlation study is a process of numerical valuation used to learn the potency of an association among two, numerically calculated, continuous variables [37, 38]. The correlation matrix can be seen in Figure 9.3 and Figure 9.4.

Correlation scrutiny is a means of arithmetical assessment used to learning the potency of an association among two, numerically deliberate, incessant variables [24, 39]. The different values of Thalessemia V/s resting blood pressure can be seen in Figure 9.4, and correlation can be seen in Figure 9.5.

Matrix Correlation Analysis: There was no particular attribute that had an incredibly high association with our objective. Also several descriptions had a pessimistic association with the objective, and various were optimistic [40, 41].

9.4 EVALUATION AND RESULT

9.4.1 Performance Measure

Five general performance measurements were used to evaluate the accuracy of the classification algorithms [25]. These measures were selected because they were widely used to assess the performance of categorization models. The first measure was memorized, as shown in Equation (9.1), which measures how good a binary
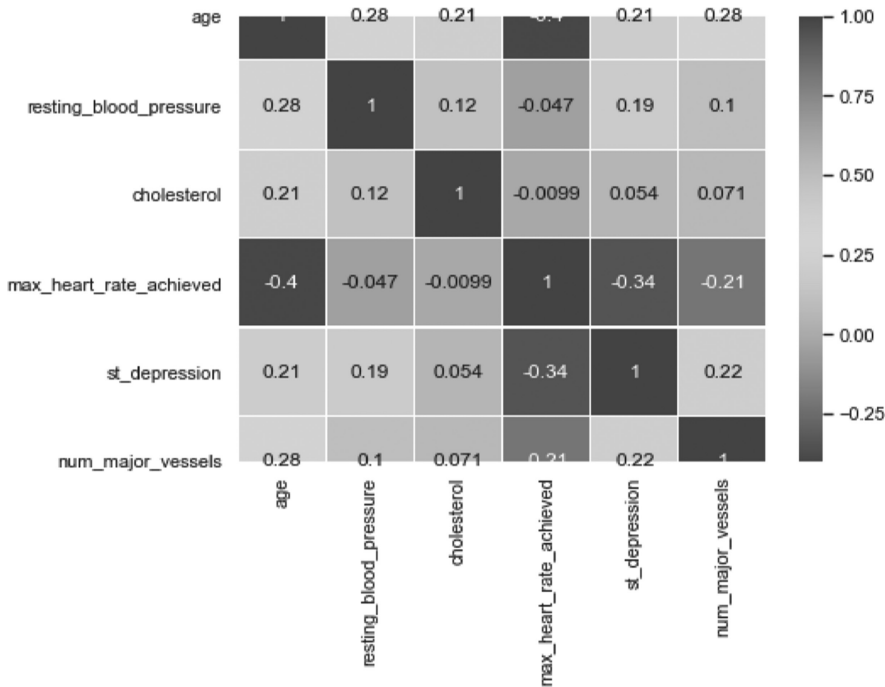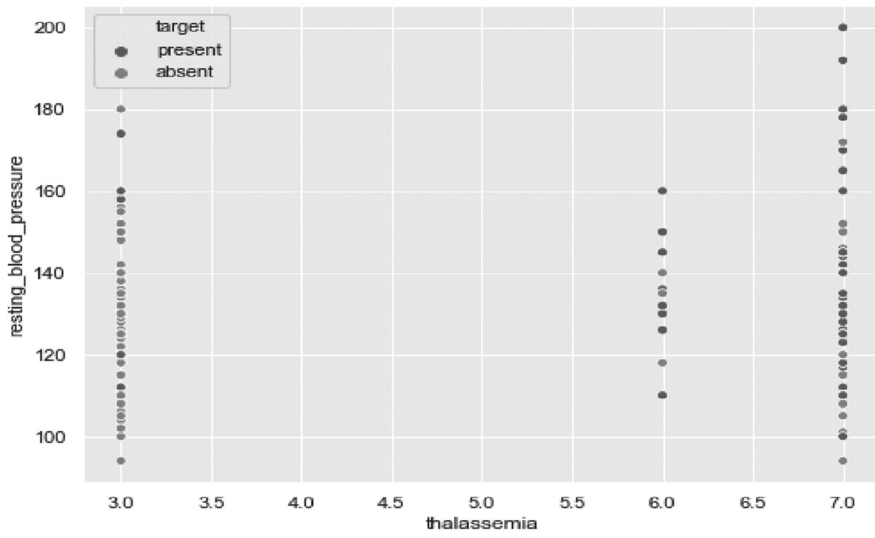

FIGURE 9.3 Correlation matrix.

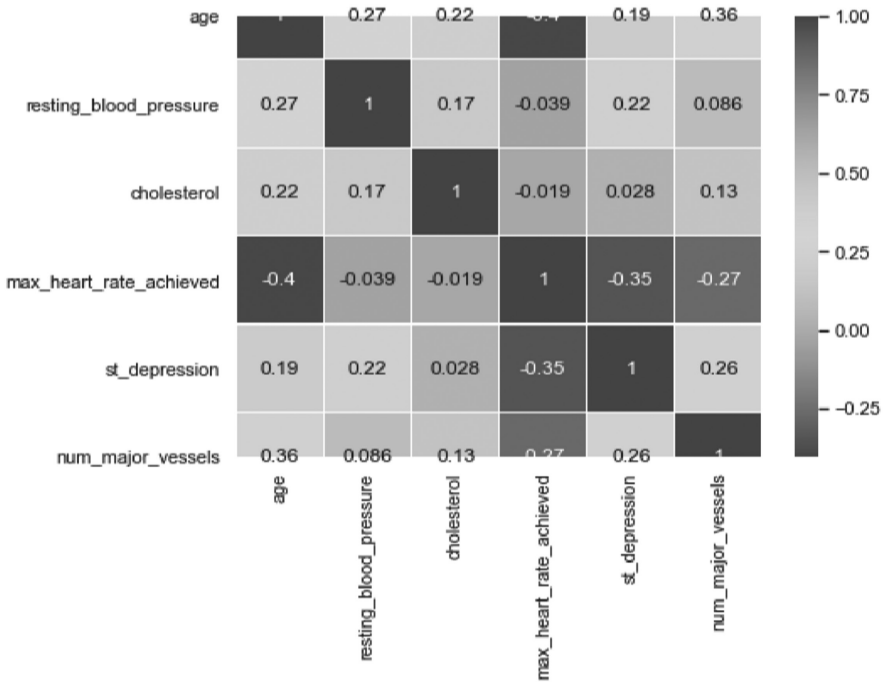**FIGURE 9.4** Thalessemia V/s resting blood pressure.

FIGURE 9.5 Correlation.

categorization test identifies the likelihood of correcting the classification members of the objective class [26]. The second measure is shown in Equation (9.2), which processes the likelihood that an optimistic prediction is accurate. The third assesses accuracy, as revealed in Equation (9.3), which measures the performance of classification. The fourth measure is uniqueness, as shown in Equation (9.4), which processes how well the binary categorization test identifies pessimistic cases [27, 42]. The last measure is the F-measure, as shown in Equation (9.5), which processes the likelihood that an optimistic prediction is accurate.

$$\text{Precision} = \text{TP} \qquad\qquad (9.1)$$
$$\text{TP} + \text{FP}$$

$$\text{Recall} = \text{TP} \qquad\qquad (9.2)$$
$$\text{TP} + \text{FN}$$

$$\text{F1} = 2*\text{Precision}*\text{Recall} \qquad\qquad (9.3)$$
$$\text{Precision} + \text{Recall}$$

$$\text{Accuracy} = \text{TP} + \text{TN} \qquad\qquad (9.4)$$
$$\text{TP} + \text{FN} + \text{TN} + \text{FP}$$

$$\text{Specificity} = \text{TN} \qquad\qquad (9.5)$$
$$\text{TN} + \text{FP}$$

Where,
   TP is figure of factual optimistic.
   TN is figure of factual pessimistic.
   FP is the figure of fake optimistic.
   FN is figure of fake pessimistic.

9.4.2 EXPERIMENTAL RESULTS

In traditional healthcare data systems, doctors rely on patients' symptoms to diagnose heart disease. However, many signs and symptoms are not significant, which can lead to unwanted errors and thus affect the quality of health service provided. To overcome these issues, expert systems were developed to simulate health decision-making with better accuracy [28, 43]. This involves correlating different pieces of patient information after obtaining specific patterns. In this research, we implemented six classification algorithms for the prediction of heart ailments based on pre-selected characteristics.

For each dataset, we ran six classification algorithms, which omitted one cross-validation. Table 9.2 shows that all the categorization algorithms were predictive in addition to giving accurate responses regarding the patient's condition. In particular,

TABLE 9.2

Performance Measure Using Statlog Dataset

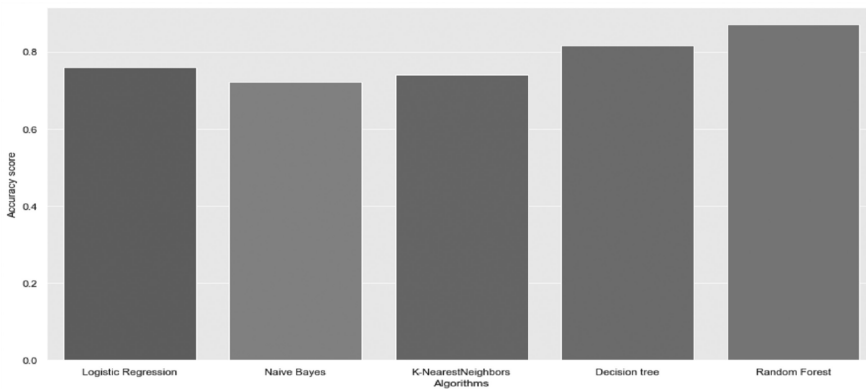| Performance Measure	Decision Tree	Naïve Bayes	*k*-NN	SVM	Logistic Regression	Random Forest
Accuracy	0.851851	0.740740	0.6851851	0.759259	0.777777	0.8703703
Support	54	54	54	54	54	54
Precision	0.85	0.74	0.69	0.76	0.78	0.87
Recall	0.84	0.74	0.68	0.75	0.78	0.86
F-Measure	0.85	0.74	0.69	0.76	0.78	0.87

FIGURE 9.6   Statlog dataset accuracy score graph.

the recall measurements suggested that all the classification models [44] could work well and the Cleveland dataset, but unfortunately it wasn't established which classification algorithm produced more accurate results. Similarly, the uniqueness measure could not tell us which classification algorithms gave better results. However, the remaining performance measurements (F-measure, accuracy, specificity) had the same trend that confirmed that the decision tree was the most correct classifier among all the classification algorithms.

The Statlog Dataset Accuracy Score Graph is shown in Figure 9.6. Similarly we ran six algorithms on the Cleveland dataset. The results obtained for this dataset were quite similar to those obtained for the Statlog dataset. The random forest was most accurate algorithm confirmed by the results of accuracy, precision, and F-measure. Both recall and support measure cannot tell us any useful feedback because they cannot distinguish among algorithm [45, 46].

Another important issue that should be emphasized in this study was the ranking stability of the classification algorithms over both datasets as well as the crossways manifold performance measure. Table 9.3 shows the performance measure using the Cleveland dataset. Position constancy to facilitate this algorithm should create the correct results and be used with all datasets and crossways assessment measures.

TABLE 9.3
Performance Measure Using Cleavaland Dataset

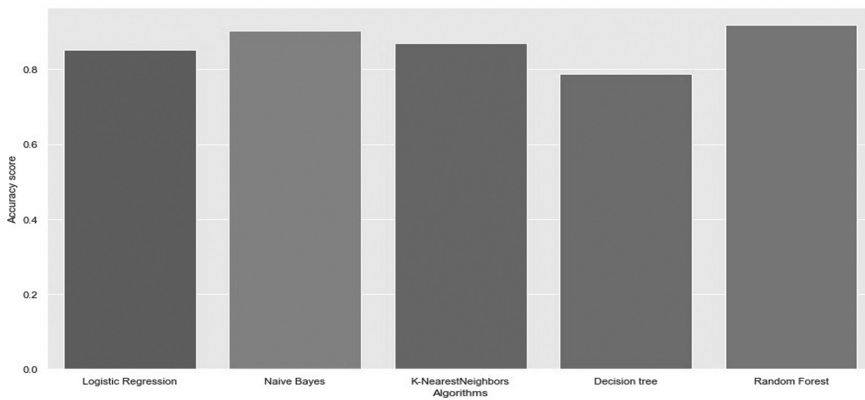| Performance Measure	Decision Tree	Naïve Bayes	k-NN	SVM	Logistic Regression	Random Forest
Accuracy	0.786885	0.901639	0.868852	0.859259	0.852459	0.918032
Support	61	61	61	61	61	61
Precision	0.79	0.90	0.87	0.85	0.85	0.88
Recall	0.79	0.90	0.87	0.85	0.85	0.87
F-Measure	0.79	0.79	0.87	0.85	0.85	0.87

FIGURE 9.7 Cleveland dataset accuracy score graph.

Figure 9.7 shows the Cleveland and dataset accuracy score graph. From our result, we could observe that the random forest was ranked first across all evaluation measures in both datasets, followed by logistic regression in the second position. This gave us an important conclusion on the role of the random forest in classifying data.

9.5 CONCLUSION AND FUTURE SCOPE

9.5.1 CONCLUSION

Motivated by the increasing rates of many different heart ailments, scientists are using fact removal strategies for the scrutiny of heart diseases. The strategy of fact removal for healthcare specialists in the scrutiny of heart ailments has been a definite achievement, the use of a data mining system recognizes the appropriate treatments for heart disease patients. The major reason of our report was to compare accurateness and analyze the reasons following the disparity of dissimilar algorithms. Accurate predictive safety measures provide correct healthcare; similarly, using comparative studies of heart disease on various datasets has demonstrated

favorable results in detecting heart disease. We used the Cleveland and Statlog datasets for heart diseases which consisted of 303 and 270 instances, respectively, and split the facts into two sections, which are the training and testing datasets. We used 14 attributes and implemented six dissimilar algorithms to analyze accurateness. For this, in the present work, we successfully achieved a high accuracy using the random forest classifier for heart disease analysis.

9.5.2 Future Scope

In the future, automated heart disease prediction systems may be implemented in remote areas such as rural areas to replicate human specialists for diagnosis. The prediction system is appropriate to support healthcare experts in the judgment of heart ailments. The characteristics of the disease should be enhanced to achieve more accurate results. With the use of these characteristics, work can be extended to detect other types of disease, such as cancer and arthritis, but early detection of any chronic disease should also be noted. The automation of heart ailment predictions can be improved by building a GUI, website, or mobile application. Genuine information from hospitals, healthcare organizations, and healthcare agencies could be used to increase the accuracy of data technologies along with other technologies for predicting heart ailments.

REFERENCES

1. N. S. C. Reddy, S. Shue Nee, L. Z. Min, and C. X. Ying. 2019. Classification and feature selection approaches by machine learning techniques: Heart disease prediction, *International Journal of Innovative Computing* Vol. 9, No. 1, pp. 39–46, doi:10.11113/ijic.v9n9.210.
2. A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour. 2015. Effective diagnosis and monitoring of heart disease, *International Journal of Software Engineering and its Applications*, Vol. 9, No. 1, pp. 143–156, doi:10.14257/ijseia.2015.9.9.12.
3. S. U. Amin, K. Agarwal, and R. Beg. 2013. Genetic neural network based data mining in prediction of heart disease using risk factors, *IEEE Conference on Information and Communication Technologies 2013*, pp. 1227–1231, doi:10.1109/CICT.2013.6558288.
4. J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle. 2007. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study, *British Medical Journal*, Vol. 335, No. 7611, pp. 136–141, doi:10.1136/bmj.39269.471806.55.
5. A. Patel, S. Gandhi, S. Shetty, and B. Tekwani. 2017. Heart disease prediction using data mining, *International Research Journal of Engineering and Technology (IRJET)*, Vol. 4, No. 1, pp. 1705–1707.
6. H. E. Kuper, M. Marmot, and H. Hemingway. 2002. Systematic review of prospective cohort studies of psychosocial factors in the etiology and prognosis of coronary heart disease, *Seminars in Vascular Medicine*, Vol. 2, No. 3, pp. 267–314, doi:10.1055/s-2002-35401.
7. T. Marikani, and K. Shyamala. 2017. Prediction of heart disease using supervised learning algorithms, *International Journal of Computer Applications*, Vol. 165, No. 5, pp. 41–44, doi:10.5120/ijca2017913868.

8. G. Shanmugasundaram, V. Malar Selvam, R. Saravanan, and S. Balaji. 2018. An investigation of heart disease prediction techniques, *IEEE International Conference on System, Computation, Automation, and Networking, ICSCA 2018*, pp. 1–6, doi:10.1109/ICSCAN.2018.8541165.

9. A. Golande, and T. P. Kumar. 2019. Heart disease prediction using effective machine learning techniques, *International Journal of Recent Technology and Engineering (IJRTE)*, Vol. 8, No. 1, Special Issue 4, pp. 944–950.

10. M. A. Jabbar, B. L. Deekshatulu, and P. Chandra. 2012. Prediction of risk score for heart disease using associative classification and hybrid feature subset selection, *International Conference on Intelligent Systems Design and Applications ISDA*, pp. 628–634, doi:10.1109/ISDA.2012.6416610.

11. N. Kumar, and S. Khatri. 2017. Implementing WEKA for medical data classification and early disease prediction. *3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, pp. 1–6, doi:10.1109/CIACT.2017.7977277.

12. S. D. Desai, S. Giraddi, P. Narayankar, N. R. Pudakalakatti, and S. Sulegaon. 2019. Back-propagation neural network versus logistic regression in heart disease classification. *In Advances in Intelligent Systems and Computing*, Vol. 702, pp. 133–144, doi:10.1007/978-981-13-0680-8_13.

13. K. S. Purushottam, and R. Sharma. 2015. Efficient heart disease prediction system using decision tree, *International Conference on Computing, Communication & Automation ICCCA 2015*, pp. 72–77, doi:10.1109/CCAA.2015.7148346.

14. K. Dsouza, and Z. Ansari. 2018. Big data science in building medical data classifier using naïve bayes model, *IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pp. 76–80. doi:10.1109/CCEM.2018.00020.

15. A. Daftaribesheli, M. Ataei, and F. Sereshki. 2011. Assessment of rock slope stability using the Fuzzy Slope Mass Rating (FSMR) system, *Applied Soft Computing*, Vol. 11, No. 8, pp. 4465–4473, https://doi.org/10.1016/j.asoc.2011.08.032.

16. C. S. Dangare, and S. S. Apte. 2012. Improved study of heart disease prediction system using data mining classification techniques, *International Journal of Computer Applications*, Vol. 47, No. 10, pp. 44–48, doi:10.5120/72280076.

17. A. Cooper, G. Lloyd, J. Weinman, and G. Jackson. 1999. Why patients do not attend cardiac rehabilitation: Role of intentions and illness beliefs, *Heart*, Vol. 82, No. 2, pp. 234–236, doi:10.1136/hrt.82.2.234.

18. L. Parthiban, and R. Subramanian. 2007. Intelligent heart disease prediction system using CANFIS and genetic algorithm, *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, Vol. 1, pp. 278–281.

19. F. A. McAlister et al. 2019. A comparison of four risk models for the prediction of cardiovascular complications in patients with a history of atrial fibrillation undergoing non-cardiac surgery, *Anaesthesia*, Vol. 2019, pp. 27–36, doi:10.1111/anae.14777.

20. S. Habib, M. B. Moin, S. Aziz, K. Banik, and H. Arif. 2019. Heart failure risk prediction and medicine recommendation using exploratory data analysis, *1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1–6.

21. S. Prakash, K. Sangeetha, and N. Ramkumar. 2019. An optimal criterion feature selection method for prediction and effective analysis of heart disease, *Cluster Computer*, Vol. 22, pp. 11957–11963, https://doi.org/10.1007/s10586-017-1530-z.

22. T. Mythili, D. Mukherji, N. Padalia, and A. Naidu. 2013. A heart disease prediction model using SVM-decision trees-logistic regression (SDL), *International Journal of Computer Applications*, Vol. 68, No. 16, pp. 11–15, doi:10.5120/11662-7250.

23. K. Vembandasamy, R. Sasipriya, and E. Deepa. 2015. Heart diseases detection using naive bayes algorithm, *International Journal of Innovative Science Engineering & Technology*, Vol. 2, No. 9, pp. 441–444.

24. P. K. Dalvi, S. K. Khandge, A. Deomore, A. Bankar, and V. A. Kanade. 2016. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression, *Symposium on Colossal Data Analysis and Networking (CDAN)*, Indore, pp. 1–4, doi:10.1109/CDAN.2016.7570883.

25. M. Gandhi, and S. N. Singh. 2015. Predictions in heart disease using techniques of data mining, *1st International Conference on Futuristic Trends on Computational Analysis and Knowledge Management ABLAZE 2015*, pp. 520–525, doi:10.1109/ABLAZE.2015.7154917.

26. A. N. Repaka, S. D. Ravikanti, and R. G. Franklin. 2019. Design and implementing heart disease prediction using naives bayesian, *International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 292–297, doi:10.1109/icoei.2019.8862604.

27. S. Tuli et al. 2020. HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments, *Future Generation Computer Systems*, Vol. 104, pp. 187–200, doi:10.1016/j.future.2019.10.043.

28. E. Yilmaz, and Ç. Kilikçier. 2013. Determination of fetal state from cardiotocogram using LS-SVM with particle swarm optimization and binary decision tree, *Computational and Mathematical Methods in Medicine*, Vol. 2013, pp. 1–8, doi:10.1155/2013/487179.

29. B. M. Al-maqaleh. 2017. Intelligent predictive system using classification techniques for heart disease diagnosis, *International Journal on Computer Science & Engineering (IJCSE)*, Vol. 6, No. 6, pp. 145–151.

30. S. Palaniappan, and R. Awang. 2008. Intelligent heart disease prediction system using data mining techniques, *AICCSA 08–6th IEEE/ACS International Conference on Computer Systems and Applications*, pp. 108–115, doi:10.1109/AICCSA.2008.4493524.

31. A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinō. 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms, *Mobile Information Systems*, Vol. 2018, doi:10.1155/2018/3860146.

32. J. S. Sonawane, and D. R. Patil. 2014. Prediction of heart disease using multilayer perceptron neural network, *International Conference on Information Communication and Embedded Systems (ICICES2014)*, pp. 1–6.

33. P. P. Sunila, and N. Godara. 2012. Decision support system for cardiovascular heart disease diagnosis using improved multilayer perceptron, *International Journal of Computers & Applications*, Vol. 45, No. 8, pp. 12–20.

34. M. M. Bhajibhakare, N. Shaikh, and D. Patil. 2019. Heart disease prediction using machine learning, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, Vol. 7, No. XII, pp. 455–460.

35. R. Roy, and K. T. George. 2017. Detecting insurance claims fraud using machine learning techniques, *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2017*, pp. 531–536.

36. J. A. A. G. Damen, L. Hooft, E. Schuit et al. 2016. Prediction models for cardiovascular disease risk in the general population: Systematic review, *BMJ*, Vol. 353, pp. 1–11, doi:10.1136/bmj.i2416.

37. S. C. Virgeniya, and E. Ramaraj. 2019. Predictive analytics using rule based classification and hybrid logistic regression (HLR) algorithm for decision making, *International Journal of Scientific & Technology Research*, Vol. 8, pp. 1509–1513.

38. O. Ye, J. Deng, Z. Yu, T. Liu, and L. Dong, 2020. Abnormal event detection via feature expectation subgraph calibrating classification in video surveillance scenes, *In IEEE Access*, Vol. 8, pp. 97564–97575, doi:10.1109/ACCESS.2020.2997357.

39. B. Tarle, and S. Jena. 2016. Improved artificial neural network for dimension reduction in medical data classification, *International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, pp. 1–6, doi:10.1109/ICCUBEA.2016.7860033.

40. P. S. Kohli, and S. Arora. 2018. Application of machine learning in disease prediction, *4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, pp. 1–4, doi:10.1109/CCAA.2018.8777449.

41. S. S. Bano, and S. Mary. 2019. Heart disease prediction system using genetic algorithm, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, Vol. 7, No. 6, pp. 2178–2182, doi:10.22214/ijraset.2019.6366.

42. C. B. C. Latha, and S. C. Jeeva. 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, *Informatics in Medicine Unlocked*, Vol. 16, doi:10.1016/j.imu.2019.100203.

43. C. Ordonez et al. 2001. Mining constrained association rules to predict heart disease, *Proc. - IEEE International Conference on Data Mining, ICDM*, pp. 433–440, doi:10.1109/icdm.2009.989549.

44. S. K. Khare, and V. Bajaj. 2020. Time-frequency representation and convolutional neural network-based emotion recognition [published online ahead of print, 2020 Jul 31]. *IEEE Transactions on Neural Networks and Learning Systems 2020*, doi:10.1109/TNNLS.2020.3008938.

45. F. Demir, A. Şengür, V. Bajaj et al. 2019. Towards the classification of heart sounds based on convolutional deep neural network. *Health Information Science and Systems*, Vol. 7, p. 16, https://doi.org/10.1007/s13755-019-0078-0.

46. F. Demir, V. Bajaj, M. C. Ince et al. 2019. Surface EMG signals and deep transfer learning-based physical action classification. *Neural Computing and Applications*, Vol. 31, pp. 8455–8462, https://doi.org/10.1007/s00521-019-04553-7.