Review Article

International Journal of Advanced Computer Research, Vol 8(38)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2018.837025

A systematic review and analysis of the heart disease prediction methodology

Animesh Kumar Dubey[1*] and Kavita Choudhary[2]
Research Scholar, Department of Computer Science Engineering, Institute of Engineering and Technology, JK Lakshmipat University, Rajasthan, India[1]
Associate Professor, Department of Computer Science Engineering, Institute of Engineering and Technology, JK Lakshmipat University, Rajasthan, India[2]

Received: 15-May-2018; Revised: 12-September-2018; Accepted: 15-September-2018

Abstract
Most of the decisions in medical diagnosis are taken on the basis of experts' opinions. In the case of heart diseases, however, the experts' decisions do not always reach a consensus since the pattern of heart disorders varies considerably among patients. Researchers have been making continuous efforts to detect heart diseases at the primary stages by using different methodologies in order to increase the chances of curing a condition that has one of the highest mortality rates in the world. The three main objectives of this study were to analyze the global impact of heart diseases on the basis of mortality rates, to assess the risk of heart diseases in different age groups, and to discuss the advantages and disadvantages of methodologies that have been used previously for predicting heart disease at the primary stage. The mortality rate due to heart diseases was assessed according to attributes such as age, population group, clinical risk factors, and geographical locations. Different methodologies were analyzed on the basis of results obtained from literature searches in IEEE, Elsevier, Springer, and other publications. The percentage of deaths due to heart diseases increase with age, indicating that the risk of developing heart disease is directly proportional to age. The analysis of various methodological approaches indicated that data mining and the combination of optimization methods can be effective in predicting heart disease at the initial stages. The current data available on heart diseases can help design better frameworks for predicting new cases. The statistics of heart disease-related death shows a worrying trend world-wide. This study concludes that a framework based on hybrid approaches consisting of the combination of classification and clustering methods of data mining, along with biological system inspired algorithms, can prove to be a landmark in the field of heart disease prediction and detection.

Keywords
Heart disease, Prediction strategies, Death rates, Data mining, Classification and clustering methods.

1.Introduction
Heart or cardiovascular diseases, according to world health organization (WHO) statistics, cause more deaths worldwide, compared to other diseases (31% of all deaths). Current trends project that India will soon rank number one in incidences of heart diseases. In the present scenario, every fifth death in India is due to heart diseases and this ratio is expected to rise up to every third death in 2020, with the majority of the afflicted among the younger age groups [1]. Hospitals all over the world collect a vast amount of data on heart diseases that can be used to predict disease rates manually. However, the data so far have not been translated efficiently to correlate them with the risk and symptoms of the disease [2].

*Author for correspondence

The majority of heart disease- related deaths (75%) occur in low to middle-income countries, most likely due to high costs of diagnosis [3]. According to the National Center for Biotechnology Information (NCBI), the total number of deaths due to heart diseases in the USA increased from 23.2 million in 1990 to 37 million in 2010, i.e., by 59% [4]. Heart disease related mortality rates are lower in the developed countries compared to the developing countries [5].

The region wise statistics of the impact of heart diseases are shown below.

Africa:
In Africa, heart disease is the leading cause of deaths in adults over the age of 30 and the leading cause of overall deaths. One of the most important drivers of heart disease in Africa's is undiagnosed and untreated hypertension, which affects approximately one out of

two Africans over the age of 25, the highest rate in any continent [6].

Australia: Heart disease is a major cause of death in Australia, accounting for 45,392 deaths (30% of all deaths) in 2015 alone. Every 12 minutes, one Australian dies due to heart disease, one out of every six Australians is affected by heart disease, and approximately 400,000 Australians have had a heart attack at some point in their lives. Over 54,000 Australians suffer a heart attack per year, i.e., one heart attack in every 10 min [7].

Europe: Heart diseases account for 3.9 million deaths (45%) in Europe and over 1.8 million (37%) deaths in the European Union annually. Heart disease affects both men and women, with 1.8 million men (40%) and 2.1 million women (49%) dying as a result of it. In 2015, 11.3 million new cases of heart diseases emerged in Europe and 6.1 million new cases were diagnosed in the European Union [8].

South-East Asia: Heart disease causes every fourth death in Asian countries. Approximately 3.7 million die annually, of which 2 million are men. Around 38% of the deaths due to heart disease in this region occur in individuals younger than 70 years. The statistics are especially grim for India which, according to a recent report [9], will soon rank the highest in heart disease-related death rates. Every fifth death in India can be attributed to heart diseases and by 2020, this will increase to every third and majority of the deaths will occur in the younger age groups [1].

United States: Approximately 610,000 people die due to heart disease annually in the United States, accounting for every fourth death in the population. According to an epidemiological report [10], heart diseases are responsible for deaths in both men and women, and 50% deaths in 2009 among men were due to heart diseases. High blood pressure, high cholesterol, and smoking are the main risk factors and about 47% of the Americans present at least one of these factors.

The statistics highlighted above establish heart disease as one of the major health concerns worldwide. This study, therefore, has three main objectives: a) To survey and analyze the incidences of heart diseases in different geographical regions on the basis of death rates, b) To analyze the risk levels and major risk factors of heart diseases in different age groups in the different regions to help predict the risk of developing heart disease in medical diagnoses, and c) To determine the advantages and disadvantages of the previously used algorithms and methods for predicting heart diseases at the early stages. These objectives will help design a better framework for predicting heart diseases at early stages and improve the chances of recovery.

2.Material and methods

Data were collected from different continents to analyze the global impact of heart diseases. The following sources are used for collecting information for our study:

World Health Organization (WHO) is a specialized arm of the United Nations that was established in 1948 in Geneva, Switzerland. The WHO deals with global public health trends.

National Center for Biotechnology Information (NCBI) was established in November 1988 by Senator Claude Pepper in the USA for processing of computerized information for conducting biomedical research.

Centers for Diseases Control and Prevention (CDC) is an operative arm of the Department of Health and Human services and deals with health issues in both USA and other countries.

Institute for Health Metrics and Evaluation (IHME) is a research institute at the University of Washington in Seattle and conducts studies in the areas of global health statistics and impact evaluation.

The Global Terrorism Database (GTD) is an open-source database in the US that includes information on global terrorist events from 1970 through 2016.

The Atlantic is a popular American magazine and multi-platform publisher that was first, established in 1857. It provides daily coverage and analysis of breaking news, politics and international affairs, education, technology, health, science, and culture.

Our World in Data was founded at the University Of Oxford, the UK by Max Roser, with the aim of generating global data overviews and to show long-term trends and changes.

Amnesty International is a London-based non-governmental organization focused on human rights.

The European Heart Network (EHN) is a Brussels-based alliance of heart foundations and non-governmental organizations that are working in the field of prevention and reduction of heart disease.

Neo Cardiab Care is the unit of Fine cure Pharmaceuticals Limited, India, under the aegis of Group Chandrans, and focuses on the areas of human health including cardiac management, diabetes, and Neuropsychiatry.

The Heart Foundation is an Australian organization with the aim of fighting heart diseases.

Data extracted from the above sources were used to calculate the total heart disease-related deaths worldwide in 2016 determine their specific causes and stratify the deaths in the different age groups. First, the death percentages of specific diseases (DPSD) were calculated by the following formula: DPSD = (Deaths due to specific diseases in the current year/Total death in the current year) ×100. This was followed by calculating the DPSDs in specific age groups (DPSDSAG), i.e., 0 to 14 years, 15 to 49 years, 50 to 69 years and 70+ years, in the same year by the following formula:
DSPDSAG = (Deaths due to specific diseases in specific age group/Total deaths in specific age group) ×100.

3.Results

We first focused on the total number of deaths worldwide in 2016, along with their causes, as shown in *Table 1*. Heart disease was responsible for around 17.6 million deaths worldwide in 2016, which was higher than deaths due to any other cause. *Figure 1* shows the percentage of total deaths worldwide attributed to different causes in 2016, and accordingly, heart diseases were responsible for around 34% of total deaths. These statistics illustrate the grim health situation all over the world due to heart diseases. The second focus of our study was to analyze the impact of heart diseases in different age groups in order to find the most susceptible age group, and also to determine the risk of heart diseases with increasing age. *Table 2* shows the total number of deaths worldwide in 2016 in the 0–14 year age group and the overall percentages of deaths in this age group are shown in *Figure 2*. Compared to other diseases, heart conditions had a very low impact on individuals aged between 0 to 14 years. The total number and percentage of deaths in the 15 to 49 year age group are shown in *Table 3* and *Figure 3*, respectively; heart diseases had a higher impact on this age group compared to the other diseases, accounting for more than 1.3 million deaths and 18% of all deaths worldwide in 2016. In the 50–69 age groups, heart diseases were responsible for more than 5.1 million deaths (*Table 4*) and 35% of all deaths (*Figure 4*) in 2016, indicating a significantly higher impact of the heart diseases compared to other disorders. Finally, in the 70+ age group, heart diseases accounted for 11 million deaths worldwide in 2016 (*Table 5*), which made up 44% of this demographic affected by heart diseases (*Figure 5*). As with the other age groups above 14 years of age, heart diseases had a significantly higher impact compared to other diseases among those over 70 years of age. The age-group stratified percentages of deaths worldwide in 2016 were compared (*Figure 6*), clearly showing that the risk of death due to heart disease increased with age.

Table 1 Total number of deaths with their causes in year 2016 over the world [11−13]

Serial no.	Causes of death	Numbers of deaths
1	Heart diseases	17650000
2	Cancers	8930000
3	Respiratory diseases	3540000
4	Diabetes and endocrine diseases	3190000
5	Dementia	2380000
6	Lower respiratory infections	2380000
7	Neonatal death	1730000
8	Diarrheal diseases	1660000
9	Road accidents	1340000
10	Liver diseases	1260000
11	Tuberculosis	1210000
12	Kidney diseases	1190000
13	Digestive diseases	1090000
14	HIV/AIDS	1030000
15	Suicide and Homicide	1406058
16	Maleria, Nutritional deficiencies and Meningitis	1406059
17	Malnutrition, drowning and maternal deaths	841941
18	Parkinsons diseases, alcohol and drug disorder	528964
19	Intestinal infection, hepatitis and fire	421578
20	Heat related, terrorism and natural disasters	97331

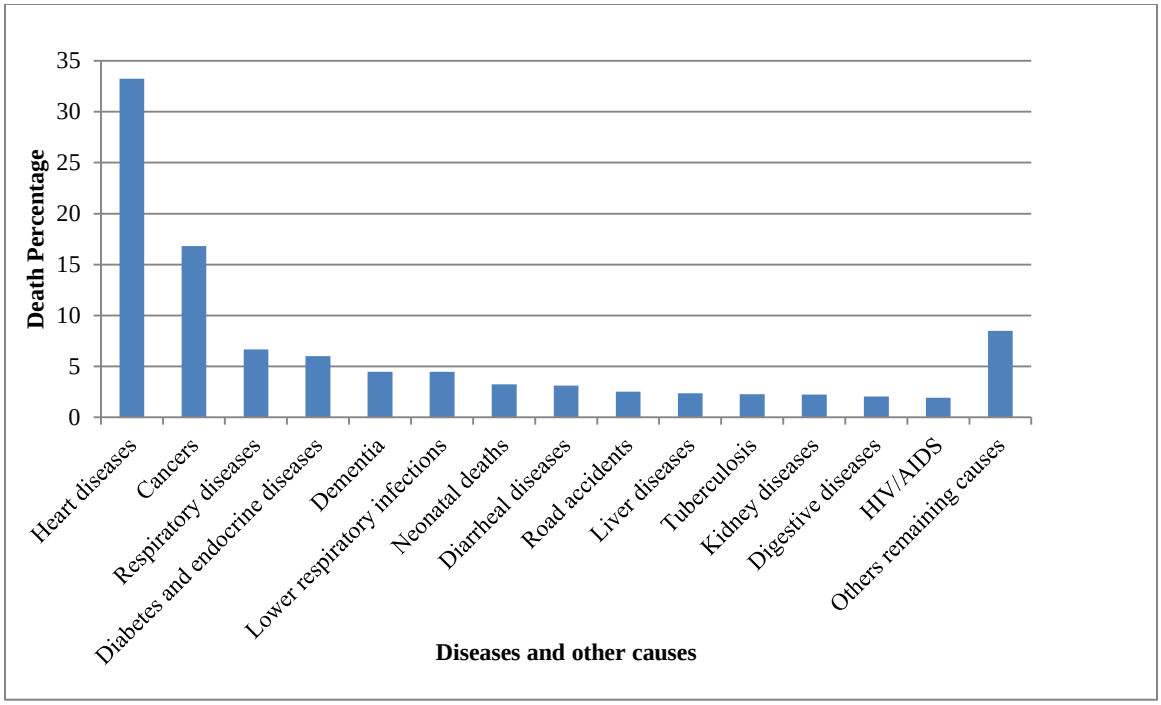

Figure1 Death percentage with respect to their causes in year 2016 over the world

Table 2 Number of deaths worldwide with respect to their causes in the year 2016 (Age between 0 to 14 years) [11, 12]

Serial no.	Causes of death	Numbers of deaths
1	Heart diseases	48915
2	Cancers	79912
3	Respiratory diseases	711528
4	Diabetes and endocrine diseases	76944
5	Congenital birth defects	415223
6	Neonatal death	1972992
7	Diarrheal diseases	497524
8	Road accidents	103811
9	Liver diseases and tuberculosis	52011
10	Kidney diseases and HIV/AIDS	128476
11	Digestive diseases and measles	101211
12	Suicide and homicide	28808
13	Maleria, nutritional deficiencies	769226
14	Drowning and drug disorders	99829
15	Malnutrition	177390
16	Hepatitis, fire and natural disasters	45367

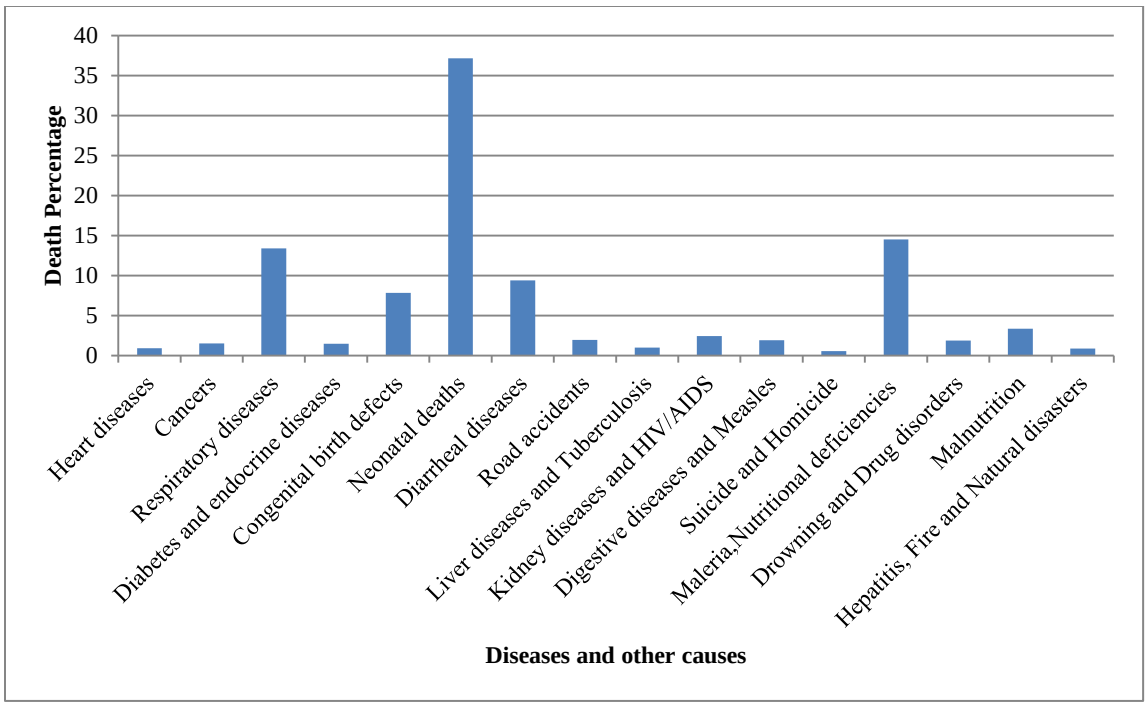

Figure 2 Death percentage with respect to their causes in year 2016 over the world (Age between 0 to 14 years)

Table 3 Number of deaths Worldwide with respect to their causes in the year 2016 (Age between 15 to 49 years) [11, 12]

Serial no.	Causes of death	Numbers of deaths
1	Heart diseases	1340000
2	Cancers	1100000
3	Respiratory diseases	173453
4	Diabetes and endocrine diseases	382958
5	HIV/AIDS and dementia	770583
6	Lower respiratory infections	194297
7	Diarrheal diseases	181117
8	Road accidents	765416
9	Liver and digestive diseases	514957
10	Tuberculosis	419584
11	Kidney and parkinsons diseases	163980
12	Suicide and homicide	781507
13	Maleria and nutritional deficiencies	100067
14	Malnutrition and drowning	136952
15	Alcohol and drug disorder	162406
16	Hepatitis and fire	98820
17	Heat related and natural disasters	20354

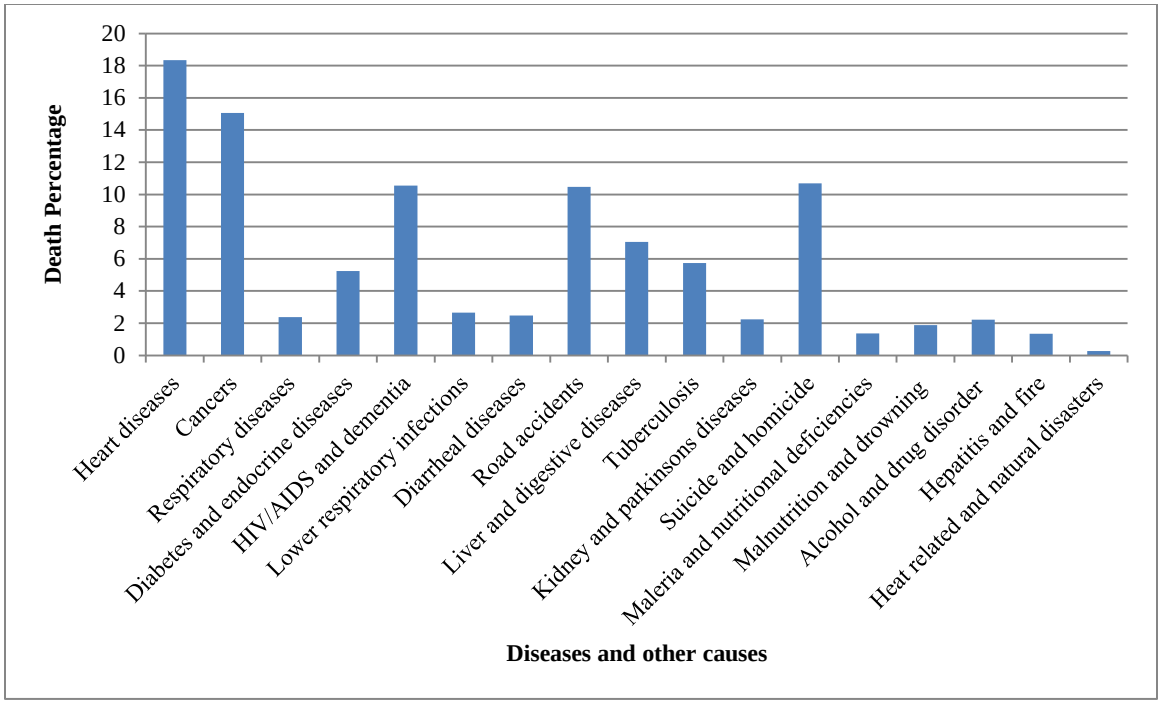

Figure 3 Death percentage with respect to their causes in year 2016 over the world (Age between 15 to 49 years)

Table 4 Number of deaths worldwide with respect to their causes in the year 2016 (Age between 50 to 69 years) [11, 12]

Serial no.	Causes of death	Numbers of deaths
1	Heart diseases	5140000
2	Cancers	3810000
3	Respiratory diseases	1000000
4	Diabetes and endocrine diseases	1030000
5	HIV/AIDS and Dementia	288649
6	Lower respiratory infections	398131
8	Diarrheal diseases	283294
9	Road accidents	324085
10	Liver and Digestive diseases	893576
11	Tuberculosis	424985
12	Kidney and Parkinsons diseases	404549
15	Suicide and Homicide	271729
16	Maleria and Nutritional deficiencies	80918
17	Malnutrition and Drowning	76646
18	Alcohol and Drug disorder	116178
19	Hepatitis and Fire	68449
20	Heat related and Natural disasters	21375

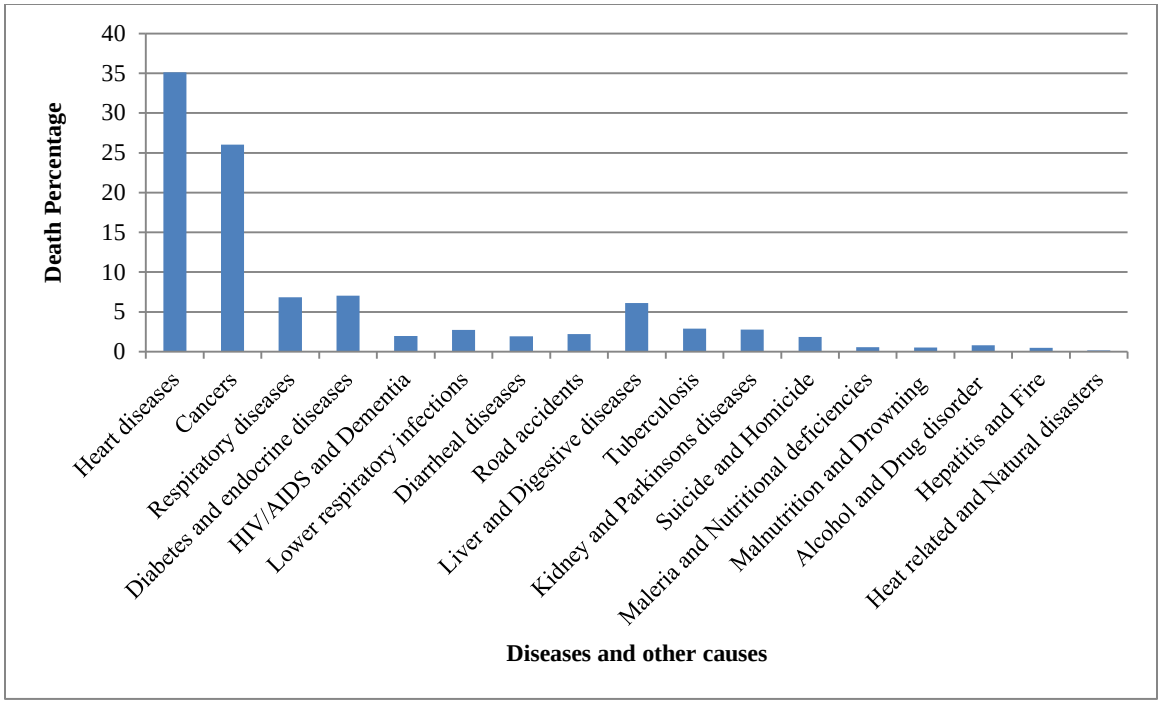

Figure 4 Death Percentage with respect to their causes in year 2016 over the world (Age between 50 to 69 years)

Table 5 Number of deaths Worldwide with respect to their causes in the year 2016 (Age 70 years and above) [11, 12]

Serial no.	Causes of death	Numbers of deaths
1	Heart diseases	11110000
2	Cancers	3930000
3	Respiratory diseases	2350000
4	Diabetes and endocrine diseases	1700000
5	HIV/AIDS and dementia	2242256
6	Lower respiratory infections	1080000
7	Diarrheal diseases	694010
8	Road accidents	148974
9	Liver and digestive diseases	885719
10	Tuberculosis	329177
11	Kidney and parkinsons diseases	813362
12	Suicide and homicide	125903
13	Maleria and nutritional deficiencies	137452
14	Malnutrition and drowning	120543
15	Alcohol and drug disorder	40559
16	Hepatitis and fire	59007
17	Heat related and natural disasters	15421

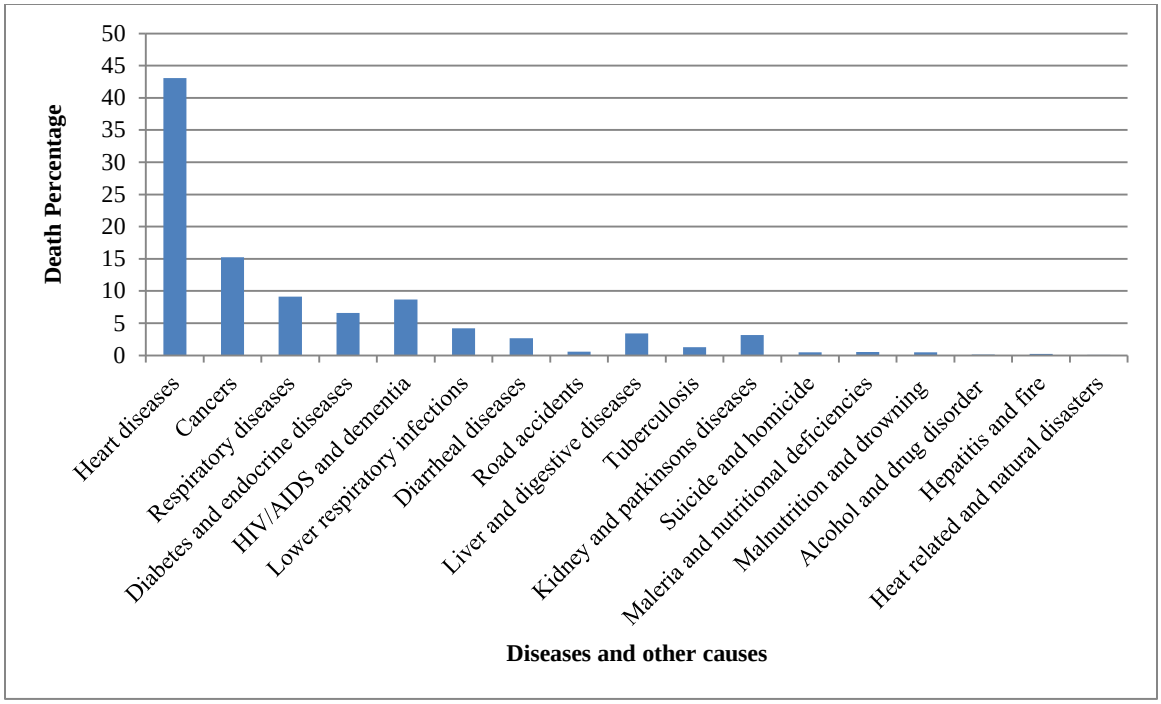

Figure 5 Death percentage with respect to their causes in year 2016 over the world (Age 70 years and above)

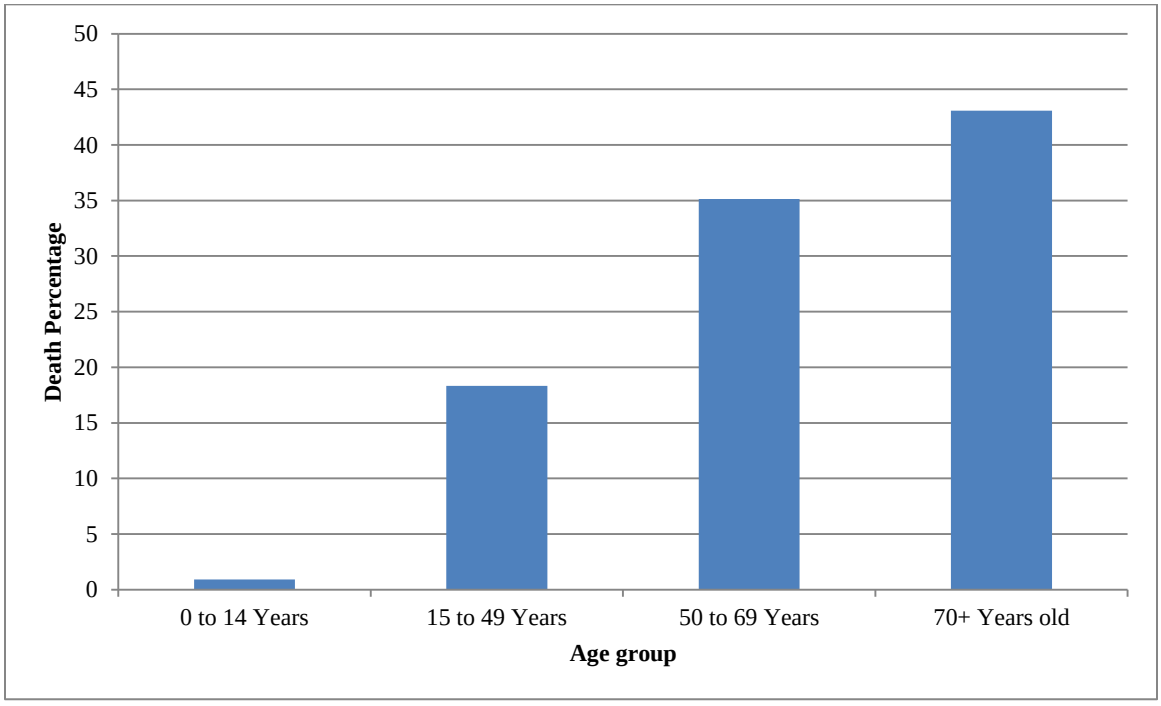

Figure 6 Overall death percentage worldwide with different age group due to heart diseases in year 2016

Next, we discussed prediction methodology based on the study of other methods that are briefly discussed below.

Shouman et al. [14] proposed a framework for diagnosing heart diseases in 2011 using the decision tree (nine voting equal frequency discretization gain ratio decision tree) method of data mining. Datasets of heart diseases were obtained from the Cleveland clinic foundation and 13 out of a total of 76 patient/disease attributes were analyzed: age (numeric), sex: (nominal) , type of chest pain (nominal), resting blood pressure (BP): (numeric), cholesterol (numeric) etc. They achieved a higher accuracy of 84.1% compared to J4.8 decision tree and bagging algorithm. The drawbacks of their methodology were the complexity of the decision tree and the time required for studying the trees, especially the large ones with many branches. Fida et al. [15] proposed a model for classifying heart diseases using a genetic optimization algorithm and a support vector machine (SVM). Four datasets were used: Cleveland dataset with 299 records and 13 attributes, Statlog with 270 records and 13 attributes having binary classes, single proton emission computed tomography (SPECT) (of University of California Irvine (UCI) machine learning repository) with 187 records and 22 attributes, and South African dataset with 462 records and 9 attributes. The data were evaluated with 10-fold cross-validation and the performance of the system was evaluated by classifiers accuracy, sensitivity, and specificity. They were able to achieve 98.63% accuracy with the Cleveland dataset and the accuracies with the Statlog, SPECT, and South African datasets were 80.79%, 93.27%, and 83.40%, respectively. Elbedwehy et al. [16] proposed a method for detecting heart disease using binary particle swarm optimization (BPSO) and SVM, in conjunction with k-nearest neighbor and 'leave-one-out cross-validation'. They applied these methods to a total of 198 heart sound signals obtained from healthy patients, and from patients with heart valve diseases like aortic stenosis, aortic regurgitation, mitral stenosis and mitral regurgitation. They used BPSO for feature selection and SVM for classification of heart signals and achieved an accuracy of 95.12%, 90.24%, and 87.80% in 20, 50 and 100 iterations, respectively, clearly showing a negative correlation between accuracy and number of iterations. Nahar et al. [17] proposed a method for determining the risk factors for heart diseases in males and females with the association rule mining concept, and the Apriori, predictive Apriori, and Tertius algorithms for rule generation. They performed their experiments on the UCI Cleveland heart disease dataset, focusing on 14 out of a total of 76 attributes. They found that a normal or hyper resting electrocardiogram (ECG) and a flat slope were potential high-risk factors for women, whereas only a hyper resting ECG was a significant risk factor for men. This indicates that resting ECG status could be a distinct factor for predicting heart disease in women. Overall, women were found to be less at risk of developing heart disease compared to men. Jabbar et al. [18] proposed a method of classifying heart disease using k-nearest neighbor and genetic algorithm (GA). They used six datasets from UCI and one from a hospital in Andhra Pradesh, India and studied 12 attributes including age, gender, diabetic status, systolic BP, diastolic BP, height, weight, BMI, hypertension, and setting (rural or urban). They achieved a 92.14% accuracy without GA and 95.73% with GA. Austin et al. [19] proposed data mining and machine learning methods for predicting and classifying heart failure-into the preserved ejection fraction and reduced ejection fraction subtypes-and compared bagging, boosting, random forests, and SVM methods with conventional regression and classification trees. Demographic characteristics, vital signs, presenting signs and symptoms, results of laboratory investigations and previous medical history were used as predictors. They achieved a positive predictive value of 69.6% with the random forest method. Alizadehsani et al. [20] proposed a data mining approach for diagnosing heart diseases using the naïve Bayes, sequential minimal optimization (SMO), bagging and neural network methods. They applied these methods on the Z-Alizadeh Sani dataset which contains the records of 303 patients, each with 54 attributes. They arranged those attributes into four groups: demographics, symptoms and examination, ECG, and laboratory and echo. The highest accuracy (94.08%) was achieved by the SMO algorithm along with the attribute selection and creation methods. Taneja [21] developed a cost-effective treatment for heart diseases using data mining technologies. They used the data collected from PGI, Chandigarh containing transthoracic ECG reports of 7,008 patients from 2008 to the first quarter of 2010. They employed the decision tree classification, Bayesian classifier, and neural network techniques of data mining. Although the neural network model achieved an accuracy of 94.85%, the most effective model to predict patients with heart disease was the J48 classifier, which showed 95.56% accuracy when implemented on selected attributes. Bohacik et al. [22] used alternating decision tree method to predict heart

diseases. In standard decision trees, only leaf nodes can be split, but in alternating decision trees, each part can be split multiple times. They used the data of Hull Life Lab, a large, epidemiologically representative dataset consisting of 2032 patients, for their study. The plasma levels of creatinine, uric acid, N-terminal pro b-type natriuretic peptide (NT-proBNP), and sodium, along with pulse rate, weight, height, gender, and age were used as predictive factors, and resulted in an accuracy of 77.66%. Persi et al.[23] proposed a fuzzy optimization technique for heart diseases prediction using the decision tree concept of data mining. Particle swarm optimization (PSO) was used for the optimization of fuzzy membership functions where the constituent parameters were altered to new positions. They used the Cleveland and Switzerland heart disease database at UCI machine learning repository. They took only 14 attributes, which were found to be more effective in diagnosing heart diseases, out of a total of 76, and used 90/250 cases from the Cleveland database and 50/122 cases from the Switzerland hospital. They achieved an accuracy of 94% after optimization with specificity and sensitivity of 92.5% and 95.6%, respectively. Yang et al. [24] developed a prediction model for the risk analysis of heart disease by optimizing an adaptive network based fuzzy inference system and linear discriminant analysis. Korean National Health and Nutrition Examinations Survey dataset was used for their study, which consisted of 4,826 patients with seven input variables and one output variable; 70% of the data was used as training data and 30% as testing data. They compared their results with four other methods and achieved an accuracy of 80.2%, which was the best among all methods. Liu et al. [25] developed an approach to detect congestive heart failure by using a combination of SVM and three nonstandard heart rate variability measures. They used the data from MIT/BIH database and compared their results with k-nearest neighbor classifier, and achieved 100% accuracy, precision, and sensitivity. Easton et al. [26] applied a data mining approach to predict very short-term versus short/intermediate term post-stroke mortality. Naive Bayes classifier was used for classifying data obtained from the UK glucose insulin in stroke trial (GIST-UK), which included patients who had suffered an acute stroke as per GIST-UK protocol criteria. They determined the mortality rate of 933 patients over a three month period and compared their method with logistic regression and decision trees. They achieved 92.6% sensitivity and 69.4% specificity. Masethe and Masethe [27] used classification algorithms like J48, naïve Bayes, reduces error pruning tree (REPTree), classification & regression trees (CART), and Bayes Net for predicting heart attacks. Data were obtained from medical practitioners in South Africa and only 11 attributes namely, patient identification number, gender, cardiogram, age, heart rate, chest pain, blood pressure level, cholesterol, smoking, alcohol consumption and blood sugar level were used predictive factors. The results showed that J48 and REPTree performed better in terms of predictive accuracy. Zapata-Impata et al. [28] developed an approach using swarm intelligence for risk evaluation in congenital heart surgery. They used the combination of PSO with k-nearest neighbor algorithm to predict the risk of congenital heart diseases. Data were obtained from Pediatric Cardiovascular Surgery Department of Cardiovascular Foundation of Colombia and consisted of 83 attributes, of this, 75% was used as a training set and the remaining 25% as a test set, which achieved a classification hit rate of 63.55%. Yang et al. [29] proposed a hybrid model for identifying risk factors for heart disease. Their dataset included discharge summaries, clinical notes and letters obtained from Partners HealthCare of i2b2 corpus, and consisted of 1,304 medical reports of 296 patients. The data from 178 patients were used as a training set and that from 118 patients as test set. They achieved an overall micro-averaged F-measure of 0.915. El-Bialy et al. [30] used fast decision tree and pruned C4.5 tree data mining methods on four datasets, namely, Cleveland, Hungarian, V.A., and STATLOG project, and used 14 attributes out of 75. They achieved a classification accuracy of 77.5% and 78% for collected datasets, and 76.3% and 75.4% for all datasets with C4.4 and Fast Decision Tree, respectively. Jonnagaddala et al. [31] developed a method for heart disease risk assessment using text data mining technique, which is a reliable method as shown by previous performance. They used 1304 records of 296 diabetic patients and categorized them into three groups. They concluded that maximum patients had Framingham risk score between 10% and 20% and all of them were diabetic. Ilayaraja and Meyyappan [32] applied association rule mining algorithm based on pattern generation on a dataset of 1000 patients to predict the risk of heart diseases. They compared their data mining method with the Apriori algorithm of data mining and achieved better results with some parameters. Basu et al. [33] proposed a supervised learning framework for predicting patient's risk-of-readmission by using dynamic hierarchical classification. They obtained data from the Washington State Inpatient Dataset and

the Heart Failure cohort of Multi-Care Health Systems. They achieved 69.20% accuracy, 24.80% precision, and 53.60% recall. Sung et al. [34] developed a stroke severity index by mining administrative data. They obtained data of 3,577 patients with acute ischemic stroke from Taiwan's National Health Insurance Research Database. They used k-nearest neighbor model and compared their results with multiple linear regression and regression tree models, and achieved better performance with their model. Paul et al. [35] proposed a genetic algorithm (GA) method based on fuzzy decision support for diagnosing heart diseases. They used Cleveland, Hungarian, Switzerland, and Long Beach datasets from the UCI machine learning repository and considered 14 attributes, with which they achieved an overall 80% classification accuracy in all datasets. Kalaiselvi [36] applied k-nearest neighbor, naive Bayes and decision tree methods for diagnosing heart diseases. They obtained data from UCI machine learning repository and tested their method with 13 and 12 attributes. Using 13 attributes, they achieved accuracies of 96.5%, 94.43% and 96.1%, respectively, with k-nearest neighbor, naive bayes and decision tree methods, and with 12 attributes; the respective accuracies were 97.5%, 90.72% and 96.62%. Kelwade and Salankar [37] proposed a method for predicting cardiac arrhythmias using PSO. They used linear and nonlinear methods like largest Lyapunov exponent, spectral entropy, Hurst exponent, SD1/SD2 ratio, and normalized Low frequency and high frequency power components extracting features. MIT-BIH arrhythmia database was used in this study. PSO with artificial neural network was compared to multi-layer perceptron method and achieved an accuracy of 95.15%. Singh et al. [38] also applied the associative classification method of data mining for heart disease prediction. To implement their method, they used the Cleveland dataset of the UCI machine learning repository, Apriori and frequent pattern-growth (FP-growth) algorithms to find association rules, and k-nearest neighbor algorithm for classification. They achieved 99.19% accuracy with the combination of k-nearest neighbor and Apriori algorithms. Kelwade and Salankar [39] proposed using PSO with radial basis function neural network for predicting heart diseases. Radial basis function neural network proved to be very sensitive to parameter of spread and was optimized by PSO in training phase. They also extracted the linear and non-linear features from datasets obtained from the MIT-BIH arrhythmia database and achieved an overall accuracy of 96.3% with their approach. Sultana et al. [40] analyzed the KStar, J48, SMO, Bayes Net and multilayer perceptron data mining techniques in predicting heart diseases with the help of Waikato environment for knowledge analysis (WEKA) software. They used predictive accuracy, receiver operating characteristics curve and area under the curve (AUC) value for measuring the performance of each technique, and found better performance of SMO and Bayes Net compared to the KStar, multilayer perceptron and J48 techniques. Purushottam et al. [41] applied the decision tree and hill climbing data mining algorithms on the Cleveland dataset and achieved an accuracy of 86.3% in the testing phase and 87.3% in the training phase. Kavitha and Kannan [42] proposed a framework for classifying heart diseases with feature extraction and data mining on UCI repository datasets. They began with a data cleaning process using outlier analysis, and used the principal component analysis for feature extraction. Wrapper filter was used to improve classification. Rajathi and Radhamani [43] used the K-nearest neighbor classification technique with ant colony optimization (ACO) to predict heart diseases. Streptococcus pyogenes was used as the dataset and divided into training (70%) and testing (30%) sets. They compared their method with decision support, SVM, and k-nearest neighbor and obtained the best results with the latter, with 70.26% accuracy and 0.526 error rate. Small et al. [44] applied text mining to electronic heart diseases procedure reports obtained from the Hospital of the University of Pennsylvania electronic echocardiogram database, with good results vis-à-vis medical diagnosis. Tayefi et al. [45] applied the decision tree technique to predict heart disease on a dataset of 2346 individuals, of which 1159 were healthy and 1187 under angiography (405 with negative and 782 with positive angiography), using 10 out of 12 variables. They achieved 96% sensitivity, 87% specificity, and 94% accuracy. Arabasadi et al. [46] proposed a hybrid of neural network and genetic algorithm to detect heart diseases. They used Z-Alizadeh Sani dataset which contained the information of 303 patients, and analyzed 54 features from each patient. GA increased performance of neural network by approx.10%, and resulted in 93.85% accuracy, 97% sensitivity and 92% specificity.

Decision tree data mining was used for diagnosing heart diseases in the UCI repository datasets with 84.1% accuracy by Shouman et al. [14]. By implementing the J48 classifier of decision tree on the data of 7008 patients (PGI, Chandigarh), Taneja et al. [21] got an accuracy of 95.56%. The same

method was applied to 2032 patient data of Hull Life Lab by Bohacik et al. [22] with 77.6% accuracy. The decision tree with hill climbing algorithm that was used by Purushottam et al. [41] on the Cleveland dataset of UCI repository with 86.7% accuracy was also used by Tayefi et al. [45] on Ghaem Hospital, Mashhad-Iran dataset with 94% accuracy. However, any methodology based on decision tree is complex and time-consuming, especially with a large dataset which consists of a large number of branches. Liu et al. [25] compared classification methods other than the decision tree method, e.g., SVM, k-nearest neighbor, naïve Bayes, neural network, etc. SVM classifier is a discriminative and therefore did not give better performance when support vectors were too large. K-nearest neighbor method can be used for both classification and regression but it also not optimum for large datasets [34]. Naive Bayes is a simple probabilistic classifier based on Bayes theorem while GA is based on a natural selection process that can solve constrained or unconstrained optimization problems. Nevertheless, for accurate prediction, using any one classification method is not sufficient; therefore, to increase predictive accuracy, it is necessary to use a hybrid of different optimization methods and classifiers, for example, the combination of SVM and GA resulted in higher accuracy compared to that obtained by either method alone on the UCI repository dataset [15]. Similarly, a higher accuracy was obtained using GA and K-nearest neighbor than the latter alone on one dataset from a hospital in Andhra Pradesh, India and six datasets from the UCI repository [18]. The same optimization algorithm combined with neural network method resulted in good prediction accuracy when applied on Z-Alizadeh Sani dataset [46]. However, using an optimization algorithm does not ensure finding global maxima and the process becomes more complex due to the evolution operators involved. Better prediction accuracy can be achieved by using a hybrid of an optimization method and a classification method. The combination of k-nearest neighbor classifier and the ACO optimization method, which is based on the food searching behavior of ants, was applied to the dataset of rheumatic fever causing streptococcus pyogenes but the resulting accuracy of 70.6% was very low [43]. The combination of SVM, PSO and k-nearest neighbor was used on heart sound data of patients and healthy controls, and the predictive accuracy decreased with increasing iterations of 20, 50 and 100 [16]. The PSO algorithm, based on the social behavior of birds or fish, was introduced by Dr. Eberhart and Dr. Kennedy in 1995. It is similar to GA except that it does not use evolution operators like crossover and mutation [47]. Combination of PSO and decision tree on the UCI machine learning repository datasets resulted in a prediction accuracy of 94% [23]. When used with multi-layer perceptron and applied on MIT-BIH arrhythmia database, PSO achieved accuracy 95.15% [37]. By using the association rule mining concept on Cleveland heart diseases dataset, a lower risk of heart diseases was seen in women compared to men [17].

Table 6 Methodological analysis

Authors	**Method used**	**Data**	**Evaluation measures /results**
Shouman et al. [14]	Decision tree (Nine Voting Equal Frequency Discretization Gain Ratio)	Cleveland Clinic Foundation heart disease dataset from UCI.	Sensitivity: 77.9% Specificity: 85.2% Accuracy: 84.1%
Fida et al. [15]	SVM and GA	Four datasets: Cleveland, Statlog, SPECT of UCI and South African dataset.	For Cleveland dataset: Accuracy: 98.63% For SPECT dataset: Accuracy: 80.79% For Statlog dataset: Accuracy: 93.27% For South African dataset: Accuracy: 83.40%

Authors	Method used	Data	Evaluation measures /results
Elbedwehy et al. [16]	SVM, k-nearest neighbor and binary PSO	Taken 70 samples of heart sound of healthy and unhealthy data, 76 samples of Diastolic Murmur and 84 samples of Systolic Murmur consisting of two classes and 88 features.	In 20 iterations Sensitivity: 95.24% Specificity: 95% Accuracy: 95.12% In 50 iterations Sensitivity: 85.71% Specificity: 95% Accuracy: 90.24% In 100 iterations Sensitivity: 95.24% Specificity: 80% Accuracy: 87.80%
Jabbar et al. [18]	GA and k-nearest neighbor	6 datasets of UCI heart disease dataset and 1 dataset taken from the hospital of Andhra Pradesh India.	Without GA Accuracy: 92.14% With GA Accuracy: 95.73%
Austin et al. [19]	Bagging, boosting, Random forests, and SVM	Number of data 3697 patients for training 4515 patients for testing	By random forest Sensitivity: 37.8% Positive predictive value (PPV): 69.6% Specificity: 89.7% Negative predictive value (NPV): 69.7%
Alizadehsani et al. [20]	Naïve Bayes, SMO, bagging and neural network	The Z-Alizadeh Sani dataset.	Accuracy by using bagging SMO: 93.40% naïve Bayes: 75.51% SMO: 94.08% neural network: 88.11%
Persi et al.[23]	Decision tree algorithm, fuzzy systems and PSO	Cleveland and Switzerland heart disease database from UCI machine learning repository	On Cleveland database: Before optimization Accuracy: 92.2% After optimization Accuracy: 94.4% On Switzerland database: Before optimization Accuracy: 86% After optimization Accuracy: 94%
Yang et al. [24]	Adaptive network-based fuzzy inference system and Linear discriminant analysis	Dataset of Korean national health and nutrition examinations survey	Accuracy: 80.2%
El-Bialy et al. [30]	Fast decision tree and C4.5 Algorithm	UCI repository	Accuracy: 77.5 (by using C4.5) and 78% (by using fast decision tree)
Basu et al. [33]	Dynamic hierarchical classification	Number of data Washington State inpatient dataset and the heart failure cohort data from multi care health systems (MHS).	Accuracy: 69.20% Precision: 24.80% Recall:53.60%
Paul et al. [35]	GA based fuzzy decision support system.	UCI machine learning repository.	Classification accuracy: 80%
Kalaiselvi [36]	K-nearest neighbor, naive Bayes and decision tree	UCI machine learning repository.	With 13 Attributes by k-nearest neighbor

Authors	Method used	Data	Evaluation measures /results
			Accuracy: 96.5% By naive Bayes Accuracy: 94.43% By decision tree Accuracy: 96.1% With 12 Attributes by k-nearest neighbor Accuracy: 97.5% By naive Bayes Accuracy: 90.72% By Decision tree Accuracy: 96.62%
Kelwade and Salankar [37]	PSO and multi-layer perceptron (MLP)	MIT-BIH arrhythmia database.	Accuracy: 95.15%
Singh et al. [38]	Nearest neighbor with Aprior associative algorithms	Cleveland heart diseases dataset from the UCI machine learning repository.	Accuracy: 99.19
Kelwade and Salankar [39]	PSO and radial basis function neural network.	The ECG data, used in this work, is collected from the standard MIT-BIH arrhythmia database.	Accuracy: 96.3%
Sultana et al. [40]	Bayes Net, SMO, KStar, MLP and J48	UCI machine learning repository	Accuracy by using KStar: 83.67% J48: 74.43% SMO: 83.92% Bayes Net: 90.2% MLP: 85.27%
Purushottam et al. [41]	Decision tree and hill climbing algorithm	Cleveland heart diseases dataset	Accuracy: 86.7%
Rajathi and Radhamani [43]	K-Nearest neighbour algorithm with ACO	The dataset used in this work is Streptococcus Pyogenes bacteria that cause Rheumatic Fever, also known as Acute Rheumatic Fever.	Accuracy: 70.26%
Arabasadi et al. [46]	Neural network and GA	Z-Alizadeh Sani dataset	Accuracy, Sensitivity and Specificity are 93.85%, 97% and 92%

To summarize, data mining plays an important role in predicting heart diseases and the hybridization of various classifications and clustering techniques with bio-system inspired algorithms like GA, ACO or PSO improves predictive accuracy. In other words, the combination of data mining and the above optimization algorithms can prove to be a powerful predictive and diagnostic tool for clinicians. Table 6 summarizes the analyses based on different methodologies and their impact on the prediction of heart disease. The latter is clinically significant since early prediction or diagnosis increases the chances of recovery. A large amount of data that is generated on a daily basis in hospitals are already used for manual prediction, but the correlation between these data and disease symptoms have not been completely mapped yet. There is a need for automated data collection and computation in order to accurately predict disease risk.

4.Discussion

The purpose of this review was to analyze and compare the different methods used for predicting heart diseases. Many researchers have used the decision tree based classification methods of data mining, which excludes the need to identify the parametric nature of the correlation between predictor variables and outcomes [19]. However, this approach is not suitable when the datasets have many predictive features [34]. We observed that the multi interval equal frequency discretization with nine voting gain ratio decision tree provides better predictive accuracy [14]. Furthermore, a slight alteration in the standard decision tree wherein each part can be split many times and to which decision nodes can be attached anywhere improves heart diseases prediction. Furthermore, if the concept of fuzzy logic is added, it helps in handling the uncertainties that exist within the datasets [22]. Text

mining has also been used by some researchers [44] on heart diseases datasets but their studies were single center-based and considered limited conditions. A better predictive accuracy may be obtained with text mining if used on multi center data. Classification algorithms have been used to classify heart diseases data. Since the k-nearest neighbor method is mainly used for continuous attributes, its efficiency depends on the number of chosen clusters. When the number of samples is reduced by clustering them according to their super classes, the process becomes faster, but loses accuracy for noisy data [36]. According to Liu et al. [25], feature selection and its combinations are the main problems with the k-nearest neighbor method, which prompted the use of SVM classifier by Elbedwehy et al. [16]. It is also observed that the women have significantly less chances of developing heart diseases compared to men, indicating that resting ECG status is a reliable predictive factor for heart diseases [17]. Some researchers have used hybrid methods by combining classification or clustering methods along with some optimization algorithms, and have obtained higher accuracies compared to single classification methods [15, 16, 18, 37, 43, 46]. Jonnagaddala et al. [31] identified diabetes as a risk factor of heart diseases with the help of data mining. Several classification algorithms can improve the predictive accuracy of heart diseases, but it is a big challenge to build a model for predicting specific heart diseases [21]. According to Masethe and Masethe [27], there were no significant differences in the predictive accuracies of different classification algorithms when used on the same dataset but may become significant when variations in the parameters and attributes are considered.

5.Conclusion

As per global statistics, heart disease has one of the highest death rates worldwide and the mortality risk increases with age. In 2016 alone, heart diseases were responsible for around 34% of total deaths worldwide. According to this study, the risk of developing heart disease is also directly proportional to age and increases significantly in individuals older than 15 years. The grim statistics of heart diseases call for effective strategies for disease management. Various methodologies have been developed for predicting and detecting the diseases in their initial stages. A review of those methodologies shows that data mining techniques like classification, clustering, and association rule mining, as well as bio-systems inspired algorithms like GA, ACO and PSO predict and detect heart diseases fairly accurately. However, no single algorithm was sufficient for accurate prediction. Therefore, hybrid prediction frameworks based on a combination of classification or clustering methods and bioinspired algorithms can prove to be a landmark in heart disease prediction and detection.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Alarming Statistics from India. http://neocardiabcare.com/alarming-statistics-india.htm. Accessed 26 December 2017.

[2] Kahramanli H, Allahverdi N. Mining classification rules for liver disorders. International Journal of Mathematics and Computers in Simulation. 2009; 3(1):9-19.

[3] Cardiovascular vascular diseases statistics, WHO. http://www.who.int/cardiovascular_diseases/en/. Accessed 26 December 2017.

[4] Prabhakaran D, Jeemon P, Roy A. Cardiovascular diseases in India: current epidemiology and future directions. Circulation. 2016; 133(16):1605-20.

[5] Kelly BB, Fuster V. Promoting cardiovascular health in the developing world: a critical challenge to achieve global health. National Academies Press; 2010.

[6] Ouyang H. Africa's top health challenge: cardiovascular disease. The Atlantic Journal. 2014:17-25.

[7] Cardiovascular disease fact sheet in Australia. https://www.heartfoundation.org.au/about-us/what-we-do/heart-disease-in-australia/cardiovascular-disease-fact-sheet. Accessed 26 December 2017.

[8] Allender S, Scarborough P, Peto V, Rayner M, Leal J, Luengo-Fernandez R, et al. European cardiovascular disease statistics. European Heart Network, Brussels, England.2008.

[9] Cardiovascular diseases, WHO India. http://www.searo.who.int/india/topics/cardiovascular_diseases/en/. Accessed 26 December 2017.

[10] Heart Disease in the United States, CDC. https://www.cdc.gov/heartdisease/facts.htm. Accessed 26 December 2017.

[11] Institute of health metrics and evaluation (IHME). http://ghdx.healthdata.org/. Accessed 26 December 2017.

[12] Ritchie H and Roser M. Causes of death. Our world in data. https://ourworldindata.org/. Accessed 2 February 2018.

[13] Global terrorism database (GTD). https://www.start.umd.edu/gtd/. Accessed 26 December 2017.

[14] Shouman M, Turner T, Stocker R. Using decision tree for diagnosing heart disease patients. In proceedings of the Australasian data mining conference 2011 (pp. 23-30). Australian Computer Society.

[15] Fida B, Nazir M, Naveed N, Akram S. Heart disease classification ensemble optimization using genetic algorithm. In international multitopic conference 2011 (pp. 19-24). IEEE.
[16] Elbedwehy MN, Zawbaa HM, Ghali N, Hassanien AE. Detection of heart disease using binary particle swarm optimization. In federated conference on computer science and information systems 2012 (pp. 177-82). IEEE.
[17] Nahar J, Imam T, Tickle KS, Chen YP. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications. 2013; 40(4):1086-93.
[18] Jabbar MA, Deekshatulu BL, Chandra P. Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology. 2013; 10:85-94.
[19] Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology. 2013; 66(4):398-407.
[20] Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, et al. A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine. 2013; 111(1):52-61.
[21] Taneja A. Heart disease prediction system using data mining techniques. Oriental Journal of Computer Science and Technology. 2013; 6(4):457-66.
[22] Bohacik J, Kambhampati C, Davis DN, Cleland JG. Alternating decision tree applied to risk assessment of heart failure patients. Journal of Information Technologies. 2013; 6(2):25-33.
[23] Persi Pamela I, Gayathri P, Jaisankar N. A fuzzy optimization technique for the prediction of coronary heart disease using decision tree. International Journal of Engineering and Technology. 2013; 5(3):2506-14.
[24] Yang JG, Kim JK, Kang UG, Lee YH. Coronary heart disease optimization system on adaptive-network-based fuzzy inference system and linear discriminant analysis (ANFIS-LDA). Personal and Ubiquitous Computing. 2014; 18(6):1351-62.
[25] Liu G, Wang L, Wang Q, Zhou G, Wang Y, Jiang Q. A new approach to detect congestive heart failure using short-term heart rate variability measures. PloS One. 2014; 9(4).
[26] Easton JF, Stephens CR, Angelova M. Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. Computers in Biology and Medicine. 2014; 54:199-210.
[27] Masethe HD, Masethe MA. Prediction of heart disease using classification algorithms. In proceedings of the world congress on engineering and computer science 2014 (pp. 22-4).
[28] Zapata-Impata BS, Ruiz-Fernandez D, Monsalve-Torra A. Swarm intelligence applied to the risk evaluation for congenital heart surgery. In international conference of the engineering in medicine and biology society 2015 (pp. 214-7). IEEE.
[29] Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease. Journal of Biomedical Informatics. 2015; 58:171-82.
[30] El-Bialy R, Salamay MA, Karam OH, Khalifa ME. Feature analysis of coronary artery heart disease data sets. Procedia Computer Science. 2015; 65:459-68.
[31] Jonnagaddala J, Liaw ST, Ray P, Kumar M, Chang NW, Dai HJ. Coronary artery disease risk assessment from unstructured electronic health records using text mining. Journal of Biomedical Informatics. 2015; 58:203-10.
[32] Ilayaraja M, Meyyappan T. Efficient data mining method to predict the risk of heart diseases through frequent itemsets. Procedia Computer Science. 2015; 70:586-92.
[33] Basu Roy S, Teredesai A, Zolfaghar K, Liu R, Hazel D, Newman S, et al. Dynamic hierarchical classification for patient risk-of-readmission. In proceedings of the international conference on knowledge discovery and data mining 2015 (pp. 1691-700). ACM.
[34] Sung SF, Hsieh CY, Yang YH, Lin HJ, Chen CH, Chen YW, et al. Developing a stroke severity index based on administrative data was feasible using data mining techniques. Journal of Clinical Epidemiology. 2015; 68(11):1292-300.
[35] Paul AK, Shill PC, Rabin MR, Akhand MA. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In international conference on informatics, electronics and vision 2016 (pp. 145-50). IEEE.
[36] Kalaiselvi C. Diagnosing of heart diseases using average k-nearest neighbor algorithm of data mining. In international conference on computing for sustainable global development 2016 (pp. 3099-103). IEEE.
[37] Kelwade JP, Salankar SS. An optimal structure of multilayer perceptron using particle swarm optimization for the prediction of cardiac arrhythmias. In international conference on reliability, infocom technologies and optimization (Trends and Future Directions) 2016 (pp. 426-30). IEEE.
[38] Singh J, Kamra A, Singh H. Prediction of heart diseases using associative classification. In international conference on wireless networks and embedded systems 2016 (pp. 1-7). IEEE.
[39] Kelwade JP, Salankar SS. Prediction of heart abnormalities using particle swarm optimization in radial basis function neural network. In international conference on automatic control and dynamic optimization techniques 2016 (pp. 793-7). IEEE.
[40] Sultana M, Haider A, Uddin MS. Analysis of data mining techniques for heart disease prediction. In international conference on electrical engineering and information communication technology 2016 (pp. 1-5). IEEE.

[41] Purushottam, Saxena K, Sharma R. Efficient heart disease prediction system. Procedia Computer Science. 2016; 85:962-9.
[42] Kavitha R, Kannan E. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. In international conference on emerging trends in engineering, technology and science 2016 (pp. 1-5). IEEE.
[43] Rajathi S, Radhamani G. Prediction and analysis of rheumatic heart disease using KNN classification with ACO. In international conference on data mining and advanced computing 2016 (pp. 68-73). IEEE.
[44] Small AM, Kiss DH, Zlatsin Y, Birtwell DL, Williams H, Guerraty MA, et al. Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease. Journal of Biomedical Informatics. 2017; 72:77-84.
[45] Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H, et al. Hs-CRP is strongly associated with coronary heart disease (CHD): a data mining approach using decision tree algorithm. Computer Methods and Programs in Biomedicine. 2017; 141:105-9.
[46] Arabasadi Z, Alizadehsani R, Roshanzamir M, Moosaei H, Yarifard AA. Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm. Computer Methods and Programs in Biomedicine. 2017; 141:19-26.
[47] Kennedy J, Eberhart R. Particle swarm optimization. In proceedings of the international conference on neural networks 1995 (pp. 1942-8).

Animesh Dubey is a PhD Research Scholar in Computer Science, at the JK Lakshmipat University, Jaipur, India. His dissertation research, focus on the prediction and detection of Heart Diseases using Data Mining and Optimization Techniques. He completed his Master's degree and B.Tech in Computer Science in 2013 and 2009. He has Published various papers in international journals and conferences. His research areas are Automaton, Data Mining and Optimization.
Email:animeshdubey123@gmail.com