Journal of Theoretical and Applied Information Technology
30th June 2023. Vol.101. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GENETIC FEATURE SELECTION AND NAIVE BAYES FOR EFFICIENT HEART DISEASE PREDICTION

Dr. RALLA SURESH[1], Dr.NAGARATNA PARAMESHWAR HEGDE [2]

[1] Assistant Professor, Department of CSE (Data Science), ACE Engineering College, Hyderabad, India
[2] Professor, Department of CSE, Vasavi College of Engineering, Hyderabad, India
Email:[1] rella.suresh@gmail.com, [2] nagaratnaph@staff.vce.ac.in

ABSTRACT

Heart disease is the leading cause of death in the world and it is easier to treat when detected early. The data in the health sector are huge but have not been used potentially because of its complexity in the system. The main reason for the complexity is the lack of adequate data analysis tools in the key patterns. Machine learning can help inthe retrieval of useful information from existing data and it also helps in the training of a model to forecast patients' health, which is faster than clinical experimentation. The Cleveland heart datasets have been used in a lot of studies. The existing K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression, Naive Bayes, and other classifiers were used to limit the number of selected attributes. The present research applied Genetic algorithm (GN) modeling to discover the 16 related features of the heart data set from hospitals, in India. The proposed model showed transparent and reliable graphical representation with other attributes which has the ability to predict the disease. The proposed model obtained 95 % accuracy better than existing models.

Keywords: *Artificial Intelligence, Genetic algorithm, machine learning, Naive Bayes Classifier.*

1. INTRODUCTION

Cardiovascular diseases are a result of a blockage in blood vessels that leads to the abnormal function of the heart. The hardening of arteries that becomes thick and inflexible is known as Arteriosclerosis. A heart attack occurs when there is a blood clot or a congestion to circulation of blood from the heart. In the human body, the heart is an important organ and its pumps blood via the circulatory system's blood vessels. The blood flow helps deliver the oxygen needed for the body's cells to function. According to World Health Organization, heart disease affects both men and women equally. Therefore, machine learning techniques are utilized at various stages in the recurrent area of Artificial Intelligence having the active type of applications and researches in the last decade.

The prediction model can be useful as a tool to identify heart disease. The input selection and the feature selection process finds the relevant inputs from the model. The feature selection technique removes and identifies irrelevant and unneeded features. Thus,those variables will not contribute to decreasing the accuracy of the predictive model. The new population in each generation is created using the algorithm that selects the individuals based on the fitness level for the problem-based features. A delay in medical therapy and an inability to provide patients with proper recommendation of a prediction model are considered as risks in clinical diagnosis. The present proposed work utilizes Genetic algorithm modelling that discovered 16 relevant attributes that are related to heart data taken from the Indian hospitals.

The input dataset for ML is form a dataset used by data scientists from the UCI repository. The dataset attributes were from developed countries and a sample of the targeted population. The classification models built on the existing dataset might lead to risk underestimation or overestimation. So previously used classification models might not be appropriate for predicting CVD for a new set of populations.

An efficient classification model is required for each set of populations, and the model obtained will not have any validity concerns. The present research for CVD prediction aims to uncover

hidden knowledge in identifying relevant attributes from a dataset collected from various hospitals in Telangana and UCI repository. Each attribute in the dataset is analyzed based on Random Forest & GA for feature selection based on the algorithm the attributes is chosen.

The operation of the research paper is given as follows: Section 2 is the literature review that discusses about the exiting methodologies. The proposed method is discussed by Section 3 and the section 4 discusses both the outcomes and the section. The conclusion and future work is discussed in Section 5.

2. LITERATURE REVIEW

The heart disease prediction has been performed for many years and various techniques have been utilized such as Neural Network, Navie Bayes, Decision Tree,and SVM showed distinct levels of accuracy.

Venkatesh [1] utilized Naïve Bayes for Big Data Predictive Analytics that showed a better prediction of disease using UCI machine learning Repository data. The existing predictive analytics failed to make better decisions and thus the detection of heart disease was done by using machine learning UCI data that showed better predictions. The developed approach faced the problem of complexity during computation of the model showed improvement evaluated in terms of CPU utilization, accuracy, and processing time. Yet, with respect to the heterogeneous environments, the model was not suitable.

Thanga Selvi and Muthulakshmi [2] diagnosed heart disease using an Optimal Artificial Neural Network (OANN). The developed model utilized approaches such as the Teaching and Learning Based Optimization (TLBO) and Misclassified Instance Removal (DBMIR) algorithm. The selected features were undergone for classification using the ANN model. The UCI repository dataset was used in the developed model where that performed prediction ofdisease dataset. The developed model still required improvement in term of computation time as it was increased.

Magesh and Swarnalatha [3] developed a Cluster-Based Decision Tree Learning (CDTL) for performing an optimal feature selection model. The developed model utilized a UCI repository and Cleveland heart samples for performing classification. The CDTL model mainly performed 5stages that made portioned through the target label distribution from the original dataset. However, the grouping determined the topographies based on the randomness values using the distributed samples and class combination for each class.

Beulah Christalin Latha [4] developed an ensemble classification model for heart disease detection. The developed model investigated heart disease risk prediction using the ensemble classification approach. The developed investigated the ensemble classification that improved the classification accuracy for the weak algorithms that combined the multiple classifiers. However, the developed model still required enhancement using feature selection algorithm which showed improvement in terms of accuracy.

Mohammad Shafenoor Amin [5] utilized various features for prediction of heart disease. The main aim of the developed model was used for the selection of features for performing correct combination showed improvement in terms of performances of the model.The combined features were fed for classifiers such as Decision Tree, Support Vector Machine (SVM), Logistic Regression (LR), k-NN, Naïve Bayes and the developed Vote hybrid technique that showed better performance.

3. METHODOLOGY

The proposed research work undergoes the process of preprocessing to remove missing values and noise in the heart disease UCI. The current research work uses the modified genetic algorithm for the selection of important features where a huge number of heart disease data is collected. Several classification methods are established for heart disease prediction from different datasets. The method performances are authenticated and compare with the real time data and the flow diagram or the proposed research is given in Figure 1.

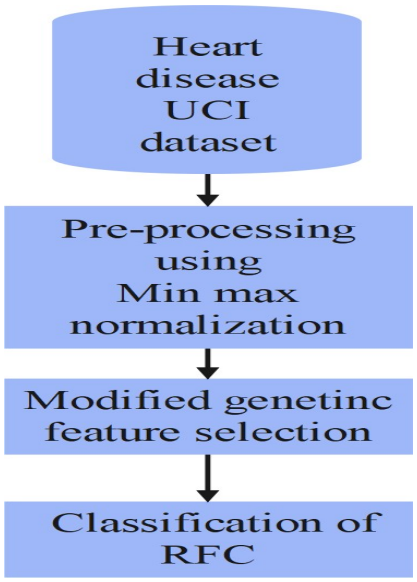

Figure. 1 Block diagram of the proposed research

3.1 Data acquisition

Table 1. Dataset sources

Data description	Cleveland dataset
Positive instances (%)	54.13
Features of count	13
Class count	2
Instance count	303
Negative instances (%)	45.13

The heart disease UCI data set contains 76 attributes that all published to the usage of 14 subgroups of them. The Cleveland database focused to attempt and differentiate the appearance and nonappearance values. Finding certain heart disease patient based on the variables is one of the key challenges in the UCI dataset. The other goal is to select and forecast several perspectives from thisdataset that can aid in better empathizing with the problems. This dataset collected of 14 subsets that are age, chest pain, exercise induced angina, all-out heart rate achieved, old peak – ST depression encouraged by exercise related to rest, sex, serum cholesterol, the slope of the peak workout ST segment, resting electrocardiographic results, fasting blood, sugar, resting blood pressure, amount of main thalassemia and vessels. Table 1 shows the Dataset sources and Table 2 shows the Attributes and the description.

Table 2. The description of the attributes

Attribute	Description of attributes
Age	years
Sex	0-Female 1- Male
Cp	Chest pain Type 1 – typical angina Type 2-atypical angina Type 3-nonangina Type 4-asymptomatic
Fbs	Fasting blood sugar
Restbp	Resting blood pressure
Thalach	Rate of heart beat
Chol	Serum Cholesterol
Oldpeak	Depression tempted by workout (ST wave)
Resstecg	Resting ECG results
Ca	Major vessels colored by fluoroscopy
Examg	Angina due to exercise
Class	Heart disease presence or Absence
Slope	Slant of the peak exercise

3.2 Preprocessing data

The preprocessing data is used to clean the data which is the most significant steps for achieving in the best from the heart disease diagnosis. The preprocessing is either improved or maintained the performance of heart disease classifiers. Min Max Normalization is a preprocessing step or scaling technique used in this study. The min max

normalization is a most widespread use in normalization. For every individual feature, the least value is changed to a 0, the highest maximum value is changed to a 1, and all supplementary values are changed to a decimal range between 0 and 1. The equation for the min max normalization process will be expressed in Equation
(1).X_{min} − Minimum value

X_{max} − Maximum value

3.3 Feature selection

Feature selection is the method of choosing a subsection of related features for structuring robust learning models through eliminating most superfluous and unrelated topographies from the data. Feature selection provides better generalization and reduces the chance of overfitting. It enhances the performance of learning methods by ModifiedGenetic Algorithm (GA). The genetic algorithm is a stochastic optimization method enthused through the method of usual selection, which is generally used to resolve various classes. GA upholds a populace of applicant resolutions via the selective process. GA is
prepared by a group of resolutions named population. Every individual explanation is known as chromosome. The chromosomes growth by Where, X_i − i^{th} data point consecutive iterations named generations. The GA uses the mutations to all individual parent chromosome, where random trading of gens happens. The modified GA applied for feature selection in the automated classification system and it is possible to increase the classification accuracy.

The modified GA supports to reduce the lots of tests which are necessity to be taken by the patients by decreasing the number of attributes. Initially, the dataset included 14 attributes for assessing the heart disease. At last, the fourteen attributes are condensed to six attributes. The algorithm will generate a new population at each generation by selecting every individual based on the value of fitness level under the problem domain. These individuals are together recombined based on the natural genetics operators that were borrowed. The performance of modified genetic algorithm reliesson a number of problems such as fitness function, mutation and crossover.

3.3.1 Mutation

The crossover operator will result in offspring that resemble the parents, which is the primary reasonfor the minimal variety photographs in a generation. The offspring will be similar to that of the parents. This will cause the newer generation to the lower diversity. Therefore, the mutation operator will solve the problem of value changing from some features that randomly generate the offspring Therefore, to check the feature is mutated or not, the random numbers are generated between the values 0 and 1. If the number is low compared to other values, then the mutation rate will be flipped. The mutation rate with respect to the standard value is obtained as 1/m. The equation for the mutation process will be expressed in Equation (2).

$$X_{norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (1)$$

$$x' = \min(\max(N(x, \sigma), a), b$$

Where, σ − Depending on the duration of the time or gap.

3.3.2 Crossover

The operator for selecting the half population has been chosen by recombining the crossover operator for an individual that selects and generates the new population. The operator is picked into random individuals that combined the features for getting four offspring with a new population. The new population is reached up to the size of the old function. The uniform crossover technique decides to be whether each of the offspring is different from one another.

3.3.3 Fitness function

Once after the initialization, the fitness value should be assigned for each of the individuals in the population. The neural network has to be trained with the network instances for evaluation of the error by a selection of instances. The selection error is generating lower values of fitness function and all these values showed higher fitness to recombine. The
rank-based fitness assignment is used for fitness assignment that shows the selection of errors of each of the sorted individuals. The values obtained for the fitness function are assigned with each of the individuals' dependent on the error selection which was sorted based on individuals. According to where they are in space, not according to the

actual selection of error, each person will be assigned to fitness values. Based on the rank-based approach, the fitness function value is assigned for each individual. The approach assigns the fitness values for each individual has the rank which is derived as shown in Equation (3).

$\Phi(i) = k \cdot R(i)$ (3)
Where,
$i = 1, \ldots, N.\Phi(i) = k \cdot R(i)$ where$i = 1, , N$

Here, the selective pressure is represented as k and the value will be ranging from 1 to 2. The selective pressure values are made with the individuals having more probability for recombination with the fittest individuals. The rank of an individual i for the parameter is represented as R(i) . Once the fitness function is assigned the operation of selection is performed to choose and recombine the individuals for the next generation. The individuals are likely to survive only to their fitness level. The selected individuals are represented
as N2 having the size of the population as size N. The selection method will replace all the individuals having the features that are selected randomly. The individuals are selected and correspondingly perform according to the recombination. The selection probability of every individual is stated in Equation (4), which is also known as fitness proportional selection

$$pi = \frac{f_i}{\sum N_{k=1}} \quad (4)$$

Where,
pi − Probability of individual i selection
fi − Fitness of possibilities
N − Number of possibilities in the population

3.4 Random Forest classifier

The Random Forest classification is a machine learning algorithm that works on non-linear propensity of data set but gives good outcomes than Decision Tree algorithm. The random forest classification using an ensemble of deep features perhaps supports enhancing performance expressively. The RF algorithm has been utilized in probability and prediction estimation. RF contains of several decision trees. Every individual tree offers a vote that specify the result about class of the object. The RF consumes a propensity to distinct non linearly reliant dataset, that supported executing this algorithm on heart disease dataset. RF requires a good adjustment to produce good outcomes, therefore by altering the parameters such as lots of trees, randomness, least number of leaf nodes and least number of splits the accuracy can be improved. In it, some samples of data are created unsystematically after the innovative dataset with spare and every individual decision tree is qualified on various models of data. Features are also nominated aimlessly while tree construction. Forecast made by several trees are merged employing a common vote. To evaluate the occurrence of heart disease the random forest classifiers are used. It achieved accuracy of 86.9% for heart disease anticipation with specificity value 82.7% and sensitivity value 90.6%. The Gini impurity (Split criterion) is calculated as Equation (5)

$$G = \sum_{j=1} p(j) * (1 - p(j)) \quad C \quad (5)$$

Where, C −Number of classes p(j) − Probability of choosing a class j data point For tree learning, it mostly put on bootstrap bagging or aggregating. For a given input, X = {x1 x2, x3, ..., xn} with replies Y = {x1, x2, x3, ..., xn} that replicate the bagging from b = 1 to B. The unnoticed models x′ is complete by calculating the forecasts $\sum_{b=1}^{B} fb(x')$ from each single tree on x′.

$$j = \frac{1}{B} \sum_{b=1}^{B} fb(x') \quad (6)$$

The indecision of prediction on these trees is done by its standard Equation, $\sigma = \sum \frac{(fb(x') - f) \; B \; b=1}{B-1}$ (7)

3.3 Random Forest classifier

The Random Forest classification is a machine as in Equation (8).

PERFORMANCE MEASURES

The results are estimated with respect to precision, recall and accuracy for evaluating the parameter.

Precision

Precision refers to the fraction of retrieved documents to that of the query which are relevant

$$Precision = \frac{TP}{FP+TP} \quad (8)$$

Recall

Recall is the ratio of the related documents that are recovered successfully as shown in equation. (9).

$$Recall = \frac{TP}{}$$

FN=TP

Accuracy

Accuracy is defined as the fraction of true positives and true negatives to that of the positive and negative observations which is expressed as shown in the Equation (10).

$$\text{Accuracy} = \frac{(TN+TP)}{TN+TP+FN+FP} \quad (10)$$

Table 3. The results obtained for the proposed research in terms of Precision, Recall, F-score, Accuracy

Classes	Precision	Recall	F1 score	Accuracy	Support
0	93	96	94	97	85
1	18	20	28	23	123
avg	92	94	94	95	208

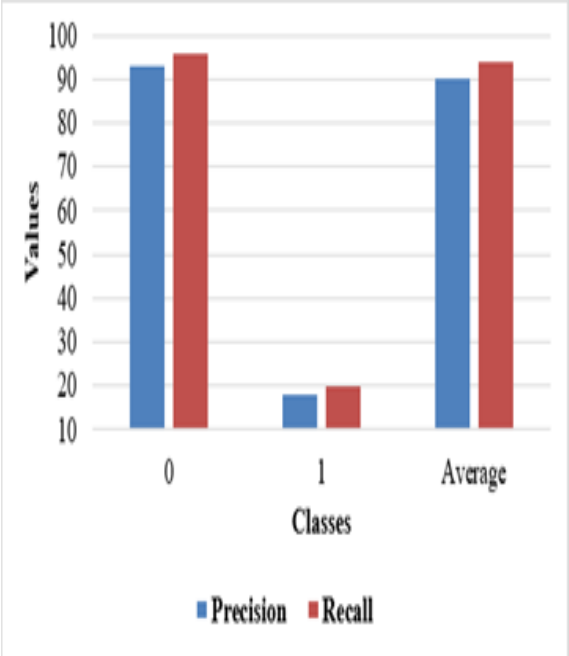

Figure. 3 Proposed method results obtained with respect to precision and recal

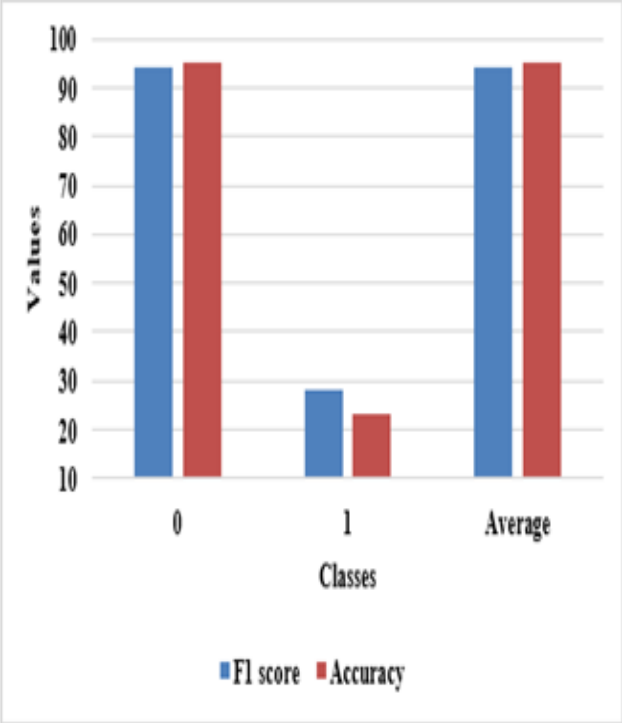

Figure. 4 The results obtained with respect to accuracy and F-score

Fig. 3 shows the proposed Genetic feature selection algorithm with Naïve Bayes showed achievement of the accuracy of 97%, F-score of 94 %, Recall of 96 %, and precision of 93 %. Similarly, to class 1, the accuracy value is obtained as 23 %, F-score of 28 %, Recall of 20%, and precision of 18%. Similarly, the average values for the precision are obtained as 92%, recall of 94%, accuracy of 95 %., F-score of 94 %. Fig 4 represents the graph plotted for the proposed GA-NB model showing improvement in terms of f-score and accuracy. The GA-NB model obtained the results that increase the computation time during model prediction. Table 3 has the proposed method outcomes evaluated in terms of recall, f-score, accuracy, and precision.

4.1 Comparative Analysis

Table 4 includes the results that are obtained for the existing and the proposed methods. Selvi [12] utilized a Random Forest model which obtained an accuracy of 82.18 % and 83.93 % of F-score. An effective tool was utilized for analyzing the big data to predict the heart disease that showed satisfying results when the number of patients' data is increased. Similarly, the Naïve Bayes with the LSTM classifier was used for improving the performances. The existing ensemble-based

classification model obtained an accuracy of 85.48% and the Vote with Naïve Bayes and Logistic regression model obtained an accuracy of 87.41 %. The existing RF with NB used only the irrelevant features for heart disease prediction resulted in the overfitting problems, Similarly, the weighted ensemble model and cluster based DT learning obtained lower accuracy values because of the computation complexity created in the system. Whereas, the proposed model used the Genetic feature selection algorithm that selected the relevant features and was undergone for the classification using the NB model. Fig 5 shows the graph of the proposed method results compared with the existing models.

Database	Author	Method	Accuracy	F score
UCI machine learning repository	Venkatesh [13]	Naive Bayes	82.83	91
	Selvi [14]	Random Forest	82.18	83.98
	Magesh [15]	cluster-based DT learning	76.70	77.0
	C. Beulah Christalin Latha [16]	ensemble classification techniques	85.48	86.6
	Mohammad Shafenoor Amin,[17]	Vote with Naïve Bayes and Logistic Regression	87.41	88.9
	Proposed	GA-NB algorithm	95	96

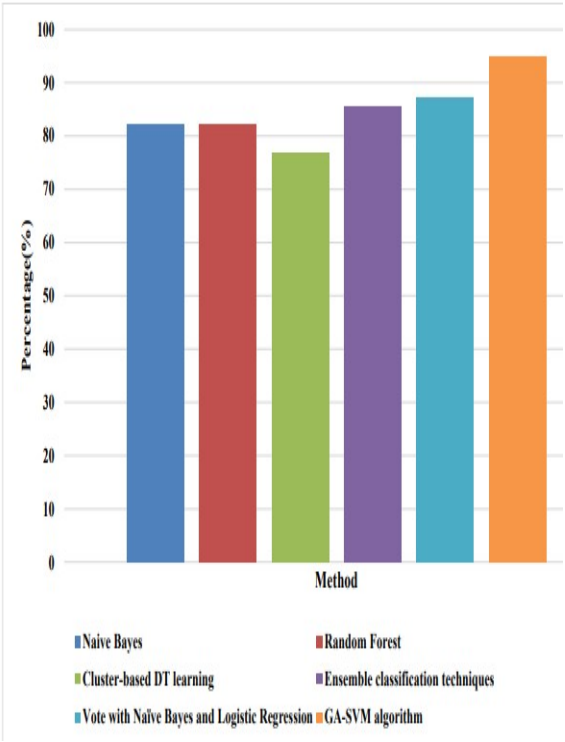

Figure. 5 Comparative Analysis

4. CONCLUSION

Heart disease prediction is a crucial component of big data analytics and is employed to address the issue of cardiovascular disease. Big data analysis has greater chances for forecasting the status of health metrics and has produced better decisions in terms of heart disease prediction. The models were consisting of more number of traffic data that showed uncertainty and difficulty

and operate it at the medical service, which accurately analyzed the patient's history data to for predicting the heart disease. Big data is used to ensure solve this issue. The dependence model's graph was displayed using the Genetic and Naive Bayes model employed in the suggested study. The model was able to quickly identify the correlations between the attributes, according to the model's acquired results. The present research used the dataset which was taken from various hospitals in India. The performances of the Genetic algorithm and Naive Bayes have outperformed the K-nearest neighbor and Support vector machine in the prediction of heart diseases. The proposed model obtained 95 % accuracy better than existing models such as Random Forest of 82.18 %, Naive Bayes obtained 82.34 % and cluster-based DT learning of

76.70 %. In the future, the same algorithms to be implemented with real-time data for estimating the effectiveness of the system.

Conflicts of Interest The authors declare no conflict of interest. Author Contributions The paper background work, conceptualization, methodology, editing draft and visualization, dataset collection, implementation, result analysis and comparison, preparing have been done by 1st author. The supervision, review of work and project administration, has been done by 2nd author.

REFERENCES

[1] S. Mai, T. Turner and R. Stocker, "Integrating Naive Bayes and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients", *InternationalConference of Data Mining & Knowledge Management Process,* 2012

[2] I. H. Witten, E. Frank, "Data Mining: Practical machine learning tools and techniques", *The Morgan Kaufmann Series in Data Management Systems,* 2005.

[3] S. Palaniappan, and R. Awang, "Intelligent HeartDisease Prediction System Using Data Mining Techniques", *IEEE/ACS International Conference on Computer Systems and Applications,* 2008.

[4] V. Sree Hari Rao, and M. Naresh Kumar, "Novel Approaches for Predicting Risk Factors ofAtherosclerosis", *IEEE Journal of Biomedicaland Health Informatics*, Vol. 17, No. 1, 2013.

[5] S. Mai, T. Turner and R. Stocker, "Using Data Mining Techniques in Heart Disease Diagnosis And Treatment", *Japan-Egypt Conference on Electronics, Communications and Computers* 2012.

[6] G. Monika, and S. Narayan Singh, "Predictions in Heart Disease Using Techniques of Data Mining", *International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, 2015.

[7] P. Theresa and R.J. Thomas, "Human Heart Disease Prediction System using Data MiningTechniques", *International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 2016.

[8] H. Mohammad, Tekieh, B. Raahemi, Importance of Data Mining in Healthcare: A Survey, "I*nternational Conference on Circuit, Power and Computing Technologies (ICCPCT)*",2015.

[9] D. Ankita, M. Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", *2nd International Conference on Computing for Sustainable Global Development (INDIACom),* 2015.

[10] A. Zoubida, M. Mourad El Yadari, A. Benyoussef, and A. El Kenz, "Study and analysis of Data Mining for Healthcare", *4th IEEE International Colloquium on Information Scienceand Technology*, 2016.

[11] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the RFRSmethod", *Computational and mathematical methods in medicine*, 2017.

[12] I.A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods", *arXivpreprint arXiv:1704.02799.*

[13] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of big data predictive analytics model for disease prediction using machine learning technique", *Journal of medicalsystems*, Vol. 43, No. 8, pp. 1-8, 2019.

[14] R.T. Selvi, and I. Muthulakshmi, "An optimal artificial neural network based big data application for heart disease diagnosis and classification model," *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12,No. 6, pp. 6129-6139, 2021.

[15] G. Magesh, and P. Swarnalatha, "Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction", *Evolutionary Intelligence*, Vol. 14, No. 2, pp. 583-593, 2021.

[16] C.B.C. Latha, and S.C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Informatics in MedicineUnlocked*, Vol. 16, pp. 100203, 2019.

[17] M.S. Amin, Y.K. Chiam, and K.D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease", *Telematics and Informatics*, Vol. 36, pp. 82-93, 2019.

[18] "Cleveland Heart Disease Dataset," accessed on03-02-2017.[Online].Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease