DOI: 10.21917/ijsc.2022.0366

CARDIOVASCULAR DISEASE PREDICTION USING ENSEMBLE CLASSIFICATION ALGORITHM IN MACHINE LEARNING

Rajarshi Sinha Roy[1] and Anupam Sen[2]
[1]*Department of Computer Science, St. Xavier's College, India*
[2]*Department of Computer Science, Government General Degree College, India*

Abstract

Cardiovascular disease includes a wide range of heart-related illnesses and has surpassed cancer as the top cause of mortality worldwide in recent decades. Many people nowadays are engrossed in their daily lives and engage in various activities while ignoring their health. As a result of their rushed lifestyles and disrespect for their health, the number of people becoming unwell is increasing every day. According to the World Health Organization, heart disease claims the lives of over 31% of the world's population. As a result, doctors must be able to predict whether a patient may develop heart illness, but the amount of data collected by the medical sector or hospitals, on the other hand, is so vast that it can be difficult to analyze at times. This research paper assessed several aspects of heart illness and develops a model based on supervised learning methods like Gaussian Naïve Bayes and AdaBoosting algorithm. The purpose of this research is to figure out how to anticipate whether a patient will develop heart disease. The AdaBoosting algorithm achieves a great accuracy score of 95%, according to the data.

Keywords:

Heart Disease Prediction, GaussianNB, Machine Learning, AdaBoosting Algorithm, Healthcare

1. INTRODUCTION

In today's fast-paced society, everybody is so preoccupied with their lives and careers that many dont even have any time to look for themselves. Because of their hurried lives, most individuals suffer from stress, anxiety, despair, and a variety of other ailments. And this also makes it easier for them to eat fast food. They are becoming unwell and suffering from severe ailments as a result of these key factors. Cardiovascular disease is the most common disease which is the leading cause of mortality in the medical world. In the chest, the heart is located slightly below the lungs. Blood is circulated throughout the body by the heart. When you have a cardiac arrest, it stops pumping blood. Its vital to keep it healthy because it performs so many tasks in our bodies. The hearts proper functioning is essential for the human bodys survival. It has been linked to a number of heart-related illnesses. Cardiovascular disease is one of them. According to the World Health Organization, 17.9 million people die each year from cardiovascular disease. According to WHO figures, heart disease is responsible for around 31% of all fatalities globally, and also heart-related illnesses cost India $237 billion between 2005 and 2015.

Within the healthcare industry, machine learning is a rapidly growing trend. ML and data mining are important in the healthcare industry for disease prediction. Data mining is the process of extracting essential information from large datasets in a variety of disciplines, including medicine, business, and education. Machine learning techniques are needed to mine vast amounts of patient data in the medical profession. These data are analyzed by healthcare professionals in order for them to make efficient diagnostic decisions. Machine learning can anticipate heart problems, allowing for a more accurate diagnosis. The era can also assist medical professionals in analyzing data to find qualities or red flags that could lead to advanced diagnostics and medication.

Machine Learning algorithms are capable of a wide range of tasks, including prediction, classification, and decision-making. Training data is required to learn the ML algorithms. Following the learning phase, a model is created as a result of the ML algorithm. After that, the model is tested and verified on a collection of previously unseen real-time test datasets. Machine learning has the potential to predict cardiovascular problems, leading to a much more accurate diagnosis at a lesser cost than the traditional method. The classification algorithms can be taught and tested in order to create predictions about a person's likelihood of developing heart disease. We used a dataset from the UCI repository for this study. For the prediction of cardiac disease, the classification model was created utilizing classification techniques and a boosting algorithm is used to modify the model more accurately. We have endeavored to analyze all the risks and factors that affect the heart and potentially contribute to cardiac disease in this study. The literature review is given in the second section. We went over the Proposed Approach and procedures we utilized to forecast cardiac disease in part 3. In section 4, the Machine learning algorithms are discussed. In addition, section 5 and 6, which are the results and conclusion. Lastly, references, have been added to our paper.

2. LITERATURE REVIEW

The authors of [1] conducted research on machine learning algorithms and methodology that were applied to a wide range of medical datasets. Researchers choose supervised-learning approaches such as SVM, KNN, Naive Bayes, Decision Trees (DT), and ensemble models. Machine learning and machine learning-based algorithms have proven to be quite accurate in forecasting cardiac disease. When combined with PCA, alternating decision trees have also performed admirably. The authors of this research highlighted that Random Forest and Ensemble models perform considerably better than other models in solving the challenge. The authors of [3] also show how to extract knowledge using data mining approaches that are already being used in heart disease prediction research. They discussed Naive Bayes, Neural Networks, and Decision Trees, and concluded that the number of features included in the prediction could alter the forecasts accuracy. In [4] constructed a prototype IHDPS using data mining algorithms such as Decision Trees, Nave Bayes, and Neural Network. IHDPS can estimate the likelihood of individuals developing heart disease based on 13 medical attributes. Also, they used classification matrices to see

if the prediction was accurate. This concept may be utilized to train and educate medical students while also saving money. The study looks at multiple classification algorithms for determining a persons threat level in terms of age, gender, bp, lipid, and heartbeat in this publication [5]. To identify the patients overall risk, classification methods like Bayesian classification, k-nearest neighbors, decision Tree algorithm, Neural Network, and many others are applied. The danger levels accuracy is extremely high when a big number of attributes are applied. They looked at several traits and statistical risks, such as exceeding 50 percent, less than 50 percent, and 0%, to predict heart disease. The K-nearest neighbors and ID3 Classification algorithms were used to find the cardiovascular risk rate, as well as the accuracy rate for several factors. In paper [6], the studys main purpose is to use the Naive Bayes machine learning modeling technique to construct an Intelligent Heart Disease Prediction System [6]. It can help doctors improve clinical decisions more than traditional decision support systems that rely on user answers. It can help doctors make better clinical judgments by providing solutions to complex questions concerning heart disease diagnosis. The use of random forest and naive Bayes to forecast heart illness was advocated in this study [7]. For feature selection, SVM-RFE and gain ratio algorithms are utilized. This strategy helps to improve accuracy while also cutting down on computation time. According to the experiments, the proposed feature selection technique improves accuracy for both models. In [8], similarity feature extraction (CFS) is utilized to evaluate subsets. The system outperforms previous techniques in the CHDD test, with an accuracy of 91.6 percent. It obtains a 97 percent accuracy in the Peoples Hospital datasets test, which is better than most other classifiers save SVM (98.9%). For the prediction of heart disease, a model that used a random forest algorithm or modified random forest is proposed, which did quite well when compared to the classic random forest technique.

The usefulness of the prominent classification methods kNN algorithm and ACO optimization are combined in this study [9] to predict the possibility of heart disease. To estimate the chances of acquiring heart disease, the k-Nearest Neighbour (kNN) method is combined with ACO. The bacteria Streptococcus Pyogenes, which cause Rheumatic Fever, also known as Acute Rheumatic Fever, were employed in this study (ARF). The kNN classification is often utilized to categorize the test data in the first step.

3. PROPOSED APPROACH

The approaches used for the research in this paper have been detailed in this section. We have discussed how we approached the experiment by considering major risk factors and the strategies we utilized to predict heart illness. The systems processing begins with data collecting, for which we employ the UCI repository dataset, which has been thoroughly confirmed by a number of researchers and the UCI authority [12].

3.1 DATASET AND ATTRIBUTES

The UCI machine learning resource was used to gather data on heart illness. There are 13 factors that can be used to predict cardiac disease. Despite the fact that the Cleveland dataset contains 76 properties, the data set supplied in the repository only contains information for 14 of them. There are 303 records in the collection.

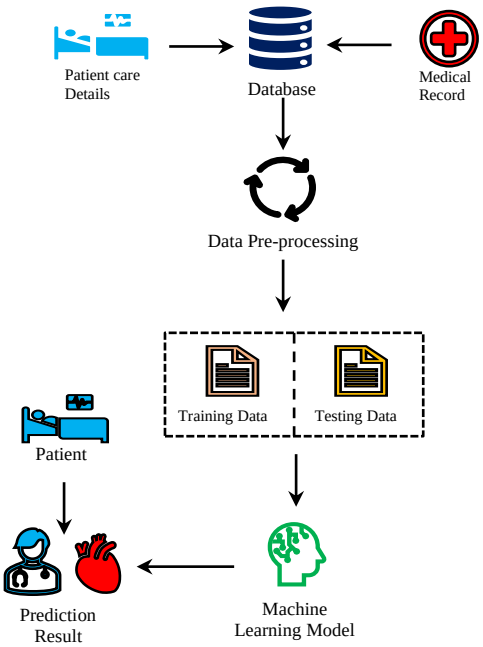

Fig.1. An architecture of Our Proposed method for predicting a heart attack

The 14 attributes are : Age - age of a person- minimum range is 29 and maximum is 77, Sex- Gender of a person – 1 for male, 0 for female, Cp – chest pain- minimum range is 0 and maximum is 3, Testbps - resting blood pressure (in mm Hg on admission to the hospital) - minimum range is 94mm Hg and maximum is 200mm Hg, Chol- serum cholesterol in mg/dl - minimum range is 126mg/dl and maximum is 564mg/dl, Fbs - fasting blood sugar & gt; 120 mg/dl – 1 for true and 0 for false, Restecg - resting electrocardiographic results - minimum range is 0 and maximum is 2, Thalach - maximum heart rate achieved – minimum is 71 and maximum is 202, Exang - exercise induced angina – 1 for yes and 0 for no, Oldpeak - ST depression induced by exercise relative to rest - minimum range is 0 and maximum is 6.2, Slope - the slope of the peak exercise ST segment - minimum range is 0 and maximum is 2, Ca - number of major vessels (0-3) coloured by fluoroscopy - minimum range is 0 and maximum is 4, Thal - thallium scan - minimum range is 0 and maximum is 3, Target - Heart disease – 1 for yes and 0 for no.

Table.1. UCI Dataset for Heart-Disease Prediction

Attributes	Min	Max	Std	Mean
age	29	77	9.082101	54.366337
sex	0	1	0.466011	0.683168
cp	0	3	1.032052	0.966997
trestbps	94	200	17.53814	131.623762
chol	126	564	51.83075	246.264026
fbs	0	1	0.356198	0.148515
restecg	0	2	0.525860	0.528053
thalachh	71	202	22.90516	149.646865
exang	0	1	0.469794	0.326733

oldpeak	0	6.2	1.161075	1.039604
slope	0	2	0.616226	1.399340
ca	0	4	1.022606	0.729373
thal	0	3	0.612277	2.313531
target	0	1	0.498835	0.544554

3.2 DATA PRE-PROCESSING

The Large numbers of incomplete and noisy data are present in real-life data. Such information needs to be pre-processed to eliminate these flaws and make confident predictions. To eliminate or alter categories, eliminate repeating groups, handle the null values and outlier data, etc., the gathered data generally includes noisy and null values. In order for the data to fall within a given range, one may need to normalize and scale it. The process of turning unprocessed data into the form or structure that is more appropriate for model development and data exploration in general is known as data transformation. It is a crucial phase in feature extraction that makes finding insights easier. The data might not always rely on a single source but rather from several, and it must be combined before being processed. It is called Data Integration. The information gathered is complex, and it must be presented in order to yield useful results. This process is called Reduction of Data.

The system was trained with 80% of the dataset and tested with 20% of the dataset.

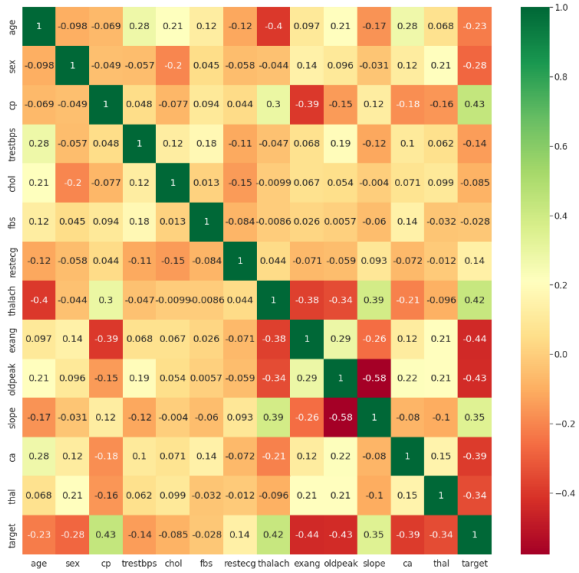

Fig.2. An illustration of Pearson correlation coefficient on UCI dataset

3.3 FEATURE SELECTION

Feature selection is a method of deleting duplicate, unnecessary, or noisy characteristics from a feature collection in order to choose a subset among the most relevant features. The key goal is to select the optimal characteristics to help the model work successfully. There are several methods for feature selection but in our research, we have used the most common technique, Pearson correlation coefficient.

The Pearson Product-Moment Correlation, *r*, can be anything between +1 and -1. A value of 0 implies that the two variables have no relationship. A positive relationship is shown by a value higher than 0 means two variables' values are linearly proportional. A negative relationship is indicated by below 0 means two variables' values are inversely proportional.

3.4 BALANCING THE DATASET

Data balancing is necessary for reliable results since we can tell from the data balance diagram that these target classes are equivalent. In Fig.4. The red one is for heart disease patients and blue one for others.

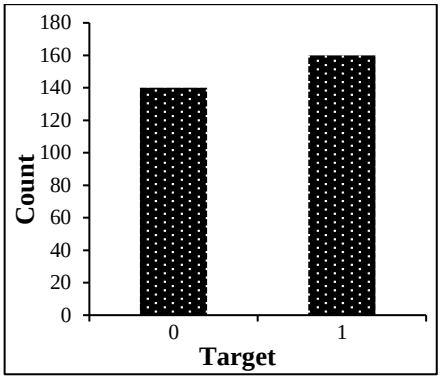

Fig.3. An illustration of Pearson correlation coefficient on UCI dataset

3.5 HISTOGRAM OF THE DATASET

The attributes of the dataset are represented by a histogram of attributes, which can be used to determine if the range is evenly distributed or not. In Fig.5. we can see that the range is evenly distributed.

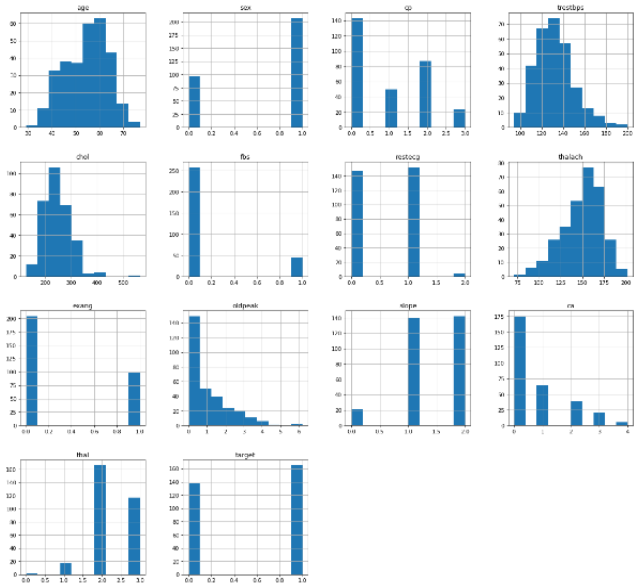

Fig.4. A histogram of 14 attributes is present in the dataset

4. MACHINE LEARNING

Machine Learning is an area of artificial intelligence where the ultimate goal of ML is to construct systems that can train and

predict the future based on past experiences. It builds a model by using ml techniques to train on a training dataset. Using the input dataset, the model predicts cardiovascular illness. It creates models by employing algorithms to find underlying trends within the dataset. Any missing values have been filled in, and the dataset has been cleansed. The models accuracy can then be tested using the dataset to predict cardiovascular disease.

4.1 SUPERVISED LEARNING

The Supervised learning is a type of ML algorithm in which systems are trained using large data sets and then used to forecast output where the dataset contained with output and all features are labeled. It includes two types: classification and regression.

4.2 UNSUPERVISED LEARNING

Unsupervised machine learning is a branch of machine learning inside which systems are trained on Unlabelled data sets and then allowed to make decisions on it without observation. Unsupervised ML tries to find the internal mechanism of a data, classify data on the basis of commonalities, and present the set of data in a minimally invasive manner.

4.3 REINFORCEMENT LEARNING

Reinforcement learning is distinct from supervised learning in that it does not involve the solution key. In this strategy, the models presentation is helped by its relationship to the surrounding environment. The model learns from its failures because it doesn't use a tagged dataset and the results aren't related to data. Because it lacks a training dataset, it is forced to learn from its mistakes.

5. MACHINE LEARNING CLASSIFICATION

The process of categorizing data into groups based on the correlation between different data pieces is referred to as classification. Categorization is used to anticipate cardiovascular problems in this scenario. There are several machine learning models available, however the suggested method can use any of the following techniques or models. To forecast the aim, we used a number of techniques. However, by utilizing an ensemble technique and the concept of hyper tuning, we achieve the best outcomes. The following are the algorithms that were used:

5.1 Naïve Bayes

Naïve Bayes classification is a type of supervised machine learning algorithm. The Bayes Theorem will be Eq.(1) for this type of dataset. The Bayes theorem is indeed a statistical concept that can be applied for probabilistic predictions. If the dataset has several features and the output is y, the corresponding values are $x_1, x_2, \ldots, x_n$ in a classification issue.

$$P\left(X_n \left| b_i \right.\right) = \frac{P\left(b_i \left| X_n \right.\right) P\left(X_n\right)}{p\left(b_i\right)} \quad (1)$$

$$P\left(X_1, X_2, ..., X_n \left| b \right.\right) = \prod_{i=1}^{n} P\left(b_i \left| X_n \right.\right) \quad (2)$$

The Nave Bayes algorithm is a non-linear, complex data classification method that is simple, easy to use, and efficient. This could execute with a Naive Bayes theory so doesn't need Bayesian methods since it is dependent on hypothesis and conditional class independent. Using all 13 attributes of our dataset, an accuracy of 88.54% was reached. Also, in our proposed method we choose the gaussian normal distribution to get more clear prediction.

5.1.1 Adaptive Boosting:

AdaBoost, or Adaptive Boosting, is a machine learning Ensemble Method that can be employed by itself to improve the effectiveness of a variety of different learning methods. The most effective way with AdaBoost is to utilize one-level decision trees or decision trees with only one split. The first step in this approach is to build a model by assigning equal weight to all data points. It therefore elevates the importance of points that have been incorrectly categorized. We have used hyper tuning in AdaBoost method also with base as GuassianNB.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In this research project, the study and evaluation of the optimum classification algorithm were completed, and the outcome is presented below. The health records in the UCI dataset are divided into two groups: train set and test set. And the Data is pre-processed after isolating the dataset, and classification algorithms like as GaussianNB and Ada-Boost are used. The whole research project is done in Google Colaboratory.

6.1 PERFORMANCE METRICS

This section explains the measures that were utilized in the analysis. Furthermore, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) scores are used to find out the performance of the algorithm. The measurement is TP, FP, TN, and FN and it is about as follows:

- TP stands for the number of persons suffering from cardiovascular disease
- TN stands for the total number of people with and without cardiac problems.
- FP stands for the number of people who have never had a heart attack or stroke.
- FN stands for the Total number of people with and without heart disease.

Precision: Precision is a performance statistic for machine learning models that quantifies the accuracy of a model's positive prediction.

$$Precision\ (P) = TP/(TP+FP) \quad (3)$$

Recall: The recall is just a metric of how well our model detects TP. As a consequence, recall gives us Total Percentage of patients accurately classified as having cardiovascular disease out of those that have disease.

$$Recall\ (R) = TP/(TP+FN) \quad (4)$$

F-Measure: F-Measure is a method for combining recall and precision together into one metric which encompasses two features. In Eq. (5) P and R: Precision and Recall.

$$F\text{-}Measure = (2*P*R)/(P+R) \quad (5)$$

ROC Curve: The ROC curve is a binary classifier problem evaluating measure that measures medical specificity and sensitivity for each feasible threshold for testing or even a collection of tests. The Roc curve analyses TPR to FPR across different threshold values, successfully isolating the signal from the noise.

6.2 RESULTS

In our model we get 95% accuracy also with that we get 92% Precision and 89% recall and 96% of F-score. The algorithm appears to have the highest level of precision and recall, which is good because we know that the higher the precision and recall, the finer the output.

Table.1. Performance Measure of Models

Model	Precision	Recall	F-Score	Accuracy
No heart disease	1.00	0.89	0.94	0.95
Heart disease	0.92	1.00	0.96	0.95

In this work, we got great Confusion Matrix *TP* of 34, *TN* of 24, *FP* of 3 and *FN* of 0.

In our project, we got the ROC curve near to the Perfect Classifier. In Fig.6, the best classifier is with 1.0 value and the worst classifier is with value of 0.5. Calculating ROC appears to be more important than precision when dealing with uneven data. Calculating ROC for models that tackle binary classification issues is a good option since it is a highly publicized performance measure that is straightforward to calculate.

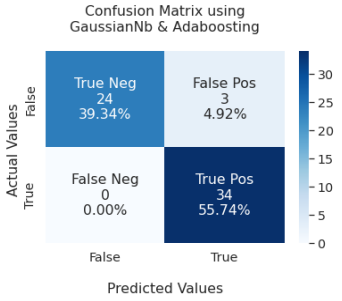

Fig.5. An illustration of Confusion Matrix of AdaBoosting

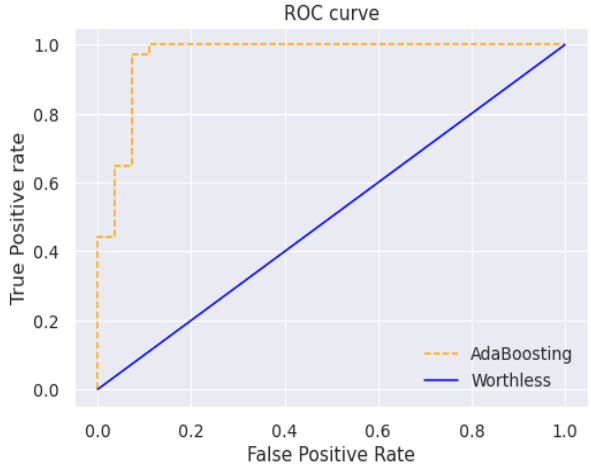

Fig.6. An illustration of ROC curve of AdaBoosting

7. CONCLUSION AND FUTURE WORK

The ML approaches will aid in the long-term preservation of human life and the early detection of heart problems. In this study, ML approaches were utilized to analyze original data and present a new and unique perspective on cardiac illness. Predicting cardiac illness is difficult yet crucial in medicine. If the sickness is detected early and preventative actions are taken as soon as feasible, the fatality rate can be considerably reduced. In this study, we use the UCI dataset and fine-tune machine learning algorithms for heart disease prediction. The Ada-Boost method has been used to improve the result that is generated by the other methods. In the future, further machine learning approaches will be utilized for the best analysis of cardiac diseases and earlier disease prediction, so that the incidence of mortality cases can be lowered through disease awareness also along with that real-time data.

REFERENCES

[1] V. Ramalingam, A. Dandapath and M. Karthik Raja, "Heart Disease Prediction using Machine Learning Techniques: A Survey", *International Journal of Engineering and Technology*, Vol. 7, pp. 684-687, 2018.

[2] World Health Organization, "Global Atlas on Cardiovascular Disease Prevention and Control", Available at https://apps.who.int/iris/handle/10665/44701, Accessed at 2011.

[3] M. Gandhi and S.N. Singh, "Predictions in Heart Disease using Techniques of Data Mining", Proceedings of International Conference on Futuristic Trends on Computational Analysis and Knowledge Management, pp. 25-27, 2015.

[4] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications*, pp. 108-115, 2008.

[5] J. Thomas and R.T. Princy, "Human Heart Disease Prediction System using Data Mining Techniques", *Proceedings of International Conference on Circuit, Power, and Computing Technologies*, pp. 1-13, 2016.

[6] S. Indhumathi and G. Vijaybaskar, "Web-Based Health Care Detection using Naive Bayes Algorithm", *Proceedings of International Conference on Futuristic Trends on Computational Analysis*, pp. 1-12, 2015.

[7] K. Pahwa and R. Kumar, "Prediction of Heart Disease using Hybrid Technique for Selecting Features", *Proceedings of International Conference on Electrical, Computer and Electronics*, pp. 500-504, 2017.

[8] S. Xu, Z. Zhang and T. Zhu, "Cardiovascular Risk Prediction Method based on CFS Subset Evaluation and Random Forest Classification Framework", *Proceedings of International Conference on Big Data Analysis*, pp. 228-232, 2017.

[9] S. Rajathi and G. Radhamani, "Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO", *Proceedings of International Conference on Data Mining and Advanced Computing*, pp. 68-73, 2016.

[10] R. Saini, N. Bindal and P. Bansal P, "Classification of Heart Diseases from ECG Signals using Wavelet Transform and

kNN Classifier", *Proceedings of International Conference on Futuristic Trends on Computing, Communication and Automation*, pp. 1208-1215, 2015.

[11] M.A. Jabbar, B.L. Deekshatulu and P. Chandra P, "Alternating Decision Trees for Early Diagnosis of Heart Disease", *Proceedings of International Conference on Circuits, Communication, Control and Computing*, pp. 322-328, 2014.

[12] Heart Disease Data Set, Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease, Accessed at 2021.

[13] Ping Cao, Bailu Ye, Linghui Yang and Qing Pan, "Preprocessing Unevenly Sampled RR Interval Signals to Enhance Estimation of Heart Rate Deceleration and Acceleration Capacities in Discriminating Chronic Heart Failure Patients from Healthy Controls", *Computational and Mathematical Methods in Medicine*, Vol. 2020, pp. 1-14, 2020.

[14] H. A. Esfahani and M. Ghazanfari, ''Cardiovascular Disease Detection using a New Ensemble Classifier", *Proceedings of International Conference on Futuristic Trends on Computational Analysis and Knowledge Management*, pp. 1011-1014, 2014.

[15] T. Vivekanandan and N.C.S.N. Iyengar, "Optimal Feature Selection using a Modified Differential Evolution Algorithm and its Effectiveness for Prediction of Heart Disease", *Computers in Biology and Medicine*, Vol. 90, pp. 125-136, 2017.

[16] M. Sai Shekhar, Y. Mani Chand and L. Mary Gladence, "Heart Disease Prediction using Machine Learning", *Lecture Notes in Electrical Engineering*, Vol. 708, No. 11, pp. 603-609, 2021.