International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

PREDICTION OF HEART DISEASE USING MACHINE LEARNING

Sonal Shakya,Prof. Jayesh Jain
Global Nature Care Sangathans Group of Institutions

Abstract

Predicting heart disease is one of the most challenging challenges in the medical industry today. Approximately one person dies from heart disease every minute in the modern era. Processing a vast amount of data in the healthcare industry requires data science. Since predicting cardiac illness is a complex undertaking, it is necessary to automate the process in order to reduce risks and warn the patient well in advance. The heart disease dataset from the UCI machine learning repository is used in this study. The suggested work uses various data mining algorithms, including DecisionTree, Knn, Logistic Regression, and Random Forest, to forecast the likelihood of heart disease and categorize patient risk. In order to compare the effectiveness of various machine learning algorithms, this paper will do so. The trial results show that, when compared to other ML algorithms used, the Random Forest approach has the highest accuracy (90.16%).

Keywords: Decision Tree, KNN, Logistic Regression, Random Forest, Heart Disease Prediction.

Introduction

The research that is suggested in this study focuses mostly on different data mining techniques used to forecast cardiac disease.The main organ of the human body is the heart. In essence, it controls the blood flow throughout our body. Any heart abnormality can be distressing to other body parts. Any type of impairment to the heart's regular operation might be categorized as a heart disease. In today's modern society, heart disease is one of the main causes of most fatalities. Heart disease can be brought on by living a sedentary lifestyle, using tobacco products, drinking alcohol, and eating a lot of fat [2]. The World Health Organization estimates that more than 10 million people worldwide pass away each year as a result of heart disease. The only means of stopping heart-related ailments are a healthy lifestyle and early detection. The provision of high-quality services and effective, accurate diagnosis is the main problem in today's healthcare [1]. Even though cardiac disorders have been identified as the leading cause of death worldwide in recent years, they are also the ones that can be effectively controlled and managed. The right moment of disease discovery determines how accurately a disease will be managed overall. The suggested approach aims to identify these heart conditions early in order to prevent negative outcomes.

Records of a sizable collection of medical data compiled by medical professionals are available for analysis and knowledge extraction. The use of data mining techniques allows for the extraction of important and hidden information from a vast volume of data. The medical database primarily contains discrete data. As a result, making decisions with discrete data is a challenging and complex undertaking. Data mining's subset of machine learning (ML) effectively manages massive, well-organized datasets. Machine learning may be used to diagnose, detect, and forecast a variety of disorders in the medical industry. The main objective of this study is to give clinicians a tool to identify cardiac disease at an early stage [5]. As a result, patients will receive effective care and serious repercussions will be avoided. In order to uncover the underlying discrete patterns and analyze the provided data, machine learning (ML) plays a critical role. After data analysis, machine learning approaches aid in the early detection and prediction of cardiac disease. In order to predict cardiac disease at an early stage, this work analyzes the performance of several ML approaches, including NaiveBayes, Decision Tree, Logistic Regression, and Random Forest [3].

Related work

Using the UCI Machine Learning dataset, extensive research has been done to predict cardiac disease. varied data mining approaches have been used to achieve varied accuracy levels, which are detailed below.

In comparison to the needs of today, certain articles that were published two to three years ago had a lower accuracy for the prognosis of heart disease. Sharma et al.'s "Efficient heart disease prediction system using decision tree"It was released in that year.using a decision tree classifier, they achieved an accuracy of 75%.

A system [2] that combined the MapReduce algorithm and data mining techniques has been suggested by T. Nagamani, et al.For the 45 instances in the testing set, the accuracy acquired using this paper's method was better than the accuracy obtained using a traditional fuzzy artificial neural network. Here, the usage of dynamic schema and linear scaling increased the algorithm's accuracy.

Using NB (Naive Bayesian) approaches, Anjan Nikhil Repaka, et al., devised a system in [4] for the classification of dataset and the AES (Advanced Encryption Standard) encryption method are used to transport data securely for disease prediction.

Theresa Princy, R., et al. carried out a survey that included various classification algorithms for heart disease prediction. The accuracy of the classifiers was examined for various numbers of characteristics using Naive Bayes, KNN (K-Nearest Neighbour), Decision trees, and Neural Networks as the classification algorithms [5].

Heart disease classification using machine learning algorithms is a topic of research for Avinash Golande and colleagues. An investigation was conducted to examine the accuracy of the classification algorithms Decision Tree, KNN, and K-Means[1]. The study found that Decision Trees had the highest accuracy, and it was concluded that by combining various methodologies and fine-tuning its parameters, it might be made more effective.

After reading the aforementioned publications, the fundamental idea behind the suggested system was to build a heart disease prediction system using the inputs presented in Table 1. Based on the accuracy, precision, recall, and f-measure scores of the classification algorithms Decision Tree, Random Forest, Logistic Regression, and Knn, we determined the best classification algorithm that may be applied to the prediction of heart disease.

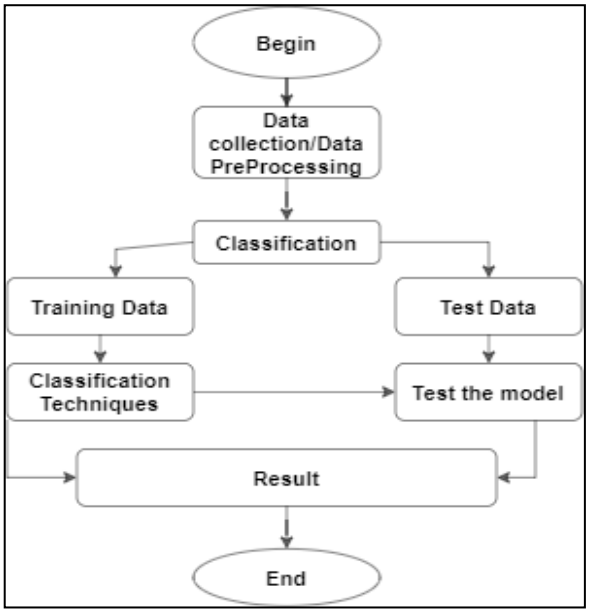


Fig. 1: Generic Model Predicting Heart Disease

A. Data Collection and Preprocessing

The dataset used was the Heart disease Dataset which is a combination of 4 different databases, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features [9]. Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis. The complete description of the 14 attributes used in the proposed work is mentioned in Table 1 shown below.

TABLE I. FEATURES SELECTED FROM DATASET

Sl. No.	Attribute Description	Distinct Values of Attribute
1.	Age- represent the age of a person	Multiple values between 29 & 71
2.	Sex- describe the gender of person (0- Female, 1-Male)	0,1
3.	CP- represents the severity of chest pain patient is suffering.	0,1,2,3
4.	Rest BP-It represents the patient's BP.	Multiple values between 94& 200
5.	Chol-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	FBS-It represent the fasting blood sugar in the patient.	0,1
7.	Resting ECG-It shows the result of ECG	0,1,2
8.	Heartbeat- shows the max heart beat of patient	Multiple values from 71 to 202
9.	Exang- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	Old Peak- describes patient's depression level.	Multiple values between 0 to 6.2
11.	Slope- describes patient condition during peak exercise. It is dividedinto three segments(Unsloping, Flat, Down sloping)	1,2,3
12.	CA- Result of fluoroscopy.	0,1,2,3
13.	Thal- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which representThallium test.	0,1,2,3
14.	Target-It is the final column of the dataset. It is class or label Colum.It represents the number of classes in dataset. This dataset has binaryclassification i.e. two classes (0,1).In class "0" represent there is lesspossibility of heart disease whereas "1" represent high chances of heart disease. The value	0,1

Proposed model

The proposed work predicts heart disease by exploring the above mentioned four classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease. Fig. 1 shows the entire process involved corresponding branch is followed to that value and jump is made to the next node.

Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic

regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables which makes logistic regression good for classification.

Result and Analysis

A. Classification

The attributes mentioned in Table 1 are provided as input to the different ML algorithms such as Random Forest, Decision Tree, Knn, Logistic Regression and The input dataset is split into 70% of the training dataset and the remaining 30% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analyzed based on different metrics used such as accuracy, precision, and recall and F-measure scores as described further. The different algorithms explored in this paper are listed as below.

a. Random Forest

Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

b. Decision Tree

Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to record's attribute. On the result of comparison.

The results obtained by applying Random Forest, Decision Tree, Naive Bayes and Logistic Regression are shown in this section. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure. Precision (mentioned in equation (2)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (3)] defines the measure of actual positives that are correct. F-measure [mentioned inequation (4)] tests accuracy.

Precision = (TP) / (TP +FP)(2) Recall = (TP) / (TP+FN) (3)

F– Measure =(2 * Precision * Recall) / (Precision +Recall) (4)

• TP True positive: the patient has the disease and the test is positive.

• FP False positive: the patient does not have the disease but the test is positive.

• TN True negative: the patient does not have the disease and the test is negative.

• FN False negative: the patient has the disease but the test is negative.

In the experiment the pre-processed dataset is used to carry out the experiments and the above-mentioned algorithms are explored and applied. The above-mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms is shown below in Table 2. The accuracy score obtained for Random Forest, Decision Tree, and Logistic Regression and is shown below in Table 3.

TABLE II.VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHM

Algorithm	**True Positive**	**False Positive**	**False Negative**	**True Negative**
Logistic Regression	22	5	4	30
Random Forest	22	5	6	28
Decision Tree	25	2	4	30

c. Knn

It is one of the simplest supervised classification algorithms and its most used algorithms. It is used to solve both classification and regression problems it's easy to understand and implement.

TABLE III. ANALYSIS OF MACHINE LEARNING ALGORITHM

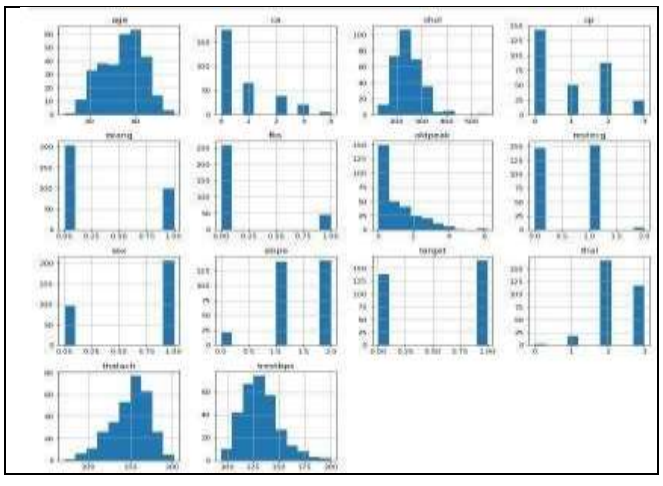
Algorithm	Precision	Recall	F-measure	Accuracy
Decision Tree	0.845	0.823	0.835	81.97%
Logistic Regression	0.857	0.882	0.869	85.25%
Random Forest	0.937	0.882	0.909	90.16%

Conclusion

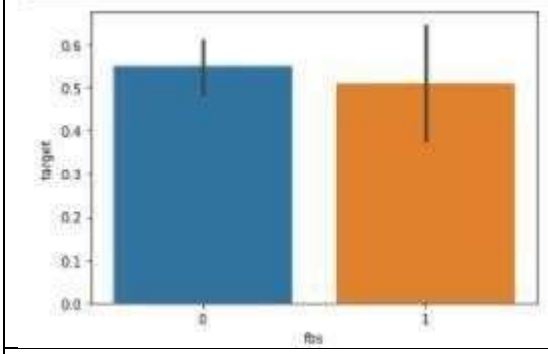
With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Knn for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Logistic Regression is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Logistic Regression as well as using a larger dataset as compared to the one used in this analysis which will help to Introduction World Health Organization has estimated 12 million deaths occurred worldwide every year due to heart disease. Half deaths in India and other developed countries due to cardiovascular diseases can aid in making decision on lifestyles changes in high-risk patients and in turn reduce the complications.

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Knn for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Logistic Regression is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Logistic Regression as well as using a larger dataset as compared to the one used in this analysis which will help to

Introduction World Health Organization has estimated 12 million deaths occurred worldwide every year due to heart disease. Half deaths in India and other developed countries dueto cardiovascular diseases can aid in making decision on lifestyles changes in high-risk patients and in turn reduce the complications.



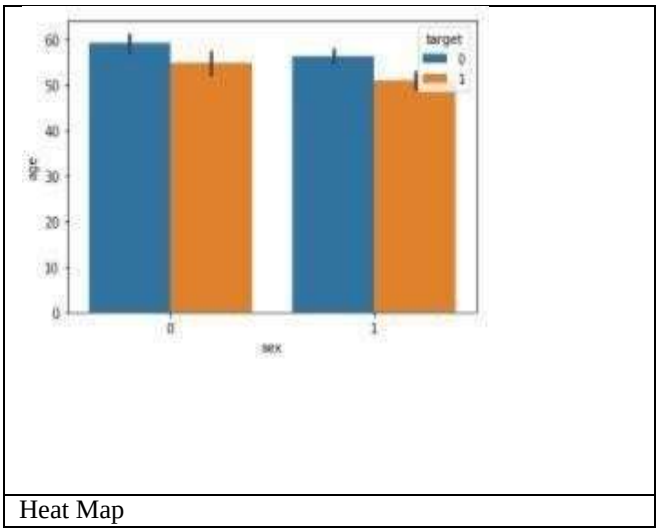
Hist That shows Features and Levels Of Persons Have Heart



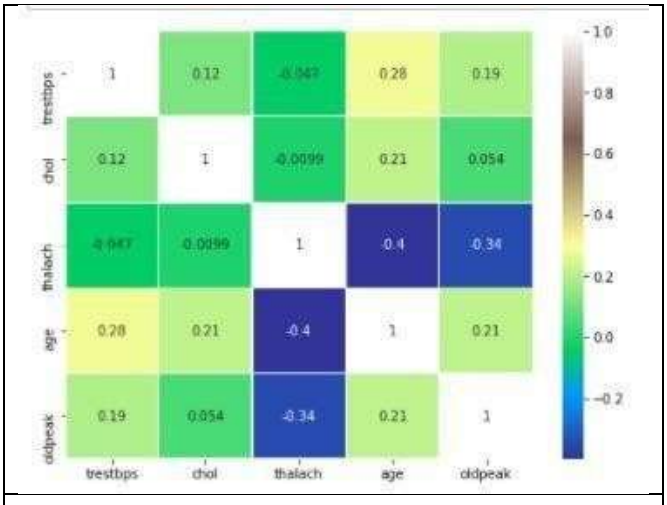
CHECK MALE AND FEMALE PATIENTS WITH MORE RATE

0 NO HEART DISEASE FOR WOMEN WITH 58% AND 48% WOMEN HAVE HEART DISEASE

0 NO HEART DISEASE FOR MEN WITH 55% AND 45% MEN HAVE HEART DISEASE

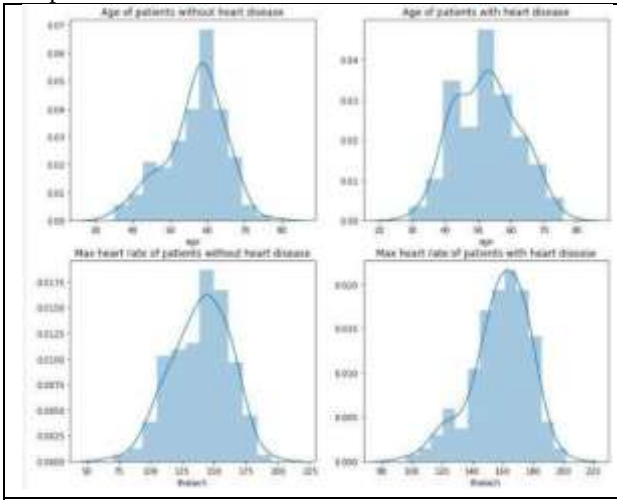


Heat Map

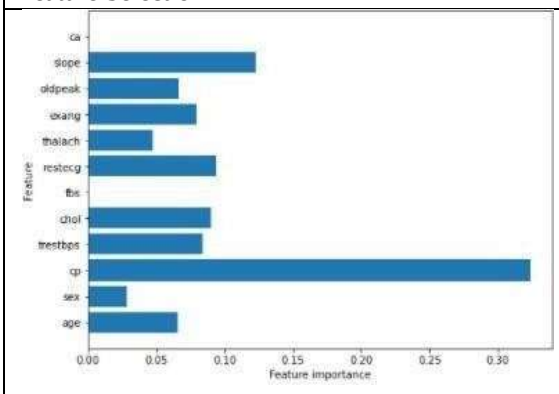


With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient MLalgorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction ofheart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis.

Displots that shows Age of patients with heart disease and no of patients who don't have heart disease.

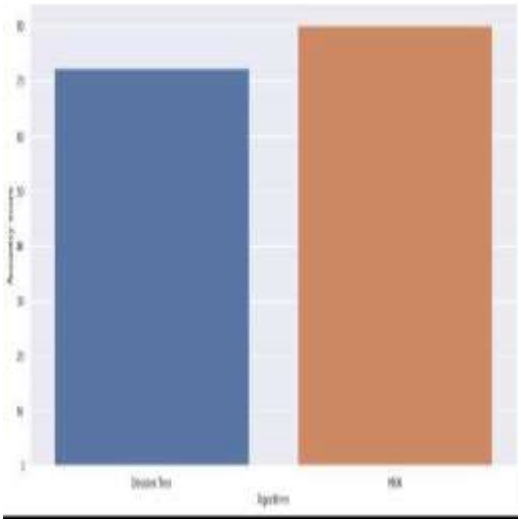


Feature Selection

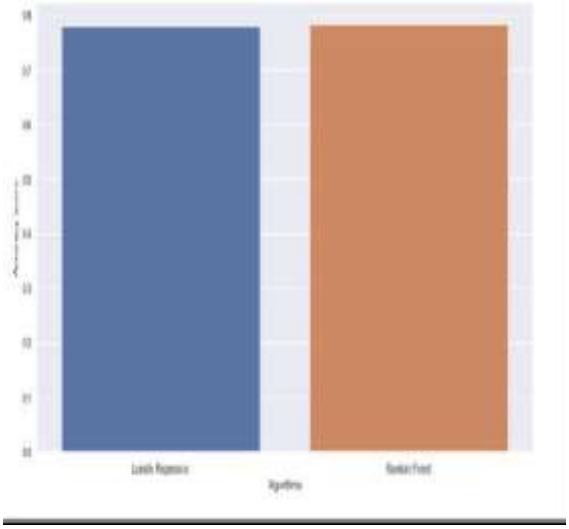


Accuracy Level Of Future Selection

Final Result Of Accuracy Compared between Decision Tree and Knn



Final Result Of Accuracy Compared between Decision Tree and Knn and logistic regression



EXISTING SYSTEM

Heart Disease is even highlighted as a silent killer which leads to death of the person without oblivious symptoms. The before all existing system works on sets of both Deep learning and Data Mining. Medical Diagnosis plays a vital role and yet complicated task that needs to be executed efficiently and accurately. To reduce the cost for achieving clinical tests appropriate computed based information and decision support should be aided.. Data Mining is the use of software techniques for finding patterns and consistency insets of data .also with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes of classes Learning of the risk components connected with

heart disease helps medicinal services experts to recognize patients at high risk of having heart disease statistical analysis has identified risk factors associated with heart disease.

PROPOSED SYSTEM

This system's data can be used to determine whether or not a patient has cardiac disease based on certain characteristics. This system's suggestion could make use of the data to develop a model that tries to forecast whether a patient has an illness or not. Using a Decision Tree, Random Forest, Logistic Regression, and Knn algorithm, the proposed system by calculating the grade with the aid of a skit library. Using random combinations of the hyperparameters, the best build model solution can be found using the random search technique. finally using model comparison to analyze the outcomes. Based on the characteristics of the patient's heart, the data we have should be divided into several structured data categories. from the data's accessibility. Using a Decision Tree, Random Forest, Logistic Regression, and Knn algorithm, we must develop a model that anticipates the patient's sickness. Importing datasets is the first step.Read through the datasets that should include several factors such as age, gender, sex, chest discomfort, and slope target. To verify the facts, the data should be examined. Build a decision tree, random forest, logistic regression, and Knn algorithm model while simultaneously creating a temporary variable. So, to aid in the graphical representation of the classified data, we apply a sigmoid function. In comparison to earlier work done in the present system, accuracy is improved by employing Random Forest and compared to other methods.

References

[1] Avinash Golande, Pavan Kumar T, ”Heart Disease Prediction Using Effective Machine Learning Techniques”, International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T.Nagamani, S. Logeswari, B. Gomathy,” Heart Disease Prediction using Data Mining with Mapreduce Algorithm”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi,” Implementation of Machine Learning Model to Predict Heart Failure Disease”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, ”Design And Implementation Heart Disease Prediction Using Naives Bayesian”, International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa Princy R,J. Thomas, ‘Human heart Disease Prediction System using Data Mining Techniques’, International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[6] Nagaraj M Lutimath, Chethan C. Basavaraj S Pol., ’Prediction Of Heart Disease using Machine Learning’, International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[7] UCI, ―Heart Disease Data Set.[Online]. Available (Accessed on May1 2020): https://www.kaggle.com/ronitf/heart-disease-uci.

[8] Sayali Ambekar, Rashmi Phalnikar, “Disease Risk Prediction by Using Convolutional Neural Network”,2018 Fourth International Conference on Computing Communication Control and Automation.

[9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[10] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics:

Conference Series, 2018

[11] Fajr Ibrahem Alarsan., and Mamoon Younes ‘Analysis and classification of heart diseases using heartbeat features and machine learning algorithms’, Journal Of Big Data,2019;6:81.

[12] Internet source [Online].Available (Accessed on May 1 2020): http://acadpubl.eu/ap