Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 7
July 2023

Genetic Algorithm Based Feature Selection to Optimize Cardiovascular Disease Prediction Using Classification Techniques

Dr. M. Rameshkumar[1], N. Jagadeesan[2] and T. Velmurugan[3]

[1]Asst. Professor & Head, [2]Assistant Professor and [3]Associate Professor,

[1]PG and Research Department of Computer Science, Government Arts College, Chennai – 600035, India

[2]PG Department of Information Technology & BCA, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai - 600106, India.

[3]PG and Research Department of Computer Science, Dwaraka Doss Goverdhan Doss Vaishnav College, Chennai - 600106, India.

Mail: proframeshkumar@gmail.co[1]; jaga.dgvit@gmail.com[2]; velmurugan_dgvc@yahoo.co.in [3]

Abstract: Background: A correct assessment of cardiac illness has the potential to save a person's life, while a poor prognosis can be fatal. Still, there are cases that are diagnosed wrongly, prognoses, then cured. People are encouraged to participate in diverse diagnostic examinations. Such diagnostics are frequently ineffective at detecting problems at this time. The purpose of the work is to find a way that effectively anticipates the incidence of cardiovascular problems using fewer variables while sparing patients' time and money spent on diagnostic procedures. The dataset, which included 14 parameters, were gathered from digital sources for this research Work. A patient undergoes fewer tests by using the genetic algorithm to identify the qualities that are most helpful in the diagnosis of heart issues. A genetic search decreased 14 traits to 6 attributes. In order to predict disease with more accuracy than before the reduction of data, four classification techniques Random Forest (RF), AdaBoost, k-Nearest Neighbor (k-NN), and Support Vector Machine (SVM) applied. The accuracy values of the Random Forest system, the AdaBoost Method, the k-NN template, the SVM framework, and the Model of Random Forest were 92.90%, 89.30%, 90.10%, and 93.70%, respectively. Following the application of the genetic algorithm, the Random Forest model's accuracy was 96.90%, followed by the AdaBoost method's accuracy of 92.30%, the k-NN algorithm's accuracy of 94.10, and the SVM model's accuracy of 98.90%. In comparison to the Random Forest (RF), AdaBoost, and k-NN classifiers, after attribute selection using genetic algorithms, the SVM classifier delivers excellent results and high metric values for heart disease prediction.

Keywords: Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Random Forest (RF) Method, AdaBoost Algorithm, Genetic Algorithm, Feature Selection Method.

1. Introduction

One of the most serious diseases in the world is coronary heart disease (CHD). The World Health Organization (WHO) identified coronary heart disease (CHD) as the top global cause of death each year. By 2030, CHD is expected to affect about 23.6 million people, according to the WHO. One in every four deaths occurs in some industrialized countries, such as the United States of America (USA) [1]. Even worse percentages exist in the Middle East and North Africa, where 39.2% of all deaths occur. As a result, reducing the number of deaths brought on by coronary heart disease requires early and accurate investigation as well as the availability of appropriate treatments. People with a high risk of developing heart disease must have access to these services [2]. CHD is becoming the leading cause of death. It affects all age groups of people [3]. Blood Pressure (BP), excessive cholesterol [4, 5], and palpitations [6] are the main causes of cardiovascular disease. Heart disease is caused by some non-modifiable variables, such as drinking alcohol and smoking. Other body organs harmed by a heart condition. Risk factors for cardiovascular disease include cholesterol, high blood pressure, age, family history, smoking, and poor diet. As a result, minimizing the incidence of coronary heart disease-related fatalities necessitates both early and accurate inquiry and the accessibility of

effective treatments. These services must be available to people who are at high risk of developing heart disease [2]. All age groups are affected by CHD, which is an increasingly major cause of death [3]. The main contributors to cardiovascular illness include high blood pressure (BP), too much cholesterol [4, 5], and palpitations [6]. Some unchangeable factors, including consuming alcohol and smoking, can also lead to heart disease. A heart problem impacts other bodily organs. Cholesterol, high blood pressure, age, family history, smoking, and poor diet are all cardiovascular disease risk factors.

A disturbance in the heart's rhythm is called an arrhythmia. Heartbeats can be regular, erratic, or both. Unusual heartbeats result from an electrical circuit short in the heart. Congenital heart disease, sometimes referred to as congenital heart abnormality, is a birth defect affecting the heart's or main vessels' structural integrity. The heart's inability to pump blood to the body's tissues is known as congestive heart failure. It was a typical term for cardiac failure. Cardiomyopathy is characterized as a change in the muscle's structure or a fatigue of the heart muscle, it may result from inadequate cardiac pumping. Risk factors for cardiomyopathy include high blood pressure, alcohol use, viral infections, and inherited diseases. Patients often refer to coronary artery disease as heart blockage, which caused by a buildup of plaques that narrows the arteries that feed blood to the heart muscle. While increased blood flow needed, such when exercising, a blockage that is severe enough to prevent the muscle from receiving the blood it needs to function may occur. This causes symptoms like shortness of breath and chest pain. Nuclear scan, like those used in exercise stress testing, are used to look for any area of the heart where the blood supply has been compromised. These tests, however, have limitations, particularly in patients who are deemed to be at high risk. The treatment, known as coronary angiography, allows you to see the shape of the real vessel in order to examine the heart blockage first hand. One to the right side and two on the left appear to be the blood vessels that run over the surface of the heart and supply it with blood. Right coronary artery refers to the one on the right. The front and major walls on the main side of the heart are supplied by the left anterior descending (LAD), which descends the front of the heart, and the sidewall is supplied by the left circumflex. If you look closely, you will see that a significant artery, known as the left main artery, supplies both the LAD and the circumflex. Mild heart obstructions was defined as less than 40%. These obstructions clearly do not restrict blood flow, hence they are not likely to cause symptoms. The progression of coronary heart disease is, however, amply demonstrated here, and patients like them need a strong focus on risk factors for the disease (such as cholesterol, diabetes, smoking, blood pressure, and so forth), appropriate medications, and better and healthier lifestyle changes like exercise, fat loss, and adherence to a healthy diet.

The typical range for mild heart obstruction is 40% to 70%. Similar to mild conditions, mild coronary artery disease focuses on risk factors, medication, and healthy lifestyle adjustment. To determine whether a heart blockage at the higher end of the medium range (50–70%) is significant and entirely responsible for symptoms, further testing may be necessary. Significantly, greater than 70% is typically used to denote serious cardiac blockage. Shortness of breath and chest pain are two symptoms that can result from this level of cardiac muscle shrinkage coupled with a drastically reduced blood flow. The significant cardiac obstruction produced symptoms that made it possible to treat with stent placement. Multiple significant blockages may necessitate bypass surgery (bypass graft). Complete blockage of a coronary artery (100%) stops all blood flow and, naturally, causes a heart attack. Heart blockages typically have significant symptoms, so it is important to get treatment as soon as you can. As can be seen in the image, a stent was inserted to treat this artery and restore regular blood flow. The cardiac muscle may perish if they are not considered (often within a few minutes, the quicker the better). Once dead, there will be little chance of survival, which will cause heart failure and a reduction in the heart's ability to pump blood.

Data mining is the technique of obtaining information from a sizable database. As the amount of actual data grows, it is becoming

increasingly important. Data preprocessing and data post processing are the first two transformational processes in the data mining process, which is a crucial part of KDD. Classification, association, and clustering are the core operations of data mining. The problem of classification is widespread and has many applications. Increased public awareness of health issues and technical advancements has led to an expansion in hospitals and medical facilities. However, many developing nations still struggle with the provision of affordable, high-quality healthcare services. Although many countries have made serious efforts to offer medical care, it is still unclear whether these services will reach the poor and the needy. In a number of medical specialties, data mining techniques have been utilized to enhance medical decision-making [6] [7]. Numerous medical facilities struggle to offer high-quality services, such as accurate patient diagnosis and treatment at reasonable cost.

Cardiovascular disease is frequently diagnosed using a recommended medical test as well as the patient's signs and symptoms. However, an effective treatment may be achievable if the condition is quickly and easily identified. The course of treatment for each of these heart patients will be determined by the results of testing, clinical documentation, and patient questionnaires [8]. In addition to delaying diagnostic procedures, all of these techniques history analysis, physical examination research, and medical expert evaluation often lead to inaccurate diagnosis and mechanical failure. It is also more expensive and computationally demanding, and analyses take a long time to complete [9]. In the past, researchers were considerably more focused on selecting pertinent qualities to use in their heart disease prediction model [10]. The objective of this study is to apply a Genetic Algorithm (GA) to find linked characteristics or features in a dataset with heart disease. Following feature selection, classification algorithms are employed to predict cardiac disease. Comparing the prediction value to the outcomes of various categorization techniques.

The following sections make up the remainder of this essay. A brief summary of related research for the diagnosis of heart disease conducted by various researchers is shown in Section 2 of this article. The materials and procedures are described in Section 3. The Results and Discussion are explained in Section 4. The conclusions are enumerated in Section 5.

2. Literature Review

Using datasets on coronary artery disease, many researchers are working hard to identify various models for the prediction of cardiovascular illness using machine learning techniques including clustering, classification, regression, and more [11] [12]. The bulk of the key characteristics and methods outlined below, together with each one's unique drawbacks and advantages, set our work apart. Overviews of numerous methodologies and review articles on machine learning methods for cardiovascular disease prediction may be found in this area.

B. S. S. Rathnayake et al. discovered a study. [13] The article "Heart Disease Prediction using Data Mining and Neural Network Techniques" describes how to predict heart disease using ECG signals and the arrhythmia dataset. The heart rate time series data that was taken into consideration for this study was successfully categorised using a neural network method. The radial basis NN is used to identify both linear and nonlinear characteristics. In the study paper titled "Predictions of heart disease using techniques of data mining," M. Gandhi et al. [14] proposed the model, which uses decision trees, neural networks, and Naive Bayes classifier to predict heart disease. But the study's exclusive focus was on classification models for data mining. Model for predicting heart disease using binary classification. In their study titled "Prediction of cardiac disease using multilayer perceptron neural network," J. S. Sonawane et al., who proposed the neural network multilayer perceptron [15], said. In which, multilayer learning was used to solve the complex problem. The algorithm's disease prediction performance was accurate when tested on the Cleveland dataset. In their study titled "Effective diagnosis of heart disease by neural networks ensembles" [16], researchers Das et al. developed a method to predict heart disease using a neural network-based ensemble model. This study suggests a strategy that combines posterior

probabilities from various predetermined models. High ensemble model accuracy was attained in this work.

In their study work titled "Prediction of Heart Disease Using Random Forest and Rough Set-Based Feature Selection," researchers predict heart disease using the Random Forest algorithm [17]. Yekkala et al., in which a roughest-based method is employed to carry out feature selection. The stat log dataset was used for the task. P.k.anooj projected the risk level and the fuzzy rule predicted the diagnosis of heart disease [18]. The process of anticipating risk is divided into two stages: the first involves creating weights using fuzzy rules, and the second involves creating decisions that are based on rules. Through the use of a fuzzy judgement, this procedure improves the prediction system. Three different datasets are used to evaluate the work: the Cleveland, Hungarian, and Swiss datasets. Using an ensemble model, researcher H. A. Esfahani et al. predicted the cardiovascular disease [19]. Feature selection for the ensemble technique was based on identifying dependencies between features and class values. In this study, the ensemble method outperformed traditional classification techniques in terms of accuracy. Support vector machine (SVM) and artificial neural network-based decision support system model for cardiovascular illness categorization was proposed by Mrudula Gudadhe et al (ANN). A decision support system that recognises cardiovascular illness is developed using a multilayer perceptron neural network (MLPNN) with three layers. The back-propagation technique is used to train the multilayer perceptron neural network, which is a computationally efficient method [20].

A scalable method for forecasting the signs of heart disease was developed and validated by Rashmi G. Saboji and colleagues. After implementing the random forest algorithm on the Spark framework to predict heart disease with as few as 600 dataset entries, they aim to look into additional healthcare ailment predictions, such as early prediction of specific types of cancer, etc. They also aim to investigate how running large trained datasets on high-performance clusters affects accuracy and performance [21]. Princy According to Theresa J. Thomas (2016), each person's risk level is determined
using a variety of classification algorithms based on their age, gender, blood pressure, cholesterol, and pulse rate. The patient's data is classified using data mining techniques such as Naive Bayes, KNN, Decision Tree Algorithm, and Neural Network. When a larger number of attributes are used, the accuracy of the risk level increases. According to the analysis model, many writers employ diverse technologies and a variety of attributes for their research. As a result, various technologies provide varying degrees of precision based on a variety of factors. The risk of heart disease was discovered using the KNN and ID3 algorithms, and the accuracy level was supplied for a variety of variables. Other techniques might be used in the future to minimize the number of features while increasing accuracy [22].

Based on clinical data from patients related to the risk factors for heart disease, Madhura Patil et al. recommended the creation of a system that will help in the prediction of heart disease. By examining medical variables including age, gender, blood pressure, overweight, and blood sugar and using an SVM classifier, we can determine whether a patient would develop heart disease. The SVM has also shows great classification accuracy, sensitivity, and specificity, making it a better choice for diagnosis. They are also doing data analysis to determine what age sickness most frequently strikes and what areas are affected. In order to prevent heart disease-related mortality, vigilance must be taken [23]. Mandavkar, Snehal Subhash, and others. In their study titled "Heart disease prediction system using classification and genetic algorithm," we discussed properties established by the genetic algorithm were employed for machine learning classification, improving the efficacy of the suggested prediction system. The Cleveland dataset, which was made available via the UCI repository, was used in the investigation. The dataset used for the study has 303 instances, 14 characteristics, and five class types. From 0 for normal to 4 for stroke, the class values are available.The experimental results showed that the Genetic Algorithm based feature selection and classification bring the best results on multiple disease threat prediction. The application is designed in which patients will be

able to predict and discuss with practitioners through the designed web application [24].

The relief method is the best feature selection approach, and the support vector machine algorithm with the linear kernel is the best machine learning algorithm, according to the research by Takci et al. [25]. The accuracy value for this combination was the highest at 84.81%. According to N. Satish Chandra Reddy et al researchers based on three distinct percentage splits, the random forest algorithm's accuracy in feature selection and classification was found to be between 90 and 95 percent. The 8 and 6 features that were chosen appear to be the minimum needed to create a more accurate performance model [26]. Researchers T Velmurugan et al. mined the significance of clinical reports and other applications in their research efforts by using the Named Entity Recognition (NER) method to find the corresponding phrases for the cardiac sickness content [27]. The classification algorithms J48, CART, and ADTree were discovered by the researchers B Padmapriya et al. [28] to analyse the data related to breast cancer. In order to forecast Parkinson's disease, researchers J Dhinakaran et al. [29] conducted a comparison study on several classification models, including Naive Bayes, k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), and Random Forest. Regarding the precision through performance, the Random Forest method does a good job at classifying the data. Farzaneh Sadeghian et al. found that the genetic algorithm and the k-NN classification technique were employed to improve the identification of two sets of autistic and healthy individuals [30]. To improve the accuracy of heart disease prediction, Chandra Babu Gokulnath et al. used a genetic algorithm in conjunction with a classification technique called Support Vector Machine [31]. According to the research by M. ANBARASI et al. [48], a genetic algorithm is used to identify the characteristics that are most helpful in the diagnosis of heart conditions, hence lowering the number of tests that a patient must do. Using genetic search, 13 attributes are condensed to 6 attributes. The diagnosis of patients is then predicted using three classifiers—Naive Bayes, Classification by Clustering, and Decision Tree—with the same accuracy as before, the reduction in the number of characteristics.

3. Materials and Method

Many scholars have examined heart disease data using classification algorithms to classify datasets with high accuracy and efficiency of learning algorithms using straightforward approaches [32]. In order to accurately forecast cardiac illness, medical analysts or practitioners may find this paper's analysis of the Random Forest (RF), AdaBoost, k-Nearest Neighbor (k-NN), and Support Vector Machine (SVM) classification approaches useful. The aforementioned algorithms are taken into consideration for this research work after examining recent publications, journals, and reviews in the fields of computer science and engineering, data mining, and cardiovascular disease. This is because, when compared to other classification algorithms, those algorithms have performed well and yield better results to satisfactorily prediction of the heart disease.

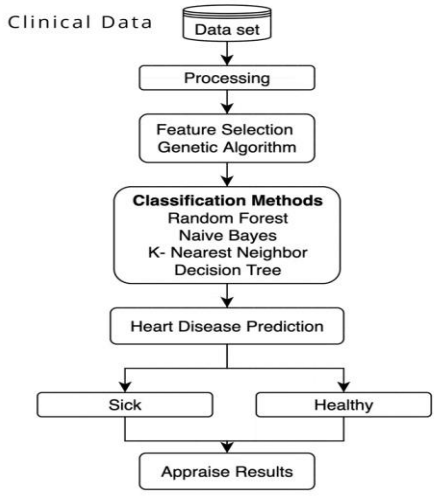


Figure 1: Architecture of the proposed work

There are several steps in the current research work to reach the goals. Figure 1 depicts the entire organizational structure of this research work clearly. The steps involved in this procedure are as follows: The first step is dataset collection, the heart disease dataset from the Kaggle database selected specifically for this research. Then, the dataset is fed into the subsequent classification algorithms. The output findings from Random Forest (RF), AdaBoost, k-Nearest Neighbor (k-NN), and Support Vector Machine (SVM) are noted, and

the same dataset is used as input for a genetic algorithm to choose key characteristics to improve the prediction accuracy. The subsets submitted as input to the same classification algorithm after the important features have been chosen. The output results contrasted with the previous findings.

3.1. Description of Dataset

The heart disease dataset for this research work was carried out by using Kaagle depository. This actual dataset contains 76 attributes, but only 14 attributes [33] have taken into consideration for this research work. These attributes play essential role in the prediction of heart disease. The remaining attributes such as patient name, address, mobile number, etc., omitted because of those attributes will not be helpful for this research. The "Target" field refers to patients' the existence of cardiac disease. It has a numeric value ranging from 0 (no presence) to 1 (presence). To maintain secrecy on patient details, the names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

The heart disease dataset used in this study has 1024 occurrences with no missing variables. This study aims to forecast cardiac disease regardless of the kind of illness. A detailed description of the dataset is provided in Table 1. The label of output feature (num) is divided into two classes to denote the presence of heart disease and absence of heart disease. These data records were created in Excel data sheet, saved in the format of .CSV. These data are processed in this research. It has taken all attributes listed in Table 1. The dataset consists 5 quantitative (numerical) attributes which are age, trestbps, chol, thalach, oldpeak, and 9 qualitative (nominal) categorical attributes which are sex, cp, fbs, restecg, exang, slope, ca, thal and target. ca is considered a qualitative categorical attribute and it only consists of 4 types of unique values.

Table 1: Description of heart disease dataset

Attribute ID	Attribute Name	Attribute Type	Description
A1	AGE	Numerical	Patient's Age in years
A2	SEX	Nominal	Patient's Gender (1 = Male, 0 = Female)
A3	CP	Nominal	Chest Pain Type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
A4	TRESTBPS	Numerical	Resting Blood Pressure (in mmHg on admission to the hospital)
A5	CHOL	Numerical	Serum Cholesterol Level (in mg/dl)
A6	FBS	Nominal	Patient's Fasting Blood Sugar >120mg/dl (1= true, 0 = false)
A7	RESTECG	Nominal	Resting Electrocardiographic Results (2 = showing probable or definite left ventricular hypertrophy by Estes' criteria, month of exercise ECG reading, 1 = ST wave abnormality, 0 = normal)
A8	THALACH	Numerical	Maximum Heart Rate Achieved by the Patient
A9	EXANG	Nominal	Exercise-induced Angina (1= Yes, 0 = No)
A10	OLDPEAK	Numerical	ST Depression Induced by Exercise Relative to Rest
A11	SLOPE	Nominal	The Slope of the Peak Exercise ST Segment (3 = downsloping, 2 = flat, 1 = upsloping)
A12	CA	Nominal	Number of Major Vessels Blocked

			(0 to 3)
A13	THAL	Nominal	The Heart Status (7 = reversible defect, 6 = fixed defect, 3 = normal)
A14	TARGET	Nominal	Diagnosis of heart disease (1 = presence, 0 = absence)

The heart disease dataset used in this study has 1024 occurrences with no missing variables. This study aims to forecast cardiac disease regardless of the kind of illness. A detailed description of the dataset is provided in Table 1. The label of output feature (num) is divided into two classes to denote the presence of heart disease and absence of heart disease. These data records were created in Excel data sheet, saved in the format of .CSV, and processed in this research work. For the study,. It has taken all attributes listed in Table 1. The dataset consists 5 quantitative (numerical) attributes which are age, trestbps, chol, thalach, oldpeak, and 9 qualitative (nominal) categorical attributes which are sex, cp, fbs, restecg, exang, slope, ca, thal and target. ca is considered a qualitative categorical attribute and it only consists of 4 types of unique values.

3.2. Preprocessing

Data preprocessing is one of the major steps in building machine learning algorithm, as only a quality and clean dataset produces a quality model that gives quality results. Data preprocessing transforms the raw data to useful and efficient format for human and machine learning process. Data reduction, transformation, and cleaning are all components of data preparation [34]. Data transformation involves data discretization, concept hierarchy, normalisation, and standardisation; data cleaning includes handling missing values and noise. Data reduction includes both dimensionality and numerosity reduction. The genetic algorithm is employed in this study to choose traits and specific qualities.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

Figure 2: After Preprocessing

The example dataset without any missing values or noises is presented in Figure 2 following the completion of the fundamental data cleaning tasks such the elimination of duplicate entries and outliers. The process of finding and removing irrelevant, tangentially relevant, or duplicated attributes or dimensions from a given data set is known as feature selection (FS) [35] [36]. Finding the smallest subset of attributes necessary will enable feature selection to provide probability distributions of data classes that are near to the original distribution generated using all attributes. One of the priciest activities in data mining assignments is comparison. In general, a data set's computing cost. D is

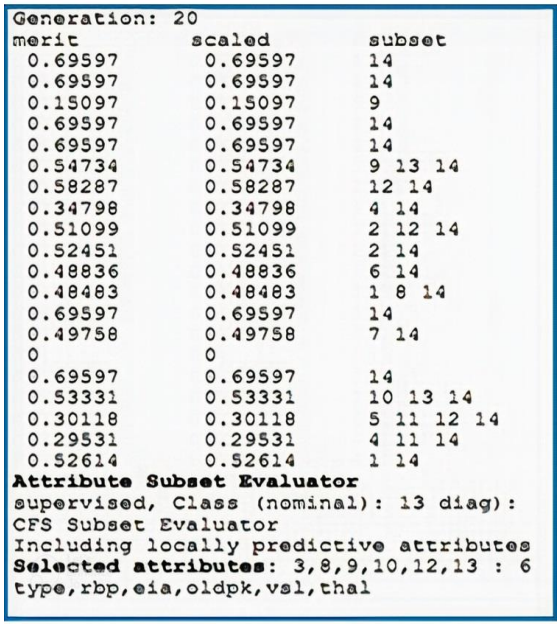
$$O (n \times |D| \times \log(|D|)) \quad (1)$$

Where, n – Number of attributes, D – Number of instances.

The number of comparisons required for m attributes and n instances is m * n2.

There are 2n subsets that can be created for a data collection D with n properties. It would

be expensive to search for the best subset, especially as n and the number of data types grow. It might not always be possible. As a result, heuristic methods make up the majority of feature selection techniques. These heuristic techniques are inherently greedy and attempt to investigate the potential reduced search space. There are two types of feature selection techniques. First feature ranking technique and second feature subset selection technique. In the former, all features are ranked by a metric like information gain, chi-square, etc. The features that do not achieve the adequate score are eliminated. In the later, the search is for the optimal subset of features that would be equivalent to the original subset of features. A subset of features are evaluated more commonly based on distance metrics like Euclidean, Hamming, etc. or filter metrics like entropy or probabilistic distance [37]. Common search approaches include greedy forward attribute selection, backward attribute selection, simulated annealing, and genetic algorithms. Genetic Algorithm (GA) incorporates natural evolution methodology.



```
Generation: 20
merit          scaled         subset
 0.69597        0.69597       14
 0.69597        0.69597       14
 0.15097        0.15097       9
 0.69597        0.69597       14
 0.69597        0.69597       14
 0.54734        0.54734       9 13 14
 0.58287        0.58287       12 14
 0.34798        0.34798       4 14
 0.51099        0.51099       2 12 14
 0.52451        0.52451       2 14
 0.48836        0.48836       6 14
 0.48483        0.48483       1 8 14
 0.69597        0.69597       14
 0.49758        0.49758       7 14
 0              0
 0.69597        0.69597       14
 0.53331        0.53331       10 13 14
 0.30118        0.30118       5 11 12 14
 0.29531        0.29531       4 11 14
 0.52614        0.52614       1 14
Attribute Subset Evaluator
supervised, Class (nominal): 13 diag):
CFS Subset Evaluator
Including locally predictive attributes
Selected attributes: 3,8,9,10,12,13 : 6
type,rbp,eia,oldpk,vsl,thal
```

Figure 3: Results of GA for feature subset selection

The initial population of the genetic search has randomly generated rules and a zero attribute. A new population is created in accordance with the fittest rules in the present population and the progeny of these laws based on the concept of survival of the fittest. children produced as a result of employing the genetic operators crossover and mutation. The generation process keeps on until it creates a population P where each rule meets the fitness requirement. The generation continued with a crossover probability of 0.6 and a mutation probability of 0.033 until the twentieth generation, from an initial population of 20 occurrences. Six out of the thirteen traits found by the genetic search are depicted in Figure 3.

3.3. Classification Algorithms

The Classification algorithm is a Supervised Learning method that employs training analysis to recognize the classification of new data. The fresh discoveries are categorized into one of several classes or groups by a programmer for classification after learning from a training sample or a set of results. Targets, labels, and categories are other names for classes. As opposed to regression, categorization produces a category as opposed to a value, such as "Green or blue," "fruit or animal," etc. Because the classification system uses supervised learning, it needs input data that has been explicitly labelled, so it has inputs and outputs. This section discusses the usage of classification algorithms like Random Forest (RF), AdaBoost, k-Nearest Neighbor (k-NN), and Support Vector Machine for the same goal as other classification approaches that are utilised by various researchers to predict and study cardiac disease (SVM).

Random Forest classification: It is a set of decision tree classifiers. It is a classifier for the ensemble approach. To decide the split, the Decision Tree is built using a random selection of characteristics at each node. During the categorization process, each tree vote returns the most popular class [38] [39].

Step 1: Select m features at random from the whole set of n features, m<n.

Step 2: Calculate the node d using the optimal split point surrounded by m features.

Step 3: For the best split, divide the node into daughter nodes.

Step 4: Repeat steps 1–3 until one number of nodes is attained.

The accuracy of a random forest depends on the strength of the individual nodes.

AdaBoost is a short form of Adaptive Boosting; it is a machine learning meta-algorithm. It is sometimes used in conjunction with many other

types of learning algorithms to improve performance. The output of the other learning algorithms on ('to weak learners') is combined into a weighted sum representing the final output of the boosted classifier. AdaBoost is adaptive in the sense that successive weak learners are biased toward cases misclassified by earlier classifiers. AdaBoost is susceptible to noise and outliers [40][41]. It may be less vulnerable to the overfitting problem than other learning algorithms in specific cases. The individual learners can be weak, but as long as each one's performance is slightly better than random guessing, the final model will prove to converge to a sharp lear

k- Nearest Neighbor (k-NN): k- NN classifier has been widely used in the area of pattern recognition [42] [43] [44]. Nearest-neighbor classifiers depend on learning by relationship, i.e., by contrasting a given test tuple and preparing tuples that are similar to it. The preparation tuples portrayed as n traits. Each tuple refers to a point in an n-dimensional space; hence, all preparation tuples are put away in an n-dimensional example space. At the point when given an obscure tuple, a k-closest neighbor classifier looks the example space for the K prepared tuples that are nearest to the obscure tuple. These k-prepared tuples are the k “closest neighbors” of the obscure tuple.

Support Vector Machine: SVM determines a hyperplane for data separation using the necessary tuples. In this instance, the two-class problem is resolved using linear separable. The disease data set is D. The groups yes and no relate, respectively, to disease. There are countless possible distances between hyperplanes. The hyperplane produces a line that may be divided into tuples of class 0 and class 1 by a straight line. Based on the data points that will be utilized to choose the appropriate hyperplane, we determine the ideal separation plane. Two potential separating hyperplanes connected by their margins were identified by the class labels. Both hyperplanes have a greater chance of correctly classifying the supplied data tuples. The biggest difference between the classes is provided by the accompanying margins. [45][46]. A splitting hyper plane is written as

$$W. X + b = 0, \quad (2)$$

W is a vector, (no. of attributes) ,
b is a scalar, X is a value of attributes.
The scalar b weight is adjustable. The hyperplane representing the margin side is expressed as,

$$H1: W0 + W1X1 + W2X2 >= 0 \quad (3)$$

$$H2: W0 + W1X1 + W2X2 <= 1 \quad (4)$$

Training tuples that fall on the hyperplaneH1or H2 satisfy equations (2) and (3) called support vectors. They equally closed to the Maximum Margin Hyper plane [MMH].

4. Results and Discussion

This section makes use of a rigorous methodology to discuss the outcomes that were obtained. Any type of research project requires a thorough analysis of the findings. To locate the patients with the selected ailment in the specified data set, this work used newly developed classification methods. Since heart disease is one of the main causes of death in humans, medical information on HD was taken into consideration in this study. Because they offer greater accuracy for medical data sets, the Random Forest, AdaBoost, k-NN, and SVM algorithms have been used in this work to examine the datasets. Preprocessing work was done based on all values of the taken attributes prior to classification. This work compared the classification accuracy of the Random Forest, AdaBoost, k-NN, and SVM algorithms. Additionally performed were the True Positive (TP) Rate, False Negative (FN) Rate, and accuracy analysis. The following list of formulas [47] is used to calculate various metrics. According to the formula, precision is the percentage of correctly anticipated positive cases.

4.1. Performance Metrics

The following criteria were used to validate the output in order to assess the outcomes of the experiments conducted in this study to predict heart disease.

Using the formula TP and FP projected to have heart disease, precision is the percentage of predicted positive instances that were accurate. People with heart conditions are TP.

$$Precision = TP/ (TP+FP) \quad (5)$$

Recall: The Peoples having heart diseases are TP and FN. The people diagnosed by the model having a heart disease are TP.

Recall = TP / (FP+FN) (6)

F-Measure: F-Measure is the balanced mean value of precision and recall.

F-Measure = 2 * Precision * Recall / Precision + Recall (7)

In this research work, the precision rate, recall value, F- Measure, and ROC values of the flowing algorithms Random Forest, AdaBoost, k-NN, and SVM are shown, respectively in Table 2 and Figure 4 for the data set having 14 attributes with 1204 records, which is shown in Figure 2 without using feature selection algorithm. The experiment was carried out on both the training and test datasets using the above-mentioned techniques and provided feature importance.

Table 2: Parameter values without feature selection (Genetic Algorithm)

Algorithms/ Parameters	Precision	Recall	F-Measure	ROC
Random Forest	0.810	0.802	0.823	0.537
AdaBoost	0.756	0.694	0.795	0.544
k-NN	0.678	0.681	0.680	0.629
SVM	0.872	0.852	0.862	0.793

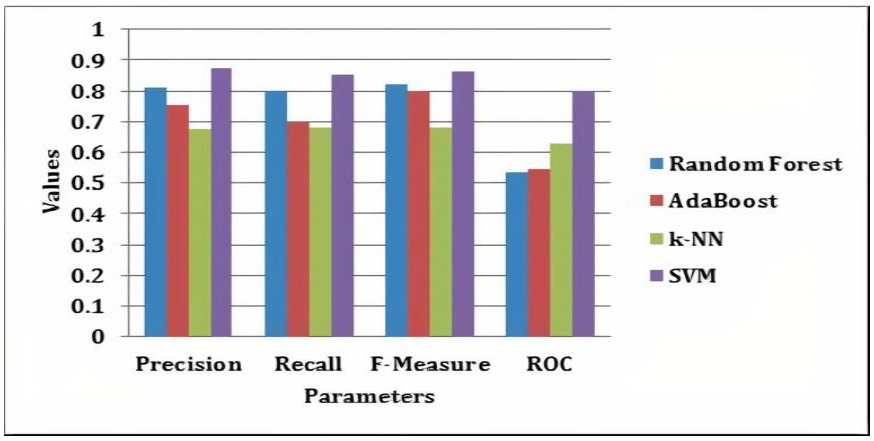


Figure 4: Parameter values without feature selection

It's clearly seen that, When compared to remaining three algorithms, the SVM approach produces the greatest Precision Rate (0.872), Recall value (0.852), F-Measure (0.862), and ROC (0.793) values

Table 3: Parameter values after feature selection (Genetic Algorithm).

Algorithms/ Parameters	Precision	Recall	F-Measure	ROC
Random Forest	0.869	0.912	0.925	0.734
AdaBoost	0.804	0.861	0.801	0.646
k-NN	0.912	0.882	0.860	0.719
SVM	0.969	0.972	0.926	0.933

Table 3 and figure 5 summarize the Precision Rate, Recall value, F-Measure, and ROC values for the flowing algorithms Random Forest, AdaBoost, k-NN, and SVM are shown respectively after selecting important features using Genetic algorithm for the same data set.

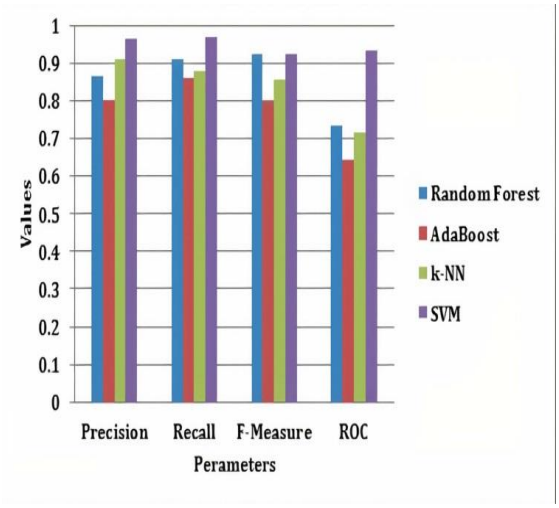


Figure 5: Parameter values after feature selection (Genetic Algorithm)

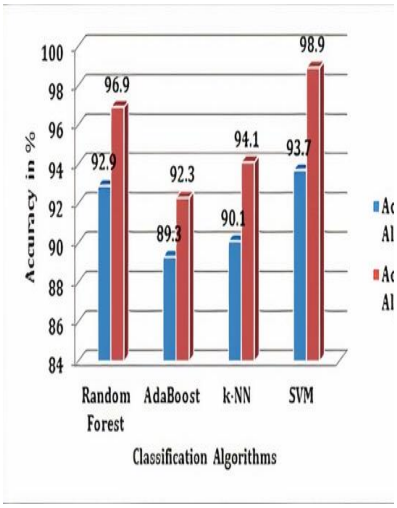


Figure 6: Results of Accuracy for Heart disease prediction

4.2. Results Analysis

This section compares the accuracy scores of the four methods, Random Forest (RF), AdaBoost, k-Nearest Neighbor (k-NN), and Support Vector Machine (SVM) for Heart Disease Prediction before and after utilising the genetic algorithms. The percentage of total predictions that were accurate is known as accuracy. Utilizing the following formula, it is determined:

$$Accuracy=TP+TN/\ (TP+TN+FP+FN) \quad (8)$$

Table 4 and Figure 6 show the comparison between genetic algorithm feature selection and genetic algorithm feature selection. To predict the accuracy, Table 4 shows the comparison results of supervised algorithms by means of correctly classified, incorrectly classified in (%) When analyzing Figure 6, before applying the feature selection algorithm (genetic algorithm), the Random Forest model got the accuracy value is 92.90%, AdaBoost Algorithm yielded the accuracy of 89.30%, k-NN yields 90.10 and SVM model yields 93.70%.

Table 4: Comparison of heart disease prediction accuracy

Classification Techniques	Accuracy Without Genetic Algorithm (%)	Accuracy With Genetic Algorithm (%)
Random Forest	92.9	96.9
AdaBoost	89.3	92.3
k-NN	90.1	94.1
SVM	93.7	98.9

However, after applying feature selection algorithm (genetic algorithm), the Random Forest model gets the accuracy value is 96.90%, AdaBoost Algorithm yields the accuracy of 92.30%, k-NN yields 94.10 and SVM model yields 98.90%. To see that, Random Forest algorithm and SVM Model attains the highest accuracy rate when comparing with the other two classification models.

5. Conclusion

By definition, heart disease unpredicted. This condition causes a heart attack and eventually death. The clinical domain has been completed, and efforts are being done to use appropriate methodologies in the prediction of heart disease. Classification algorithms used to forecast cardiac disease. It can assist forecasting heart disease based on clinical data from patients with the causes of heart disease. The present system analyzes heart disease prediction using two classification techniques: Nave Bayes and Random Forest. When compared to Nave Bayes, Random Forest produces superior outcomes. The Genetic Algorithm feature selection employed in this suggested system to decrease the characteristics of the heart disease data set. The features have been lowered while the prediction accuracy levels

are enhanced. After attribute selection, the Support Vector Machine (SVM) classifier method was employed to predict heart disease, along with the Random Forest (RF), AdaBoost, and k-Nearest Neighbor (k-NN) prediction algorithms. The SVM classifier generates good results and high metrics values for heart disease prediction when compared to the Naive Bayes, Random Forest, and SVM algorithms. Future studies will categorise and predict cardiac disease diagnosis using the deep learning-based LSTM technology. As a multivariate time series with multiple label categorization, the problem is presented.

List Of Abbreviationss

RF = Random Forest
k-NN = k-Nearest Neighbor
SVM = Support Vector Machine
GA = Genetic Algorithm
CVD = Cardio Vascular Disease
AUC = Accuracy

Conflict Of Interest

The authors confirm that this article content has no conflict of interest.

Acknowledgements

The authors thank the management for permitting us to do the research work.

References

[1] Murphy SL, Xu J, Kochanek KD, Arias E., "Mortality in the United States", 2017. NCHS Data Brief. Vol. 1, No. 8, 2018. PMID: 30500322.

[2] Maji S, Arora S., "Decision tree algorithms for prediction of heart disease" In Information and communication technology for competitive strategies, pp. 447–454, 2019.

[3] KaanUyar, Ahmet Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks". 9th international conference on theory and application of soft computing, computing with words and perception, Vol. 120, pp. 24-25, 2017. doi: https://doi.org/10.1016/j.procs.2017.11.283

[4] J. Nahar, T. Imam, K. S. Tickle, and Y.P. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females", Expert Systems with Applications, Vol. 40, No. 4, pp. 1086–1093, 2013.

[5] Mirmozaffari, Mirpouya, Alireza Alinezhad, and Azadeh Gilanpour, "Data Mining Apriori Algorithm for Heart Disease Prediction. International Journal of Computing", Communications & Instrumentation Eng., Vol. 4, No. 1. pp. 20-23, 2017.

[6] M. Shouman, T. Turner and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment", Conference on Electronics, Communications and Computers, pp. 173-177, 2012. doi: 10.1109/JEC-ECC.2012.6186978.

[7] Bhatla, Nidhi, and Kiran Jyoti, "An analysis of heart disease prediction using different data mining techniques", International Journal of Engineering Vol. 1, No. 8, pp. 1-4, 2012.

[8] S. N. Blair, Wang Y et al, "An Overview of Non-exercise Estimated Cardiorespiratory Fitness: estimation Equations, Cross-Validation and Application", Journal of Science in Sport and Exercise, Vol. 1, No. 1, pp. 94-95, 2019.

[9] K. Vanisree and J. Singaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptomsusing neural networks", International Journal of Computer Application, Vol. 19, pp. 6-12, 2011.

[10] Amin, Mohammad Shafenoor et al. "Identification of significant features and data mining techniques in predicting heart disease." Telematics Informatics Vol. 36, No.1, pp. 82-93, 2019. doi:10.1016/j.tele.2018.11.007

[11] N. Jagadeesan, T.Velmurugan, "Impact of Classification Algorithms for the Prediction of Heart Disease: A Survey", International Symposium on innovation in Information Technology and Application, Vol. 5, No. 1, pp. 417-430, 2021.

[12] Shalev-Shwartz, Shai, and Shai Ben-David, "Understanding machine learning: From theory to algorithms", Cambridge university press, 2014.

[13] B. S. S. Rathnayakc and G. U. Ganegoda, "Heart Diseases Prediction with Data Mining and Neural Network Techniques", 3rd International Conference for Convergence in Technology (I2CT), Pune, India, 2018, pp. 1-6.

[14] M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015, pp. 520-525.

[15] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network", International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, 2014, pp. 1-6.

[16] Das, Resul, Turkoglu, Ibrahim and Sengur, Abdulkadir. "Effective diagnosis of heart disease through neural networks ensembles". Expert Syst. Appl.36., 2009, pp.7675-7680. doi: 10.1016/j.eswa.2008.09.013.

[17] Yekkala, Indu & Dixit, Sunanda, "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection", International Journal of Big Data and Analytics in Healthcare, 2018, vol. 3, pp. 1-12, doi: 10.4018/IJBDAH.2018010101.

[18] P K, Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", Journal of King Saud University - Computer and Information Sciences, 2012, vol. 24, pp. 27-40, doi: 10.1016/j.jksuci.2011.09.002.

[19] H. A. Esfahani and M. Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier," 2017, 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017, pp. 1011-1014.

[20] Mrudula Gudadhe, Kapil Wankhade, SnehlataDongre, "Decision Support System for Heart Disease based on Support Vector Machine and Artificial Neural Network", 2010, IEEE doi: 978-1-4244-9034-/10.

[21] Rashmi G Saboji, Prem Kumar, Ramesh, "A Scalable Solution for Heart Disease Prediction using Classification Mining Technique", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017 doi: 978-1-5386-1887.

[22] Theresa Princy. J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", International Conference on Circuit, Power and Computing Technologies [ICCPCT] , 2016, doi: 978-1-5090-1277.

[23] Madhura Patil, Rima Jadhav, vishakha Patil, Aditi Bhawar, Mrs. Geeta Chillarge, "Prediction and Analysis of Heart Disease using SVM Algorithm", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 7, Issue 1, 2019, ISSN: 2321-9653.

[24] Snehal Subhash Mandavkar, Priyanka Pandurang Shinde, Anushka Sukhdev Jagdale, Prof. Vijaya Pinjarkar, "Heart diseases prediction system using classification and genetic algorithm", 4th International Conference on Advances in Science & Technology (ICAST), 2021, doi: http://dx.doi.org/10.2139/ssrn.3866565

[25] Takci, Hidayet, "Improvement of heart attack prediction by the feature selection methods", Turkish Journal of Electrical Engineering and Computer Sciences, Volume. 26, No. 1, 2018. doi: doi.org/10.3906/elk-1611-235.

[26] Chandra Reddy, N. S., Shue Nee, S., Zhi Min, L., & Xin Ying, C., "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction", International Journal of Innovative Computing, Volume. 9, No.1, 2019. doi: https://doi.org/10.11113/ijic.v9n1.210

[27] T Velmurugan, U Latha, "Classifying Heart Disease in Medical Data Using Deep Learning Methods", Journal of Computer and Communications, Vol. 9, pp. 66-79, 2021. doi: https://doi.org /10.4236/jcc.2021.91007

[28] Padmapriya, B., and T. Velmurugan. "Classification algorithm based analysis of breast cancer data." International Journal of Data Mining Techniques and Applications, Vol. 5. No. 1 43-49, 2016.

[29] J. Dhinakaran and T. Velmurugan, "Accuracy based Performance Analysis of Classification

Algorithms using Parkinson Disease Dataset", International Conference on Applied Artificial Intelligence and Computing, pp. 927-933, 2022. doi: 10.1109/ICAAIC53929.2022.9792981.

[30] Sadeghian, Farzaneh, Hadiseh Hasani, and Marzieh Jafari, "Feature Selection Based on Genetic Algorithm in the Diagnosis of Autism Disorder by fMRI", Caspian Journal of Neurological Sciences Vol. 7, No. 2, pp. 74-83, 2021.

[31] Gokulnath, C.B., Shantharajah, S.P., "An optimized feature selection based on genetic approach and support vector machine for heart disease", Cluster Comput 22, pp. 14777–14787, 2019. doi : https://doi.org/10.1007/s10586-018-2416-4

[32] Murphy SL, Xu J, Kochanek KD, Arias E., "Mortality in the United States", 2017. NCHS Data Brief. Vol. 1, No. 8, 2018. PMID: 30500322.

[33] N G, Bhuvaneswari, "Cardiovascular disease prediction system using genetic algorithm and neural network", Computer Commination and Application, Vol. 5. 2012. doi: https:// doi.org/10.1109/ICCCA.2012.6179185

[34] H. Benhar, A. Idri, J.L. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review" Computer Methods and Programs in Biomedicine, Vol. 195, 2020.

[35] Ghaheri A, Shoar S, Naderan M, Hoseini SS. "The Applications of Genetic Algorithms in Medicine." Oman medical journal Volume 30, No. 6, 2015. doi:10.5001/omj.2015.82

[36] Shokoufeh Aalaei, Hadi Shahraki, Alireza Rowhanimanesh, and Saeid Eslami, "Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets." Iranian journal of basic medical sciences Vol.19, No. 5, 2016. PMID: 27403253

[37] Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath, "Feature subset selection using cascaded GA & CFS: a filter approach in supervised learning", International Journal of Computer Applications Vol. 23, No. 2, pp. 1-10, 2011.

[38] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Medical and Health Sciences Volume 1, No 5, 2007.

[39] Patil P, R. Kinariwala, A.S., "Automated Diagnosis of Heart Disease using Random Forest Algorithm", International Journal of Advance Research Ideas and Innovations in Technology, Volume 3, No. 3, pp. 579-589, 2017.

[40] Gomez-Rios, Anabel, Julian Luengo, and Francisco Herrera, "A study on the noise label influence in boosting algorithms: AdaBoost, GBM and XGBoost." In International Conference on Hybrid Artificial Intelligence Systems, Springer, pp. 268-80, 2017.

[41] Olatunde, Yusuf, Lawrence Omotosho, and Caleb Akanbi. "Comparison of adaboost and bagging ensemble method for prediction of heart disease." Annals. Computer Science Series, Vol. 17, No. 1, 2019.

[42] Jabbar MA, "Prediction of heart disease using k-nearest neighbor and particle swarm optimization", Journal of Biomedical Research, Volume 28, No. 9, pp. 4154-4158, 2017.

[43] Enriko I. K. A., Muhammad, Suryanegara. and Dadang, Gunawan, "Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters", Journal of Telecommunication Electronic and Computer Engineering, Vol. 8, No. 12, pp. 59-65, 2016.

[44] N. S. C. Reddy, S. S. Nee, L. Z. Min, and C. X. Ying, "Classification and feature selection approaches by machine learningtechniques: heart disease prediction", International Journal of Innovative Computing, Vol. 9, 2019.

[45] Sandhya, Yamala, "Prediction of Heart Diseases using Support Vector Machine", International Journal for Research in Applied Science and Engineering Technology, Volume 8 No 1, pp. 126-135. doi: 10.22214/ijraset.2020.2021.

[46] Katarya, Rahul, and Polipireddy Srinivas, "Predicting heart disease at early stages

using machine learning: a survey." International Conference on Electronics and Sustainable Communication Systems, IEEE, 2020. doi: 10.1109/ICESC48915.2020.9155586

[47] Setiawan, Wahyudi & Damayanti, Fitri, "Layers Modification of Convolutional Neural Network for Pneumonia Detection", Journal of Physics: Conference Series, 2020. doi: 10.1088/1742-6596/1477/5/052055.

[48] Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. "Enhanced prediction of heart disease with feature subset selection using genetic algorithm", International Journal of Engineering Science and Technology, Vol.2, Issue 10, pp.5370-5376.