

A Study on Prediction of Diabetic Coronary Heart Disease Using Machine Learning Algorithms

S. Madhumalar¹, S. Sivakumar²

¹Assistant Professor of Computer Science, CPA College, Bodinayakanur, India

²Principal, Associate Professor, CPA College, Bodinayakanur, India

E-mail: ¹madhumalar.s@cpacollege.org, ²sivaku2002@yahoo.com

Abstract

One of the most prevalent illnesses in today's fast-paced culture is coronary heart disease in people with diabetes. It is challenging to foresee the onset of diabetic heart disease, a chronic ailment that affects people all over the world. One common misconception about machine learning is that it is an inaccessible field of study reserved for nerdy academics. However, concerning accuracy to predict disease based on symptoms, many studies have turned to machine learning methods, and this study provides a comparative analysis of these methods appeared in published research articles. This paper takes a close look at the research into predicting diabetic coronary artery disease using machine learning techniques, as well as the many different combinations of learning algorithms that have been employed to do so.

Keywords: Diabetic coronary heart disease, machine learning

1. Introduction

The presence of diabetic coronary artery disease significantly worsens a patient's long-term prognosis. Over time, elevated blood sugar may harm the heart's blood vessels and the nerves that regulate cardiac function. Additionally, diabetics are mortal heart disease due to numerous additional risk factors. The increased blood flow caused by high blood pressure is harmful to the arteries. Diabetic coronary artery disease is one of the noted causes of sudden death across the world now. There are 17.5 million annual deaths caused by diabetes-related heart disease [1]. Old people have high liability for diabetic coronary artery disease which is vital for several areas of recuperative and medical practice. A liability prediction model can be developed by the correlation analysis of a prolonged research. [2-3]. Purpose of

analyzing data, hospitals and clinics rely extensively on machine learning methods. These techniques appeal to be the correct dataset to help with model building meaningful conclusions and inferences. Many factors, including age, gender, and race, contribute to a diagnosis of diabetic coronary heart disease.

Many major vessels are colored on a fluoroscopy scan, if chest pain is experienced when taking a resting electrocardiogram (a test that measures electrical activity in the heart). Increasing salt intake has been shown to increase blood pressure (high blood pressure). A person's risk of developing heart disease can be determined by their results on the Thalach (maximum heart rate achieved), ST(STEMI) depression (discovered on an ECG, trace in the ST segment is abnormally low below the baseline), and painloc (location of chest pain; substernal=1, otherwise=0) tests. Exercise-induced angina (Exang), cigarette use, and hypoglycemia (low blood sugar), diet, weight, and hypertension [4] are a few causes of heart disease. Tabular presentation of the most common forms of heart illness has been given below.

Table 1. Types of heart disease [5]

Arrhythmia	Any of these: irregular, too slow, or too fast heart rates indicate a problem.
Cardiac arrest	Suddenly, the victim's heart stops beating, and he or she loses consciousness and breathing.
Congestive heart failure	Heart failure is a long-term disorder in which the organ fails to adequately pump blood throughout the body.
Congenital heart disease	Congenital heart defect; cardiac defect present at birth.
Coronary artery disease	The major arteries and veins are vulnerable to injury and disease.
High Blood Pressure	A heart attack can occur if the blood pressure in the arteries is high.
Peripheral heart disease	Limited blood flow to the extremities is a hallmark of cardiovascular disease, which causes arterial narrowing.
Stroke	If blood flow to the brain is suddenly cut off, it causes stroke. An irreversible damage can occur.

1.1 Machine Learning

Classifying private data in the input database and developing models are fundamental to machine learning algorithms [5]. It is able to make accurate forecasts in relation to databases that are completely novel to algorithms. Big data and predictions are of no problem for the algorithms and programs used in machine learning [6]. Schematics of machine learning techniques and its algorithms are given in Fig.1

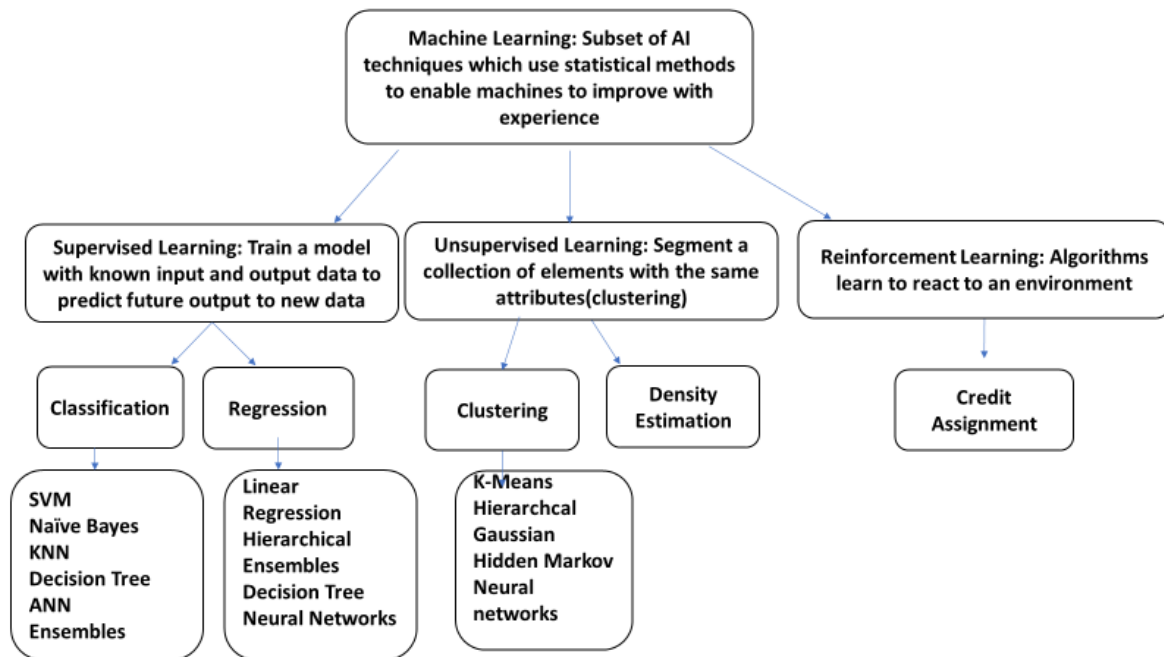


Figure 1. Machine learning Techniques

1.2 Classification

A record's classification is determined by the algorithms. For questions with a restricted number of possible responses, this type of inquiry is appropriate. It is called binary classification when there are only two possible outcomes, and multi class classification when there are more than two. [7].

1.2.1 Naive Bayes

To classify data, a Naive Bayes classifier uses Bayes' theorem. This model is a probabilistic machine learning classifier used to distinguish between objects based on their characteristics.

1.2.2 Artificial neural network

An interconnected group of nodes that mimics the structure of the brain is used. In this illustration, the link between neurons is shown as an arrow. The most popular ANN-based approach is the gradient descent and back propagation technique.

1.2.3 Decision tree

One of the methods of predictive modeling employed by ML and statistics is Decision Tree (DT). Gain and entropy are two metrics used to partition the dataset in this approach. It is widely employed in the fields of classification and regression, and is both supervised and non-parametric.

1.2.4 K-nearest neighbor

Assigning a weight to each data point, or "neighbor", is the key to the K Nearest Neighbor algorithm's core idea.

1.2.5 Support Vector Machine

Data categorization and regression analysis can be carried out using support vector machines by employing the technique called Support Vector Machine (SVM). In this procedure, information is first classified into binary categories.

1.2.6 Ensemble learning

The goal of this technique is to produce a highly accurate prediction model by integrating many simpler models.

1.2.7 Regression

An ensemble of machine learning techniques permits to make predictions about a continuous outcome variable when one or more input values are given.

1.2.8 Clustering

The process entails discovering data clusters in an automated fashion. In simple words, this algorithm does nothing more than reading the data that is given to it, and it organizes into meaningful categories.

1.2.9 K-means

This technique, which has its roots in signal processing, is a form of vector quantization that seeks to divide observations into groups, each of which has a prototype based on the mean of the observations closest to it.

1.2.10 Hierarchical

It's a method for classifying data into clusters based on shared features. The result is a collection of clusters, each of which is distinct from the others and whose members share several characteristics in common.

1.2.11 Gaussian

The form is a function for any three non-zero real constants a , b , and c . It's a bell curve because the same number of readings is typically expected above and below the mean throughout any given experiment.

1.2.12 Hidden Markov

To simplify, it is a statistical model in which the studied system is thought of as a Markov process with hidden (or unobservable) states. A necessary condition is the existence of a measurable process whose results are "affected" by the results of another process in a specified way.

1.2.13 Density estimation

Estimating the parameters of a density function based on a sampling of data and selecting a common distribution.

1.2.14 Credit assignment

To attribute an outcome to a specific activity or collection of events within an episode.

2. Review of Literature

Many of them, according to a review of previous studies, have devoted significant time and energy in analyzing the disease's medical data. As another example, the use of data mining and machine learning techniques to the building of analytical and predictive models

have been discussed. The most closely connected researchers between 2015 and 2021, together with the datasets and tools they utilized, are listed in Table 2. Many studies in hospitals have tried to forecast illness by employing data mining and machine learning techniques.

Heart disease risk assessment was performed using WEKA, and the K Star, j48, SMO, and Bayes Net models were developed by Maria et al. [8]. When compared to K Star, Multilayer perception, and J48, the performances of SMO and Bayes Net with k-fold cross validation were superior for several reasons. The accuracy achieved by these algorithms remains subpar. It's a win-win because it improves the precision with which diseases are diagnosed.

Naive Bayes, Decision Trees, Support Vector Machines, and Artificial Neural Networks (ANN) are only few methods that can be used to mine data from historical medical records and make predictions for chronic diseases, as described by S. Seema et al. [9]. A comparative study was conducted to establish which classifier yields the best results. Among the methods tested, SVM performed best, whereas Naive Bayes performed best when predicting diabetes.

The algorithms Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM, and ANN were advocated by Ashok Kumar Dwivedi et al. [10]. The Logistic Regression technique outperformed other approaches in terms of accuracy.

Prediction of cardiac illness by means of data mining was proposed by MeghaShahi et al. [11]. Software developed by the World Health Organization called WEKA is used in hospitals and clinics worldwide to aid in disease diagnosis and enhance patient care. The study used SVM, Naive Bayes, Association rule, KNN, ANN, and Decision Tree algorithms. Since SVM is faster and more accurate than competing approaches, it is recommended that it be utilized in this research.

Methods for the detection and prediction of heart problems in databases were advocated by Chala Beyene et al. [12]. The primary objective was to anticipate the onset of heart illness so that a diagnosis can be made promptly and mechanically. Experts with the same level of experience and knowledge are essential to the planned procedure. Some medical indicators for cardiovascular illness include blood sugar and heart rate, for instance. Dataset analyses are calculated with the help of WEKA software.

R. Sharmila et al. [13] developed a non-linear classification method for cardiac disease prediction. For optimal attribute set prediction of heart disease utilizing big data techniques, Hadoop Distributed File System (HDFS), Map Reduce, and Support Vector Machines (SVM) were employed. The purpose of the research was to determine if data mining techniques could aid in the prediction of heart illness. It suggested using HDFS to store data distributed across many nodes, and then executing the SVM prediction algorithm in parallel on each of those nodes. Compared to sequential SVM, processing times were reduced when SVM was parallelized.

Data mining and machine learning algorithms for the diagnosis of heart illness were proposed by Jayami Patel et al. [14]. This initiative intends to use data mining methods to discover hidden relationships. When compared to LMT, the UCI-based algorithm J48 performed best.

In order to reliably forecast heart illness, Purushottam et al. [15] suggested a data mining-based strategy. In this manner, the physician can make an educated decision based on a narrow metric. Due to its ability to test and train a single parameter, it achieves an impressive 86.3% accuracy during the testing phase.

K. Gomathi et al. [16] proposed using data mining techniques to forecast the occurrence of many diseases simultaneously. Data mining is now indispensable for forecasting the course of many different diseases. Data mining methods can help cut down on the number of necessary experiments. The primary focus of the research was on the relationships between obesity, diabetes, and breast cancer.

In a proposal made by P.SaiChandrasekhar Reddy et al. [17], with the use of an ANN algorithm, data mining can be used to make accurate predictions about cardiovascular disease. As the cost of diagnosing heart disease has increased, the need for a new system that can predict heart illness has emerged. Using a patient's vitals, such as their pulse and blood pressure, as well as their cholesterol level, a diagnosis can be made. Correctness of the system was shown in Java.

Ashwini Shetty et al. [18] referred that, medical records be used to create a heart disease prediction system. The method was developed considering the risks associated with 13 different input attributes. When the dataset was being analyzed, it was cleaned and integrated.

Jaymin Patel et al. [19] illustrated that data mining and machine learning allow for the forecasting of cardiovascular illness. There were two main objectives in cardiac system prediction. First, the system makes no assumptions about the patient's history. Next, since there will be so many records, the system needs to be scalable. It can be used in the WEKA software. The testing phase of the project makes use of WEKA's classification tools and explorer mode.

Analysis of data mining methods for predicting and diagnosing cardiovascular illness, by Boshra Brahmi et al., [20] evaluated classification strategies including J48, Decision Tree, KNN, SMO, and Naive Bayes. The next step was to evaluate results using several measures of accuracy, precision, sensitivity, and specificity. J48 and a decision tree are the most effective tools for predicting cardiovascular disease.

Artificial neural networks were proposed for the diagnosis of cardiac disease by Noura Ajam [21]. To evaluate the performance of the model, back propagation learning methods have been applied. The use of the appropriate function improved classification accuracy to 88% and led to the use of 20 neurons in the hidden layer. Improvements in cardiac disease prediction using ANN were substantial.

To foresee cardiac arrests, Prajakta Ghadge et al., [22] proposed utilizing big data. It is the intent of this piece to outline the steps necessary to create a prototype by utilizing big data and data modeling techniques. Heart disease-related patterns and correlations can be extracted from datasets using the method. The second database in the system is an updated version of the first. Reliable service was provided to consumers by employing HDFS, a java-based file system. It's possible that doctors and other medical staff can benefit from this technology by getting additional data before making judgments about their patients. Automatic processing of this kind might be useful.

The DNFS method was suggested by S. Prabhavathi et al., [23] to analyze and forecast cardiovascular conditions. The paper reviewed studies that looked into how to identify heart problems at an early stage. The DNFS refers to a neural fuzzy system that makes decisions using a decision tree. The goal of the study was to develop a system that is both more efficient and less expensive than the current one. More specifically, data mining methods were used to enhance heart illness prognosis. Both SVMs and neural networks performed very well as predictors of cardiovascular illness. The ability of data mining methods to reliably forecast cardiac issues remains debatable.

Predictions of cardiovascular disease risk were made using k-means and naive bayes by Sairabi H.Mujawar et al. [24]. A system that draws on a past medical record of cardiac diagnosis was created. The system's design took into account 13 criteria to determine whether a patient has heart disease. The information stored in a database can be mined for useful information by employing methods like clustering and classification. Three hundred records were used to represent thirteen different qualities, all of which were taken from the Cleveland Heart Database.

Monica L. Sharan et al., offered cardiovascular disease as the subject for study. In the research, data mining methodologies were developed for disease prediction. The evaluation of existing methods for data extraction from datasets will be helpful to healthcare practitioners. The time required to construct the decision tree is a good indicator of the system's efficiency. The primary objective was to streamline the requirements for making reliable condition forecasts.

Sharma et al., [26] reported that the prediction of heart disease can be made utilizing the c45 principles and the partial tree method. A general set of guidelines for estimating patients' danger levels in relation to one particular health indicator can be determined. Metrics like precision and error can be derived from the obtained results and rules. Then, C4.5 was compared to a subtree. Based on the results, improved forecasting and higher productivity are possible.

Table 2 displays a variety of cardiac diagnosis and treatment options. This table contains the numerous combinations of machine learning algorithms and techniques used in the dataset from 2017 to 2021. There are various algorithms and tools developed for improving the high level of accuracy.

Table 2. Comparisons of various classification algorithms in literature (2017-2021)

Published year	Authors	Motive	Techniques	Dataset	Percentage obtained
2021	Hisham et al.	Predicting CAD techniques using ML	SVM algorithm	South African dataset	73.82%
2020	Amir et al.	Creating ML model using features selection and PCA	SVM-PCA	Cleveland dataset	97.2%
2020	Yamala et al.	Heart disease prediction using SVM	SVM	UCI Repository	85.9%

2019	Nagaraj et al.	Predicting the heart disease using SVM algorithm	SVM	UCI Repository	90.23%
2018	Sharmila et al.	Improving the accuracy of predicting heart disease using data approaches	SVM in parallel fashion	UCI Repository	82.35%
2018	Chala Bayen et al.	Data mining used to heart disease prediction	J48, NB, SVM	UCI Repository	87%
2017	Jayami et al.	Predict the heart disease using data mining and machine learning algorithms	LMT, UCI	Cleveland dataset	88%
2017	P. Sai et al.	DM used to predict the heart disease	ANN	Cleveland dataset	98%

The following table details the 2015-2016 dataset classification algorithms. Typical methods and procedures are employed in this dataset. The Naïve Bayes algorithm was the most widely utilized algorithm in the published research.

Table 3. Comparisons of various classification algorithms in literature (2015-2016)

Published year	Authors	Motive	Techniques	Dataset	Percentage obtained
2016	Marija et al.	Weka tool to predict the heart disease	NB ANN J48 SMO	Cleveland HD dataset	76.12% 85.23% 86.14% 87.35%
2016	S. Seema et al.	Predict chronic disease	NB DT SVM	Cleveland HD dataset	94.6% 83.4% 87%
2016	Ashok Kumar et al.	ML methods used to predict the heart disease	NB KNN LR SVM	UCI	83.12% 85.34% 80.15% 77.65%
2016	Gomathy et al.	Multi disease prediction	Naïve Bayes J48	Cleveland HD dataset	79.45% 83.12%
2016	Jaya et al.	Combined heart disease prediction using ML and DM	J48 LR	UCI	56.78% 75.8%
2016	Ashwin et al.	Efficiently predict heart disease	WEKA	UCI	88.23%

2016	S. Prabhavathi et al.	DNFS techniques used to predict heart disease	DT C4.5 ANN SVM	Cleveland	85.4% 82.4% 92.1%. 87.2%
2016	Sharan L et al.	Predict heart disease using DM.	J48 NB, CART	UCI	91.4% 88.5% 92.2%
2015	Noura et al.	ANN used prediction of heart disease	ANN	Cleveland HD	97.3%
2015	Sharma et al.	Evaluating the prediction of heart disease	C4.5 NB	UCI	83.4% 80.12%
2015	Sairabi et al.	Modified heart disease prediction	K-means NB	UCI	93%. 89%.
2015	Boshra et al.	DM used for predicting and diagnosis the heart disease	J48 NB KNN SMO	Cleveland HD	87% 82.12% 85.1% 86.34%
2015	Sharma et al.	DT based heart disease prediction	Decision tree classifier	UC	87.3%

3. Observations and findings

Most of the relevant publications analyzed made use of data from UCI, the Cleveland HD dataset, and South African sources. Medical organizations should make more data publicly available to enhance ML prediction in order to keep up with future expectations and encourage additional research. Many machine learning techniques that are discussed and briefly explained are those employed in the majority of the reviewed papers. The Naïve Bayes, Decision tree, Support Vector Machine, K-Nearest Neighbor, Artificial Neural Network, and Ensemble Algorithms (Boosting, Bagging, Stacking, and Voting) are all examples of classifiers. The Naive Bayes and the KNN algorithms both yield superior accuracy, as seen in table 3. Table 2 shows that the combination of the SVM with the ANN algorithms improves the percentage of final result.

It is suggested that decision trees, ensemble methods, multiple classifier voting, and artificial neural networks are all methods that can be tried. Accuracy in algorithms can be increased by employing more pertinent feature selection strategies. While some progress has been made in developing a predictive model for people with diabetic heart disease, it is believed that convoluted models are required to better certainty of early symptoms efficient prediction of diabetic coronary heart disease. Moreover, the findings suggest the following

considerations for future research into the use of machine learning algorithms for the prediction of diabetic coronary artery disease with high accuracy and precision. Prediction models have often been trained using a very limited and consistent dataset across studies.

4. Conclusion and Future work

Predicting diabetic coronary heart disease is a significant difficulty in the healthcare industry, and this work has analyzed the existing machine learning classification algorithms for doing so. Several hundred publications on the topic of classifying data with machine learning have been surveyed and organized. The work briefly describes how supervised, unsupervised, and reinforced machine learning methods can be used to forecast diabetic heart disease. The quality of the proposed models varies with the resources and dataset employed, as well as the number of attributes and the number of entries. Adding more data to the database will allow it to make better decisions. By investigating several avenues, it is feasible to improve the system's scalability and precision. After completing this study, the prediction algorithms will be trained and tested using data from a sizable sample of diabetic individuals in our country who also have coronary artery disease. The accuracy of the prediction models must be checked by applying them to sizable datasets.

References

- [1] Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Research Gate Publications, July 2017, pp.2137-2159.
- [2] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review", International Journal of Computer Applications, February 2016.
- [3] Guizhou Hu, Martin M. Root, "Building Prediction Models for Coronary Heart Disease by Synthesizing Multiple Longitudinal Research Findings", European Science of Cardiology, 10 May 2005.
- [4] T. Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, "A Heart Disease Prediction Model using SVM, Decision Trees- Logistic Regression (SDL)", International Journal of Computer Applications, vol 68, 16 April 2013
- [5] <https://www.medicalnewstoday.com/articles/257484.php>.

- [6] Nimai Chand Das Adhikari, Arpana Alka, and rajat Garg, “HPPS: Heart Problem Prediction System using Machine Learning”.
- [7] K. Polaraju, D. Durga Prasad, “Prediction of Heart Disease using Multiple Linear Regression Model”, *International Journal of Engineering Development and Research Development*, ISSN:2321-9939, 2017.
- [8] Marjia Sultana, Afrin Haider, “Heart Disease Prediction using WEKA tool and 10-Fold cross-validation”, *The Institute of Electrical and Electronics Engineers*, March 2017.
- [9] Dr.S.Seema Shedole, Kumari Deepika, “Predictive analytics to prevent and control chronic disease”, <https://www.researchgate.net/publication/316530782>, January 2016.
- [10] Ashok kumar Dwivedi, “Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation”, *Springer*, 17 September 2016.
- [11] Megha Shahi, R. Kaur Gurm, “Heart Disease Prediction System using Data Mining Techniques”, *Orient J. Computer Science Technology*, vol.6 2017, pp.457-466.
- [12] Mr. Chala Beyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, *International Journal of Pure and Applied Mathematics*, 2018.
- [13] R. Sharmila, S. Chellammal, “A conceptual method to enhance the prediction of heart diseases using the data techniques”, *International Journal of Computer Science and Engineering*, May 2018.
- [14] Jayami Patel, Prof. Tejal Upadhyay, Dr. Samir Patel, “Heart disease Prediction using Machine Learning and Data mining Technique”, March 2017.
- [15] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, “Efficient Heart Disease Prediction System”, 2016, pp.962-969.
- [16] K.Gomathi, Dr.D.Shanmuga Priyaa, “Multi Disease Prediction using Data Mining Techniques”, *International Journal of System and Software Engineering*, December 2016, pp.12-14.
- [17] Mr.P.Sai Chandrasekhar Reddy, Mr.Puneet Palagi, S.Jaya, “Heart Disease Prediction using ANN Algorithm in Data Mining”, *International Journal of Computer Science and Mobile Computing*, April 2017, pp.168
- [18] Ashwini Shetty A, Chandra Naik, “Different Data Mining Approaches for Predicting Heart Disease”, *International Journal of Innovative in Science Engineering and Technology*, Vol.5, May 2016, pp.277-281

- [19] Jaymin Patel, Prof. Tejal Upadhyay, Dr.Samir Patel, “Heart Disease Prediction using Machine Learning and Data Mining Technique”, International Journal of Computer Science and Communication, September 2015-March 2016, pp.129-137.
- [20] Boshra Brahmi, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164- 168.
- [21] Noura Ajam, “Heart Disease Diagnoses using Artificial Neural Network”, The International Institute of Science, Technology and Education, vol.5, No.4, 2015, pp.7-11.
- [22] Prajakta Ghadge, Vrushali Girme, Kajal Kokane, Prajakta Deshmukh, “Intelligent Heart Disease Prediction System using Big Data”, International Journal of Recent Research in Mathematics Computer Science and Information Technology, vol.2, October 2015 – March 2016, pp.73-77.
- [23] S.Prabhavathi, D.M.Chitra, “Analysis and Prediction of Various Heart Diseases using DNFS Techniques”, International Journal of Innovations in Scientific and Engineering Research, vol.2, 1, January 2016, pp.1-7.
- [24] Sairabi H.Mujawar, P.R.Devale, “Prediction of Heart Disease using Modified K-means and by using Naïve Bayes”, International Journal of Innovative research in Computer and Communication Engineering, vol.3, October 2015, pp.10265-10273.
- [25] Sharan Monica.L, Sathees Kumar.B, “Analysis of CardioVascular Disease Prediction using Data Mining Techniques”, International Journal of Modern Computer Science, vol.4, 1 February 2016, pp.55-58.
- [26] Sharma Purushottam, Dr. Kanak Saxena, Richa Sharma, “Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree”, Springer, Computational Intelligence in Data Mining, vol.2, 2015, pp.285-294. IJCATM : www.ijcaonline.org
- [27] Nagaraj M.Lutimah, Arathi B” Prediction of heart disease using SVM “July 2019
- [28] Yamala Sandhya, Sreenidhi “Prediction of heart disease using SVM”, February 2020
ijraset2020-2021
- [29] Anna Karen, Amir Hajjam” Classification models for heart disease prediction using feature selection and PCA” Elsevier Informatics to medicine 2020
- [30] Hisham khdaif, Naha Dasari, ”Exploring Machine learning techniques for coronary heart disease prediction” International Institute of Business of Information Technology vol.12, No.5, 2021.