DOI : https://doi.org/10.56452/7-2-551

ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

Predicting Heart Disease Using Machine Learning in Healthcare Based on Several Data Sources

Nisha Chandran S

Associate Professor, School of Computing, Graphic Era Hill University, Dehradun, Uttarakhand India 248002,

Abstract: In this paper, cardiovascular illnesses have been the leading cause of mortality across all socioeconomic groups in the globe over the last few decades. The mortality rate from heart disorders may be lowered with early identification and close clinical monitoring. However, it is not feasible to precisely monitor patients every day, and 24-hour medical consultation is not accessible due to the increased sagacity, time, and experience required. For this project, we used Machine Learning methods such as backward elimination algorithm, logistic regression, and REFCV to analyse a publicly available dataset on the Kaggle website in order to predict the onset of heart disease in patients. The results were evaluated by means of a confusion matrix and cross validation. It would be a huge step forward in medicine if doctors could accurately predict the course of cardiovascular disease in order to let high-risk patients make choices about lifestyle adjustments that would lower the likelihood of serious problems.

Keywords: Machine Learning, Logistic regression, CrossValidation, Backward Elimination, REFCV, Cardiovascular Diseases.

Introduction

The World Health Organisation estimates that 12 million people die from heart disease each year. Worldwide, the prevalence of cardiovascular disease has been steadily rising over the last several years[1]. Numerous studies have been undertaken to identify the most important risk factors for cardiovascular disease and to provide reliable risk assessments. Heart disease is sometimes called a "silent killer" since there are often no outward signs that anything is wrong until it is too late. The ability to make informed choices about lifestyle adjustments in high-risk individuals is greatly aided by an early diagnosis of heart disease. With the use of machine learning algorithms, this initiative hopes to foresee the onset of cardiac disease in individual individuals[2]. The main focus is on forecasting using ML methods. These days, machine learning is applied in a wide variety of e-commerce and other commercial applications. One of the many applications of machine learning is prediction; in this case, we're interested in predicting the likelihood that a certain patient would develop heart disease by analysing data from a dataset comprised entirely of medical records. Python, the most popular programming language, is being utilised in this project's machine learning model because of its extensive library[3]. Machine learning is a branch of AI that makes use of sophisticated computer programmes and neural networks to learn new tasks. There are many different organs in the human body, and they all perform important tasks. The heart is an example of an organ that is vital to life because it circulates blood throughout the body. Heart disease is a leading cause of death in the modern world. That's why it's so important to take care of our hearts and brains and everything else

DOI : https://doi.org/10.56452/7-2-551

that makes us human[4]. Cardiovascular illnesses are a major global health problem. An early diagnosis of these disorders is crucial to preventing costly medical interventions and, more importantly, preserving lives. Predicting cardiovascular illnesses using data mining methods may be quite useful. By mining data for previously unseen patterns and trends, analytical models may be constructed. Machine learning is a technique that may aid in early cardiac disease detection, which can prevent further harm to the patient[5]. Machine learning, a relatively new scientific and technological discipline, may classify whether or not a person has heart disease. When compared to other potentially fatal diseases, heart attacks have the highest incidence rate. Multiple surveys are conducted by medical experts to collect data about heart patients, their symptoms, and the progression of their condition[6-8]. Heart attacks are more prevalent and sometimes deadly. There were warning signs in the form of symptoms. Increases in healthcare quality are largely attributable to the effective application of technical advances by the medical community. This advancement in technology has paved the way for more accurate medical diagnosis and prognosis. You may be able to improve the accuracy of your heart disease predictions by using machine learning. That's why three algorithms are being put into action. Logistic regression, decision tree, and random forest are all methods included in this set. These three techniques also provide substantially quicker and more reliable results. Due to developments in technology, forecasting is becoming easier. Today's global populace enjoys the finer things in life thanks to their tireless pursuit of wealth and celebrity. Due to their busy lives, people often neglect their health[9],[1].

Types of Cardiovascular Disease

Heart disease, or cardiovascular disease (CVD), is a cluster of conditions affecting the circulatory system. Myocardial infraction, sometimes known as a heart attack or angina, is one kind of heart disease that may be further broken down into subtypes. Another kind of heart disease is coronary heart disease (CHD), which occurs when a waxy material called plaque builds up within the coronary arteries. The heart muscle will rely on these coronary arteries to get oxygenated blood. Atherosclerosis is the buildup of plaque, a waxy material, in these arteries[10]. Plaque buildup within the artery will continue for a long time. If plaque development is not detected in its earliest stages, it may cause the heart to harden or burst. Harder plaque builds up over time, narrowing the coronary artery and reducing the amount of oxygen-rich blood that can reach the heart. A blood clot will develop on the plaque's surface if it hardens. When a blood clot forms in the coronary arteries, its size may sometimes totally obstruct blood flow. The blood circulation must be maintained rapidly. If not, the affected area of heart muscle will begin to atrophy. When heart disease strikes, immediate medical attention is required to prevent further complications or even death[11-13].

Risk Factors of Heart Disease

The causes of the progressing obstruction are the risk factors. These danger indicators are split into two classes: those that can be adjusted and those that cannot. Gender, age, and genetics are examples of inborn predispositions that cannot be changed. These fixed risk factors will be the primary drivers of cardiovascular disease. Risk variables that can be altered by individual action are called modifiable risk factors. Habitual behaviours, stress levels, dietary intake, and other biochemical and environmental variables are all changeable risk factors. Coronary heart disease, atherosclerosis, rheumatic heart disease, congenital heart disease, myocarditis, angina, and arrhythmia are all different forms of heart disease. Factors that increase vulnerability to cardiovascular disease are known as risk factors. In addition, these risk factors may hasten the progression of an existing illness. Factors that increase a person's likelihood of developing cardiovascular disease include 1) smoking and 2) a history of cardiovascular disease in the family. three) hypertension; four) cholesterol; five) diabetes;

six) obesity; seven) inactivity; and eight) stress. Heart disease is a broad category that includes any condition that affects the heart or its blood vessels. 'Cardio' is the medical name for heart[14][9]. Therefore, we lump them all together under the umbrella term cardiovascular disease. Obesity, high blood pressure, low HDL cholesterol, diabetes, and high mortality rate are all linked to a metabolic cluster that becomes severe enough to cause cardiovascular disease and obesity. The most important socio-psychological factor is the level of schooling and the persistence of depressive symptoms[15].

Various Prediction Methods for Heart Disease

In the event of a heart attack, prompt medical intervention is crucial for minimising potential cardiac damage and prolonging the life of the patient. In order to help heart disease patients recover, doctors are increasingly using cutting-edge technology to keep continual tabs on their medical records and provide ongoing guidance. As a result, there has been a rise in the collection and analysis of medical data with the use of computers. Heart disease diagnosis and prognosis in hospitals may be achieved using a number of different methods. The healthcare industries have a wealth of data at their disposal, most of which includes conceptual information that can be used to make informed choices. Data mining methods should be heavily used so that the right choices may be made with the available information. The illness prediction model was built using data mining methods, and it uses a variety of factors to identify whether or not a certain individual is experiencing a heart attack. The same holds true for predicting cardiac disease, which shortens the time it takes to do so, increases the accuracy with which the condition is diagnosed, and decreases the frequency with which heart attacks occur. By evaluating the data, data mining methods may shed light on the past and provide glimpses into the future. Machine learning, database technology, and AI are just few of the many disciplines that have come together to form these data mining methods. There are many uses for data mining methods, but one of the most significant is its role in early illness prediction[16]. The overall method for forecasting heart disease is analogous to the feature extraction procedure used to extract the important facts from the massive quantity of data. The second step is to train the extracted data using the chosen dataset and then feed that data into the testing phase. Various kind of categorization techniques are used for this purpose. Knowledge Discovery in Data is another name for this approach. Heart disease forecasting makes use of a wide range of methods. Decision trees, Support Vector Machines, Neural Networks, and K-Nearest Neighbours are the four most used data mining methods. In addition to these measures, several more are used in the process of predicting cardiovascular disease.

Literature Review

J. Rethna Virgil Jeny et.al.,(2020) Numerous individuals suffer with heart disease (HD), an umbrella term for a variety of various cardiovascular disorders. Numerous individuals deal with cardiac problems. The United States continues to have the highest rates of heart disease deaths worldwide. Cigarette smoking, excessive body fat, alcohol use, high blood cholesterol levels, high blood pressure, etc. are the primary causes of heart disease. Algorithms that collect data in the form of datasets are increasingly useful in the healthcare business, where they are put to use in a wide range of cutting-edge applications. Four different machine learning classification methods were employed in this work. The first thing we did was to implement a Support Vector Classifier (SVC). Is put to use in making forecasts about the model in question. Secondly, we have LR, or Logistic Regression. It's purpose is to characterise categorization issues according to given inputs. The Naive Bayes (NB) classifier comes in third, followed by the DT Algorithm. To improve precision and efficiency, we experimented with a variety of classifiers. Our approach uses machine learning classifier methods to increase diagnostic precision and speed up the process of diagnosing cardiac disease.

DOI : https://doi.org/10.56452/7-2-551

Shuying Shen et.al.,(2021) Predicting cardiac disease accurately has the potential to save thousands of lives and dramatically reduce health care costs. The accuracy-cy of our predictions may be improved by combing through data from various sources. However, the usefulness of many sources is not taken into account by existing machine learning-based prediction systems. Combining data from four sensors and an EMR, this paper uses support vector machines (SVMs) and a convolutional neural network (CNNs) to make predictions about heart disease. Each of the four sensors—the medical sensor, the activity sensor, the sleeping sensor, and the emotion sensor—uses its own unique feature extraction methods. Analyses show that using the suggested strategy improves the precision with which cardiovascular disease is predicted.

Rahul Katarya et.al.(2020) It has always been a crucial and difficult responsibility for medical professionals to recognise and predict cardiac disease. Expensive treatments and surgeries are available at hospitals and other clinics to address cardiac ailments. Therefore, it would benefit people everywhere if heart disease could be predicted in its early stages so that preventative measures could be taken before the condition worsened. The biggest causes of heart disease nowadays are unhealthy lifestyle choices including drinking alcohol, smoking cigarettes, and not getting enough exercise. The vast amounts of data generated by the healthcare sector allow machine learning to make accurate judgements and forecasts over time. This heart disease prediction makes use of many supervised machine learning approaches, including artificial neural networks, decision trees, random forests, support vector machines, naive Bayes, and the k-nearest neighbour algorithm. Additionally, a summary of each algorithms' respective performances is provided.

The worldwide increase in death rate is mostly attributable to heart disease. An early and precise diagnosis of heart disease might significantly cut down on deaths caused by the condition. According to the World Health Organisation, 17 million people worldwide lose their lives to heart disease annually. Artificial intelligence (AI) is crucial in making accurate and timely cardiac disease predictions. Cost savings, improved patient satisfaction, earlier detection of cardiac disease, higher-quality care, and individualised therapy are just some of the many benefits of AI. As more variables become available, feature or attribute selection has become a primary focus in most applications. Classifier models' output precision may be improved by careful feature selection. In addition, it is crucial when looking for relevant subsets to use in feature prediction. A clinical practitioner, for instance, can employ feature selection methods from classification approaches to arrive at a judgement about the severity of sickness. Optimising feature selection improves prediction accuracy. These days, evolutionary algorithms are used whenever a top performance is required. Traditional feature selection approaches like chi-square, data pick-up, and shred data work well for reducing the number level but have accuracy difficulties. In addition, the efficiency of the system will be impacted by the emergence of Non-Polynomial (NP) hard problems brought about by the incorrect selection of features. The size issue has been fixed, and classification execution time has been reduced by switching the final destination to a survey label that may be climbed. The most important aspect of this study is the feature selection that initiates the feature selection technique's goal of detecting cardiac problems. Data mining pre-processing is often used in the grey wolf and fire fly procedures to eliminate or greatly decrease noisy data. Using differential evolution, the grey wolf and firefly algorithms will analyse the characteristics from the chosen data to get the competent ones. Next, the GF-DE approach evaluates the feature selection, which is crucial since it allows for the rapid, optimal selection of features and the efficient upkeep of those features. One kind of meta-heuristic approach designed to improve ANN performance is HNN-Heuristic optimising the weight by artificial neural network. Technically speaking, ANN promotes brain-like conduct in order to improve the system, and it aids in keeping prediction-related issues under control. The ANN is made up of many different neurons, and in order to link them together, we'll be making use of a special kind of weight known as synaptic weight. Back propagation is used to update the ANN weights based on an evaluation of the mistakes that are gained from the initial data set created using ANN principles. There is a lag in ANN's learning rate because of its complexity and its many uses. By gradually expanding their training space, ANNs learn to use back propagation by fine-tuning the change in weights at each training cycle. Evidently, training the ANN using the back propagation algorithm takes some time, and this is crucial in advancing accuracy prediction in ANN in order to inspire the development of a new method for optimising performance in terms of both accuracy and speed. Dimensionality reduction will be accomplished by feature selection or feature extraction. The random selection of features to reduce the feature set size is the primary focus of this study, which uses the differential evolution-based grey wolf and firefly algorithms. A subset will be chosen using the computed methods, and its correctness will be verified using the original data set's classifiers. From datasets, the method computes a population size and exemplifies processes of selection, crossover, mutation, and extinction. To acquire the set of trial and error or to use the other approach, such grid or random search, the hyper parameters are employed. When the amount of the hyper parameter is large, this approach is shown to be challenging and boring. Consequently, there has been a rise in interest in finding the ideal hyper parameter and creating a system to track it. The sample size is then increased by iteratively refining the mutation selection and operator of cross over until the terminal scenario is reached, also known as the optimal configuration. When it comes to acquiring a large hyper-parameter space, these methods are simple, effective, and paralysed. Therefore, the primary focus of this study is to combine the grey wolf and firefly algorithms based on feature selection obtained through differential evolution, and then to use an improved classification technique to obtain an accurate heart disease prediction. In order to determine the efficacy of the proposed model, it will be tested on two datasets, namely Cleveland and Statlog. This is useful for determining how well the proposed model can predict cardiac disease in various datasets.

Proposed Approach

Figure 1 depicts the suggested method for identifying heart disease. The heart disease dataset will be given as the primary input for preprocessing. The given dataset will be standardised via the pre-processing procedure for generalizability and noise reduction. In this procedure, the quality of the dataset is improved through pre-processing, and then feature selection is achieved through the application of the proposed algorithm, GF-DE, which is a hybrid of the grey wolf and firefly algorithms with the aid of differential evolution. In place of unnecessary characteristics, the integrated grey wolf with firefly algorithm chooses the best ones. After that, categorization procedures are used. Here, the chosen features will be labelled using the updated ANN weight according to the modified hyper parameter. Multiple classifiers, with and without a feature selection strategy, will be compared to the proposed method.

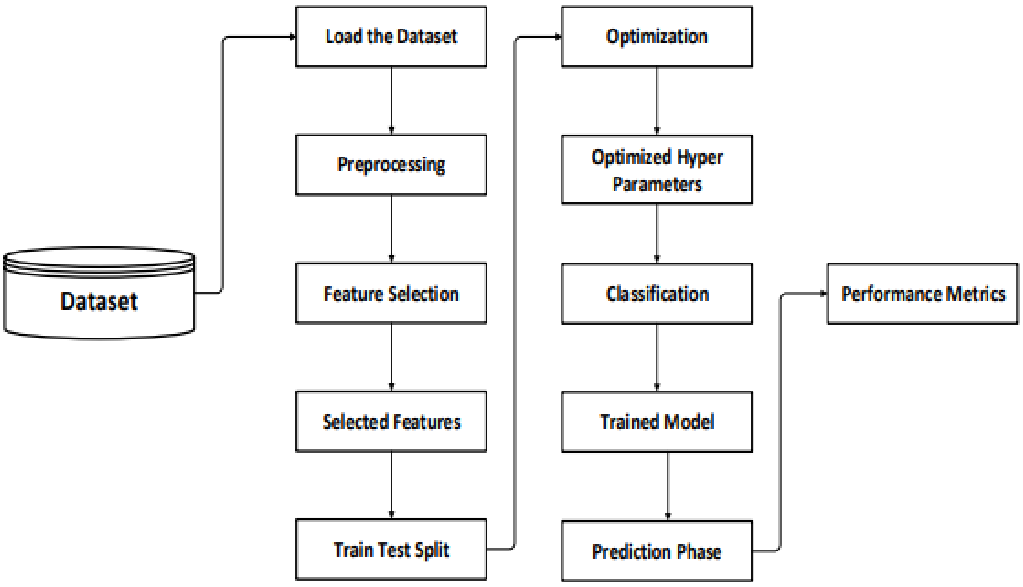
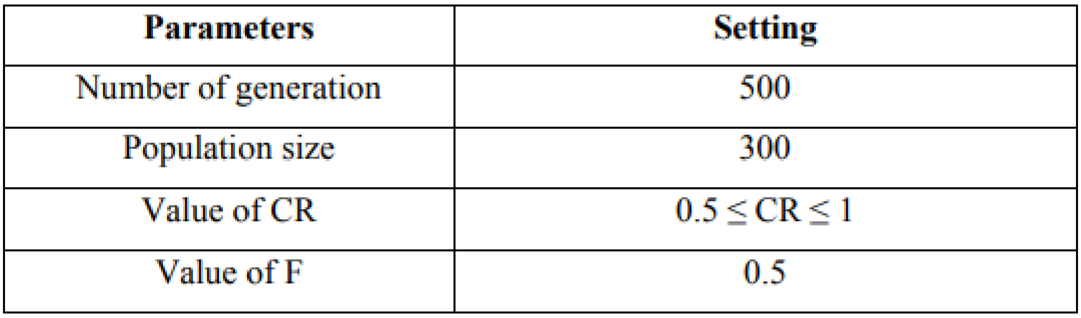


Figure 1 Flowchart of the proposed methodology

Differential Evolution (DE) for Optimal Parameters

DE was created by Price and Storn to assist with real-value optimisation problems. The DE is a search strategy that uses a random approach. DE's method is user-friendly in that it requires little input and quickly converges on a solution while being resilient and adaptable. A quicker DE convergence result will also lead to a greater searching chance for a regional optimum and an earlier time of convergence.

This DE is analogous to the process of mutation, in which an extra sum is obtained by comparing the genetic makeup of two people picked at random from the present population. The user is represented by the factor F, and the decision variables N and CR describe the problems that must be solved. The data for this is shown and summarised in Table 1.

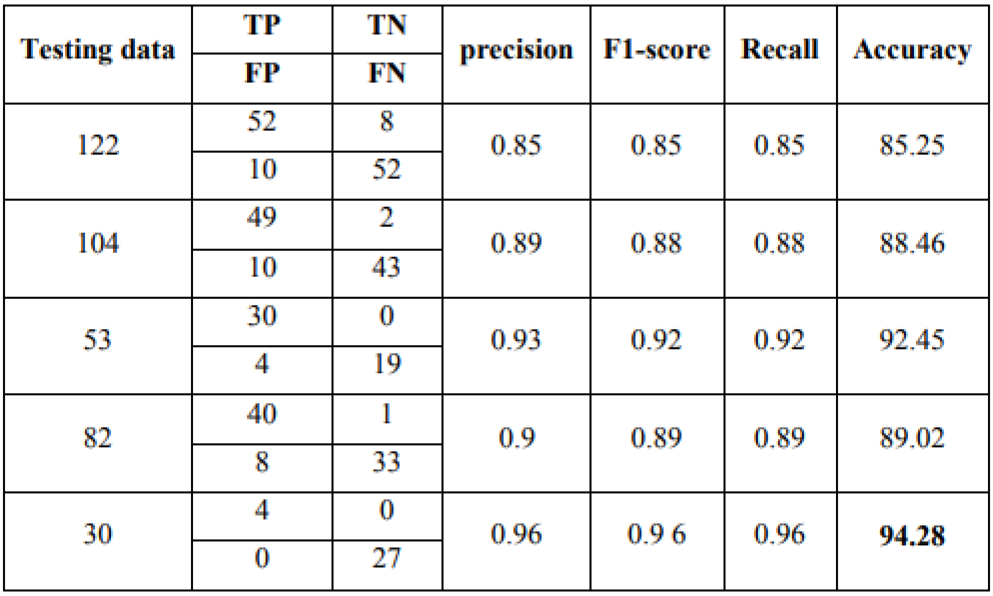


Parameters	Setting
Number of generation	500
Population size	300
Value of CR	$0.5 \leq CR \leq 1$
Value of F	0.5

Table 1 The factors given by the users

In this case, the suggested model is used to foretell whether or not a certain performer would get coronary disease. In order to accurately categorise as many positive samples as feasible, we need to improve the prediction rate in order to provide a label to the person who has heart disease. That's why we're putting all our eggs in the F1 Score basket.

Table 2. Performance of the prediction system based on the testing data the proposed systems



Testing data	**TP**	**TN**	**precision**	**F1-score**	**Recall**	**Accuracy**
	FP	**FN**				
122	52	8	0.85	0.85	0.85	85.25
	10	52				
104	49	2	0.89	0.88	0.88	88.46
	10	43				
53	30	0	0.93	0.92	0.92	92.45
	4	19				
82	40	1	0.9	0.89	0.89	89.02
	8	33				
30	4	0	0.96	0.9 6	0.96	**94.28**
	0	27				

Based on 122 test results, the percentages of real negatives and positives are 8 and 52, respectively. Similar results are obtained for the false-negative and false-positive rates: 52 and 10, respectively. The suggested model achieves similar results with 104% testing data: a true positive rate of 49.0%, a false positive rate of 10.0%, a negative rate of 0%, a negative rate of 43.0%, a precision of 0.89, an F1-score of 0.88, a recall of 0.88, and an accuracy of 88.46. Afterwards, with 53% testing data, the findings were 30/4 for true positives, 0/19 for false negatives, 0.93 for precision, 0.92 for F1-score, 0.92 for recall, and 92.45% for accuracy. The results of the remaining 82% of the tests showed a precision of 0.90, an F1-score of 0.89, a recall of 0.89, and an accuracy of 89.02%, with a true positive/false positive ratio of 40/8 and a true negative/false negative ratio of 1/33. Then, 30% of samples were used for testing, yielding false negative and positive rates of 27 and 0%, respectively, and a genuine negative and positive rate of 0%.Accuracy, precision, recall, and F1-score are all calculated to be 0.96, for a total of 94.28%. That's why in 30% of the test data, we get a better accuracy range. Thus, the results of the experiments have shown that the proposed technique is more accurate than the current methods in predicting the outcomes of specific parameters.

Figure 2 shows accuracy testing results for the suggested procedure.

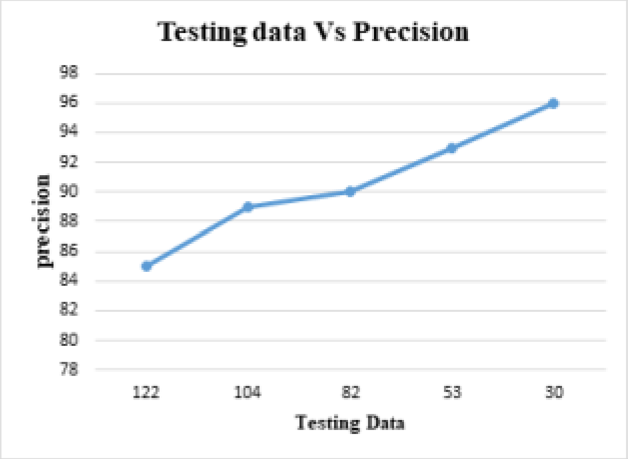


Figure 2 Analysis of precision for the proposed methods

Conclusion

In this study, we use a number of machine learning and data mining categorization methods to the problem of accurately predicting cardiac conditions. One-third of all fatalities in the world are attributable to cardiovascular disease, making it a subject of intense interest in modern medicine. The suggested system used this data in an effort to develop a reliable model for predicting (via data mining and analysis) whether or not patients suffer from this condition. Predicting cardiovascular disease as precisely as possible is important to this effort. This dataset was gathered from the UCI ML data archive. Predictions of heart disease are made using the Cleveland database. Here, a data mining classification approach was used, with the Logistic Regression dataset preprocessed, and the Sklearn package analysing the resulting score. The proposed method, logistic regression, has a 96.89% success rate.

References

1. J. R. V. Jeny, N. S. Reddy, P. Aishwarya and Samreen, "A Classification Approach for Heart Disease Diagnosis using Machine Learning," *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, Solan, India, 2021, pp. 456-459, doi: 10.1109/ISPCC53510.2021.9609468.
2. S. Shen, "A Multi-source Based Healthcare Method for Heart Disease Prediction by Machine Learning," *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, Stanford, CA, USA, 2021, pp. 145-149, doi: 10.1109/CONF-SPML54095.2021.00036.
3. R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 302-305, doi: 10.1109/ICESC48915.2020.9155586.
4. P. C. Kaur, "A Study on Role of Machine Learning in Detectin Heart Diseas," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 188-193, doi: 10.1109/ICCMC48092.2020.ICCMC-00037.
5. A. Nikam, S. Bhandari, A. Mhaske and S. Mantri, "Cardiovascular Disease Prediction Using Machine Learning Models," *2020 IEEE Pune Section International Conference (PuneCon),*

Pune, India, 2020, pp. 22-27, doi: 10.1109/PuneCon50868.2020.9362367. K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018, pp. 910-914, doi: 10.1109/ICECA.2018.8474918.

6. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
7. A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, Fez, Morocco, 2019, pp. 1-5, doi: 10.1109/WITS.2019.8723839.
8. N. C. Pereira, J. D'souza, P. Rana and S. Solaskar, "Obesity Related Disease Prediction from Healthcare Communities Using Machine Learning," *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, 2019, pp. 1-7, doi: 10.1109/ICCCNT45670.2019.8944798.
9. D. Krishnani, A. Kumari, A. Dewangan, A. Singh and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Kochi, India, 2019, pp. 367-372, doi: 10.1109/TENCON.2019.8929434
10. K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857.
11. S. Ganiger and K. M. M. Rajashekharaiah, "Chronic Diseases Diagnosis using Machine Learning," *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, Kottayam, India, 2018, pp. 1-6, doi: 10.1109/ICCSDET.2018.8821235.
12. B. D. Kanchan and M. M. Kishor, "Study of machine learning algorithms for special disease prediction using principal of component analysis," *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, Jalgaon, India, 2016, pp. 5-10, doi: 10.1109/ICGTSPICC.2016.7955260.
13. I. Yekkala, S. Dixit and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Bengaluru, India, 2017, pp. 691-698, doi: 10.1109/SmartTechCon.2017.8358460.
14. F. D. Mulla and N. Jayakumar, "A Review of Data Mining & Machine Learning approaches for identifying Risk Factor contributing to likelihood of Cardiovascular Diseases," *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2018, pp. 631-635, doi: 10.1109/ICICT43934.2018.9034257.
15. P. S. Kumar and S. Pranavi, "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics," *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, Dubai, United Arab Emirates, 2017, pp. 508-513, doi: 10.1109/ICTUS.2017.8286062.
16. K. Farooq *et al*., "A novel cardiovascular decision support framework for effective clinical risk assessment," *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, Orlando, FL, USA, 2014, pp. 117-124, doi: 10.1109/CICARE.2014.7007843.