Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.47750/jptcp.2023.30.11.018

Web-Based Machine Learning Algorithms for Personalized Cardiovascular Disease Risk Assessment

Shanmugaraj G[1*], Vetri Velan B[2], Senthil Kumar T[3], Surendar R[4]
[1,2,3,4]Velammal Institute of Technology, Chennai,India
***Corresponding author:** Shanmugaraj G, Velammal Institute of Technology, Chennai, India, Email: gsraj76@gmail.com

Submitted: 11 March 2023; Accepted: 10 April 2023; Published: 09 May 2023

ABSTRACT

Heart disease is the leading cause of death globally, claiming a life every minute. Early detection of heart disease can be challenging, but machine learning can accurately identify illnesses in the healthcare sector. This study used medical databases to analyze various heart disease situations. The data was analyzed using Python and the Random Forest algorithm. By predicting future patients based on past patient data, lives can be saved. The study developed a reliable method for predicting heart disease using the Random Forest algorithm. A technique was employed to utilize patient data from a CSV file and create a useful forecast of the likelihood of a heart attack. The prediction tool is web-based and enables users to input their information to determine their risk of developing heart disease. The method has numerous advantages, including high success rates, good performance and accuracy rates, as well as flexibility and adaptability.

Keywords: *ML ,Web-Based, Accuracy, CSV, Random forest, Datasets*

INTRODUCTION

The heart is vital to body because it transports arterial blood to all of organs. Blood that has been deoxygenated and is carrying metabolic byproducts from the body is gathered by the heart and sent towards the lungs for oxygen provision. The appropriate heart function results in a healthy life; on the other side, when the heart doesn't function regularly, it results in human death. Blood vessel irritation results from it. Improper blood flow caused by heart issues can be extremely harmful for patients [1].

There are several factors, such as diabetes, high blood pressure, excessive cholesterol, and smoking can lead to coronary heart disease. At the moment, smoking is a common habit among both young and elderly people, and it is also developing among teens. Smoking causes constriction of the heart's arteries and leads to arrhythmia or irregular heart rhythms. Blood pressure is also increased. High blood pressure can lead to a number of issues. It thickens the lower-left coronary heart chamber and makes the coronary heart has to exert more effort to circulate blood throughout body due to smoking, increasing risk of coronary artery disease. The nerves in your body or the blood arteries that regulate your heart may be impacted by high blood sugar levels. Cardiovascular disease is brought on by excessive cholesterol buildup within the cellular walls of arteries that have developed atherosclerosis. High cholesterol causes arteries to harden cardiac condition.

Blood flow is reduced and obstructed when the vessel narrows. These are the most prevalent and widespread viral diseases. Nowadays, four out of five people are illness[2].While some coronary heart conditions are extremely dangerous to health and may even result in death, in the past, many were unaware of the significance of coronary heart conditions or even that such conditions even existed. At the moment, coronary heart failure is the main cause of death. Every day, a significant number of individuals succumb to coronary heart failure, resulting in loss of life. According to World Heath Organizations calculations, 32% among all fatalities worldwide in 2019 were due to cardiovascular illnesses. [3]. Cardiovascular diseases, such as heart disease and stroke, accounted for 85% of all deaths. This result people's lack of knowledge on the causes of coronary heart ailments.

In emerging countries, the loss of skilled medical personnel and advancements in examination technology have made it increasingly challenging, time-consuming, and complicated to diagnose coronary heart disease using standard clinical techniques. As a result, the world is going through major difficulties. Doctors may recommend a number of procedures, including blood screening, electrocardiograms (ECG), workout stress tests, coronary angiograms, echocardiograms (ultrasound), and nuclear cardiac strain tests, depending on the patient's health. These diagnostic procedures are carried out only after a comprehensive evaluation of the patient's medical records and an analysis of their symptoms.

To detect irregularities in a person's heartbeat and diagnose coronary heart failure, doctors use an ECG machine that records the electrical impulses of the heart by attaching small sticky dots on the hands, legs, and chest. On the other hand, when doctors want to evaluate the heart's functioning, they recommend an MRI, which employs magnets and radio waves to generate precise images of the heart, which are then saved on a computer. Following a heart attack, a diagnostic test known as coronary angiography is performed. This involves inserting a catheter into the arteries in the wrist, arms, or groin and maneuvering it through the blood vessels. While the procedure is being done, The doctor takes a chest X-ray of the heart's chambers to detect any clogged arteries. [4].

Despite the use of various diagnostic methods, these techniques have their limitations and cannot provide a complete understanding of the intricacies of a medical condition. Recent global research conducted in 2019 revealed that coronary heart disease has become the leading disease worldwide, highlighting the critical role of a healthy heart in maintaining overall bodily function, Should it malfunctions, an individual will die as a result. Due to the slow diagnosis and limited medical expertise in treating coronary heart disease, the cost of treatment can be quite high, which underscores the growing importance of early detection and prompt intervention as time progresses. Modern diagnostic methods are unable to foresee the complete cause of the condition, unlike the ECG, which can occasionally be insufficient to detect cardiovascular illness disease. Further screening are necessary to detect cardiovascular disease. A coronary angiography may affect a convalescent's renal function among other things, it might harm the arteries and create allergies.

Clinical techniques play a critical role in diagnosing coronary heart disease, aiming to determine whether a patient is either not following appropriate practices or suffering from a heart issue. One method that has gained attention is machine learning[16]. It involves the use of algorithms[20] and statistical models to analyze data without explicit instructions, enabling the creation of computer systems that can analyze vast amounts of information. Data preparation is a information collection method ,converts irrelevant info into valuable information. Data cleaning, the first phase in the preparation of data, including replacing missing data with the value that is most likely to be true and eliminating noisy data employing grouping, regression, other splitting techniques. The following phase, data transformation, involves selecting the appropriate attributes and converting the data into a usable format by normalizing it. After data preparation, final step is information reduction, which includes aggregation of data cubes and selection of attribute subsets. There are various techniques

J Popul Ther Clin Pharmacol Vol 30(11):e153–e161; 09 May 2023.

used in data reduction, including numerosity reduction and dimensionality reduction .Algorithms are utilized to train data and test its accuracy in classifying patients as either at chance of developing cardiac disease. This approach is efficient in terms of conserving memory and resources [5].

LITERATURE SURVEY

Heart diseases are a leading cause of death, and researchers have studied various.Numerous studies have focused on diagnosing heart diseases due to their status as a leading cause of mortality. Researchers have used various approaches for diagnosing heart diseases, and the probabilities they discovered varied. Data mining and classification algorithms were used by researchers . Decision trees, Logistic Regression, Support Vector Machines, K-Neighbour classification and Random Forests are used to predict cardiac diseases. In an experiment using a dataset, a researcher created a model using neural networks[15] and hybrid intelligence techniques. The outcomes demonstrated that these techniques produce the best results and enhance prediction accuracy. Machine learning[17]-[19] algorithms are extremely useful in some methods and can predict risk before it occurs.

Researchers have employed various machine learning methods, such as Decision Trees, Logistic Regression, Support Vector Machines, K-Nearest Neighbor classification, and Random Forests, to forecast the occurrence of heart disease. The study reveals that SVM [6]provides the highest accuracy of approximately 81% followed by Decision trees (85%), Random Forest(86%), K-Neighbour classification(63%) and Logistic regression (85%). The best classification strategy for predicting cardiac disease, according to the experts, is the method of backpropagation [7]. The natural algorithm optimizer, which they suggested as a replacement for the backpropagation method, has the drawback of becoming trapped in local minimums, they discovered. They claimed that by employing this strategy, future results will be 100% accurate and error-free.

S. Prakash et al. conducted a study in 2017 to compare the two techniques, Rough Set Feature Selection on Information Entropy (RSFS-IE) and Optimality Criterion Feature Selection (OCFS), for predicting heart disease. The study included various types of datasets to evaluate factors such as prediction accuracy, speed of calculation, and error rate. Based on their findings, the researchers concluded that OCFS is a better method compared to RSFS-IE as it operates faster.[8]

Researchers conducted a study using a patient record database for developing neural network in diagnosing heart disease. The network was trained and tested using 13 factors, such as age, blood pressure, and angiography results. The researchers recommended using backpropagation training and supervised networks for optimal results[9]. The system was able to recognize unknown data and compare it to trained data when a doctor entered unknown data, creating a list of potential diseases that could threaten a patient. The output produced by the system was highly accurate. Kim and Kang [10] created a neural network-based approach for diagnosing heart problems. They conducted sensitivity analysis on the attributes to identify the most important features for diagnosis. They determined that features with a high sensitivity level of greater importance compared to with low sensitivity level. After eliminating unnecessary qualities, related features were discovered by comparing the responsiveness of the features to changes within a single feature's worth.

A research study assessed various computational models to determine their effectiveness in predicting the probability of heart disease, and it was found that regression was the most effective method [11]. The study involved a dataset containing 1000 values, which was split into two halves. During the testing process, 70% of the data was employed for training purposes, while the remaining 30% was used for testing. Regression was found to be superior to other models based on the results. Mafizur Rehman conducted an independent study in 2020, where the Random Forest method was employed to forecast heart disease with a precision rate of more than 97%.[12].

Proposed Work Limitation Of Previous Work And Our Contribution

Table 1 displays the previous research conducted by various researchers who have successfully predicted cardiovascular disease by various methods.

TABLE 1: Evaluation of earlier techniques and findings

Approach	Year	Method Used	Results
Vincy Cherian et al	2017	Naïve Bayes	86%
Rani et al	2021	Logistic Regression	86.60%

The researchers faced some limitations in their previous studies. This article proposes a solution to overcome these limitations by using support vector machines (SVM) to predict cardiovascular disease. A machine-learning technique called SVM have successfully used on various biological applications. The article also introduces machine learning approach to predict heart disease. This approach utilizes SVM, which is a powerful machine learning technique that can address complex categorization problems in bioinformatics.

METHODOLOGY

The article utilizes the Random Forest technique to predict coronary heart disease. This area of research is currently quite active in the medical field and is expected to be extensively used in the future in biomedical systems. Random Forest is set of supervised learning methods which analyze both simple and complex datasets. One of the advantages of Random Forest is its ability to effectively operate in high-dimensional spaces, without the need for linearity. This makes it a valuable tool for disease diagnosis and community screening, as no specific standards need to be followed to diagnose the condition[13]

The emerging approach of federated learning offers a potentially beneficial alternative to traditional machine learning methods in terms of security and cost savings. This approach involves submitting each dataset to a single server [14]. Decentralized edges and servers are then utilized to train the data, with each storing local data samples confidentially.

Preprocessing

The technique for predicting coronary heart disease described in the article begins with data preparation. A suitable dataset obtained from Kaggle is used for this purpose. The preprocessing phase involves cleaning and transforming the raw data into a format that can be used to train machine learning models in ML. Data reduction, transformation, and purification are all integral parts of the preprocessing process. Data reduction involves reducing the amount of data, while data transformation involves normalization and aggregation. Accounting in absent numbers, smearing noisy information, and eliminating outliers are all parts of data cleaning. Multiple Kaggle datasets are utilized in the preparation phase.During the data preprocessing phase, some data may be lost or removed in order to eliminate inefficiencies and increase accuracy. The final machine learning model's accuracy is significantly influenced by the extent and quality of the dataset used.

Dataset description

The researchers selected a dataset from Kaggle that contains 919 rows and 12 columns which are shown below . The coronary heart disease column contains values of "1" for patients with a current diagnosis of the disease and "0" for patients who do not have the disease. However, the dataset is unbalanced, meaning that there are more samples of one class than the other. Therefore, the researchers applied preprocessing techniques to the dataset to balance it. Table 2 provides a detailed description of the dataset. Table 3 provides a sample data.

TABLE 2: Attributes & Discerption of Heart Disease Dataset.

AGE	Age in years
SEX	Male (or) Female.
CP	Chest Pain
REST. BPS	Resting blood pressure
S. CHOL	Serum cholesterol in mg/dl
FBS	Fasting blood sugar
REST. CG	Resting Electrocardiographic Results
M.HAACH	Maximum heart rate achieved
EX. ANG.	Exercise Induced Angina
OLD PEAK.	ST depression induced by exercise relative to rest
SLOPE ST	The slope of the peak exercise ST segment
CA.	Number of major vessels (0-3) colored by fluoroscopy
THAL.	0 = normal; 1 = fixed defect; 2 = reversible defect
TARGET	refers to the presence of heart disease in the patient (1=yes, 0=no)

TABLE 3: Sample of Dataset

AGE	52	58	71	43
SEX	1	0	0	0
CP	0	0	0	0
TRESTBPS	125	100	112	132
CHOL	212	248	149	341
FBS	0	0	0	1
RESTECG	1	0	1	0
THALACH	168	122	125	136
EXANG	0	0	0	1
OLDPEAK	1	1	1.6	3
SLOPE	2	1	1	1
CA	2	0	0	0
THAL	3	2	2	3
TARGET	0	1	1	0

PROPOSED MODEL

The Random Forest approach is suggested in this research as a way to increase the model's precision, which outperforms previous implementations and investigations

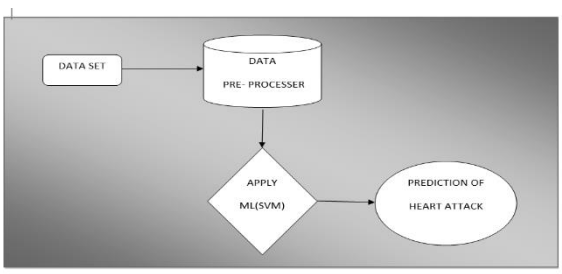


FIG. 1: Recommended Cardiovascular Disease Detection System

J Popul Ther Clin Pharmacol Vol 30(11):e153–e161; 09 May 2023.

Figure 1 shows that the dataset is preprocessed and labeled prior to the application of the trained Random Forest model to enhance accuracy, and the Confusion Matrix is utilized for analysis and validation of the results. The dataset is first preprocessed and labeled using a different format in this model, then the trained Random Forest model is applied to it for more accuracy, and finally, these results are evaluated confirming with the confusion matrix.

Implementation Detail

The experiments and analyses described in this paper were carried out using the following:

NumPy version: 1.22.4

Matplotlib version: 3.5.3

Python version: 3.8.10

Pandas version: 1.3.5

Stimulations and results

random forest

The chosen method for analysis in this case was Random Forest, a set of techniques used for supervised learning tasks including classification and regression. A discovery contractor was hired to execute this methodology. One of the advantages of Random Forest is that it can be applied in high-dimensional spaces, even when there are more dimensions than samples. The results of applying Random Forest on the dataset can be seen in Figure 2 below.

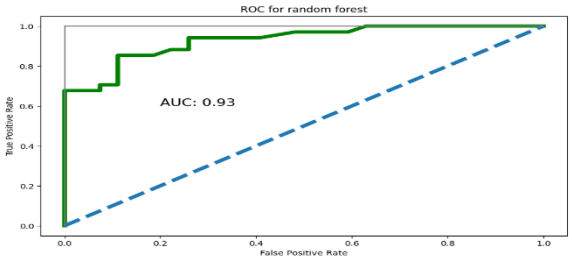


FIG. 2: ROC graph of Proposed Random Forest Trained Model

In Figure 3, the Random Forest algorithm's performance at different thresholds is presented using the AUC-ROC curve. The chart depicts the algorithm's capacity to differentiate between positive and negative instances, as it plots the true positive rate versus the false positive rate. The ROC graph is the underlying curve, while the AUC represents the class score or scale. This demonstrates the model's capacity to differentiate between multiple categories, with a higher AUC indicating a more accurate representation of 0 as 0 and 1 as 1. By the context of metrics, the model performs superior than the AUC at differentiating between people who have the illness compared to those that are healthy

Confusion Matrix

Confusion matrix makes it easy to understand yet powerful tool for evaluating model performance. It provides a comprehensive evaluation of the effectiveness of a classification model. It's a N x N matrix table, in which N is the total amount of target groups overall. The machine learning model's predictions are compared to the actual target value in this matrix, allowing for a detailed analysis of the model's performance and shortcomings. Fig 4 displays the results of the Confusion Matrix.

$Accuracy = (TP+TN) /P+N = (TP+TN) /TP+TN+FP+FN$

$$PPV = TP/(TP + FP)$$

$TNR = TP /TP+FP$

$TPR = TP /TP+FN$

$F1\ Score = 2 \times PPV \times TPR / PPV \times TPR$

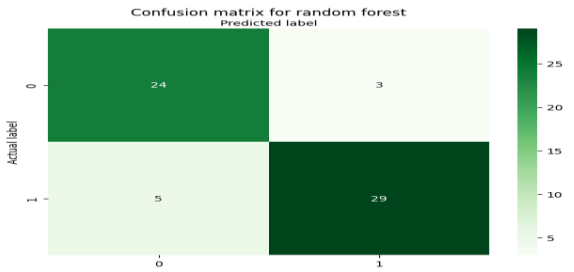


FIG 3: Random Forest Model's confusion matrix

Other Models Are

Decision Tree

Decision trees come in a variety of forms. The primary distinction is that the class attribute was given higher importance earlier in the decision

J Popul Ther Clin Pharmacol Vol 30(11):e153–e161; 09 May 2023.

tree building process. In an entropy-based system, the attribute chosen as the root of the decision tree uses information gain to decrease entropy. Before choosing the attribute with the greatest information gain as root of decision tree, the information gain of each attribute in the dataset is evaluated. After that, the characteristic utilising information gain will be given.

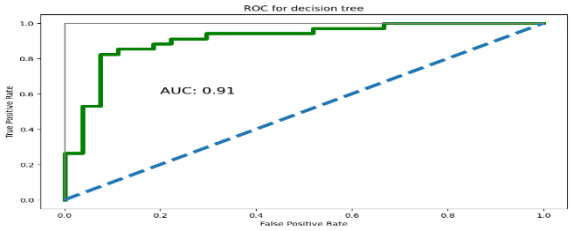


FIG. 4: ROC graph of Decision trees Trained Model

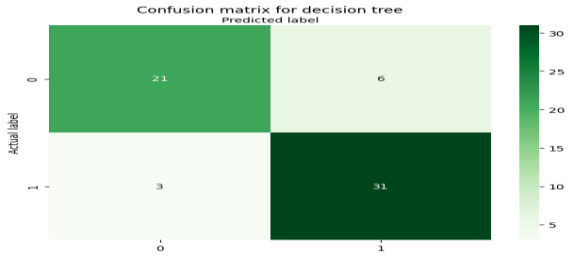


FIG .5: Confusion matrix of Decision trees Model

Logistic Regression

Another approach to machine learning includes logistic regression that is utilized to forecast binary results, such as "Yes" or "No" using one or more input variables.. The algorithm uses a logistic function to model the relationship between the input variables and the output. The logistic function transforms any real-valued input a number that ranges from 0 to 1. The algorithm finds best values for coefficients of the logistic function so that the predicted output matches the actual output for the training data as closely as possible. To make predictions, the logistic regression algorithm calculates a value called z from the input variables, which is then passed through the projected output, which is a probability between 0 and 1, using the logistic function. The projected output can be understood as the probability of the binary outcome being true based on the input variables. The logistic regression method provides a probability score between 0 and 1 for the anticipated binary outcome. If this probability value is greater than a predefined threshold , the algorithm predicts a value of 1 for the binary outcome, otherwise, it predicts a value of 0. Logistic regression is commonly used in various applications such as credit scoring, fraud detection, and medical diagnosis. Logistic regression can be applied in various scenarios, such as credit scoring, where it can forecast the probability of a borrower defaulting on a loan based on their credit history and other pertinent factors. Based on numerous transactional characteristics, logistic regression can be used in fraud detection to estimate the likelihood that a specific transaction is fraudulent. In medical diagnosis, logistic regression can estimate the probability of a patient having a specific ailment based on their medical history and symptoms.

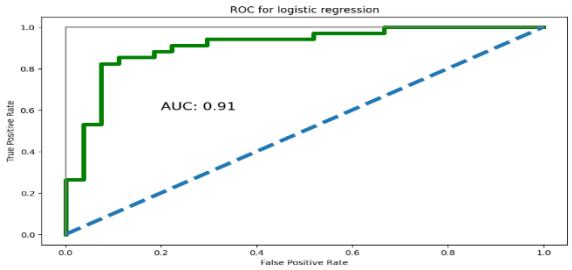


FIG.6: ROC graph of Logistic regression Trained Model

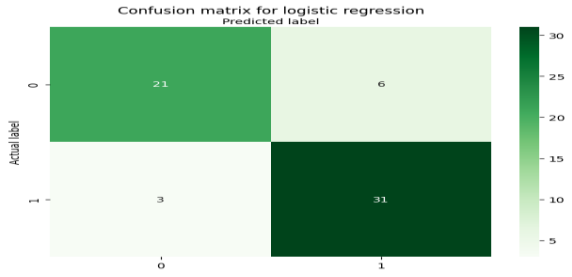


FIG.7: Logistic regression model's confusion matrix

K- Neighbors Classifier

This method of classification is one of the easiest and most useful ones. As the customer is unaware of any reliable constant controls for probability

J Popul Ther Clin Pharmacol Vol 30(11):e153–e161; 09 May 2023.

densities, understanding them during quality assurance is difficult. Thus, these kinds of computations are performed using the KNN classification algorithm. Using training datasets, the location of Knearest's neighbour is predicted. Euclidean distance is used to calculate how close the training dataset is to the objective. The k nearest neighbours should be given to the group of rows being analysed. Repeat the procedure for the unfinished rows in the target set. In this application, the largest value of K can be selected, and the software then automatically builds an identical parallel model on top of it

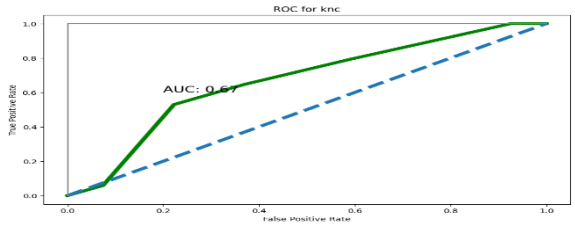


FIG.8: ROC chart for the trained KNN model

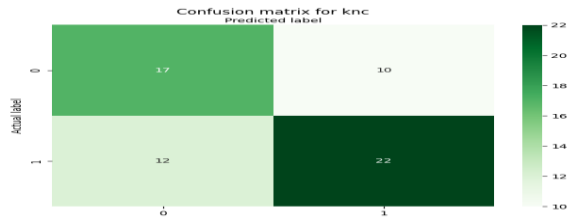
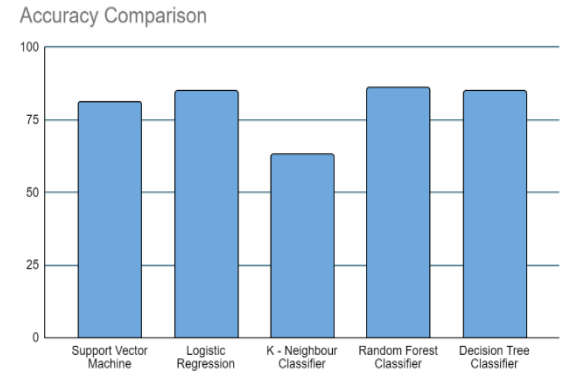


FIG.9: KNN Model's confusion matrix

Accuracy Comparison

The Accuracy of all five algorithms are compared and plotted in a Bar Graph.



CONCLUSION

Experiments have shown that the COVID-19 pandemic has caused heart damage in a significant number of people, making it crucial to develop an effective diagnostic method that can detect heart failure early and prevent fatalities. This necessitates a diagnostic approach that incorporates data from previously identified cases of cardiac disease and concentrates upon the number of cases of cardiac failure. In this context, an Random Forest approach was utilized to develop a model that has an accuracy of 90.47 percent. As the model receives more training data, its ability to correctly identify heart disease improves. Different methods can be used to break down the data, and their outcomes can be compared. Additional methods to combine trained model in ML and DL cardiac models with specialised multimedias have been developed to aid patients and doctors..The Random Forest algorithm is a potent ensemble learning method for classification and regression tasks. After building N decision trees, the method produces a class that reflects the average result of each decision tree. Early forecast accuracy is thus attained. The study of healthcare data, particularly data pertaining to the heart, will aid in the early diagnosis of heart disease or other aberrant cardiac disorders, saving long-term mortality. Predicting heart illness is a difficult task in the current world. By entering the report values into the web-based service, the patient or user can utilise this application to anticipate disease even if they are not close to a doctor.

REFERENCES

1. Sahoo, P. & Jeripothula, P. (2020) “Heart Failure Prediction Using Machine Learning Techniques” SSRN Electronic Journal.
2. Javeed, A., Rizvi, S., Zhou, S., Riaz, R., Khan, S. & Kwon, S. (2020) “Heart Risk Failure Prediction Using a Novel Feature Selection Method for Feature Refinement and Neural Network for Classification”.Mobile Information Systems 2020, 1-11.
3. Chicco, D. & Jurman, G. (2020) “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone.” BMC Medical Informatics and Decision
4. Shu, T., Zhang, B. & Tang, Y. (2017) “Effective Heart Disease Detection Based on Quantitative

Computerized Traditional Chinese Medicine Using Representation Based Classifiers." Evidence-Based Complementary and Alternative Medicine 2017, 1-10.
5. Goel, R. (2021) "Heart Disease Prediction Using Various Algorithms of Machine Learning" SSRN Electronic Journal.
6. Latah, C. & Jeeva, S. (2019) "Improving the accuracy of prediction of heart disease risk based on ensemble classification"techniques.Informatics in Medicine Unlocked 16, 100203.
7. Nanekar, G. (2021) "Heart Disease Prediction using Neural Network".International Journal for Research in Applied Science and Engineering Technology 9, 1907-1910.
8. Anon. (2022) "Improving Heart Disease Prediction Using Feature Selection Approaches" Ieeexplore.ieee.org.Https://ieeexplore.ieee.org/abstract/document/8667106/ [accessed1January 2022].
9. Gavhane, A., Kokkula, G., Pandya, I. and Devadkar, K., 2018,March."Prediction of heart disease using machine learning'. In 2018 second international conference on electronics, communication and aerospace technology (ICECA) (pp. 1275-1278). IEEE.
10. Rani, P., Kumar, R., Ahmed, N. & Jain, A. (2021) "A decision support system for heart disease prediction based upon machine learning". Journal of Reliable Intelligent Environments 7, 263-275.
11. Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., Singh, P. & kumar, N. (2021) "Latest trends on heart disease prediction using machine learning and image fusion. Materials Today". Proceedings 37, 3213-3218.
12. Pavithra M., M. (2022)" Effective Heart Disease Prediction Systems Using Data Mining Techniques" Annalsofrscb.ro. Https://www.annalsofrscb.ro/index.php/journal/article/view/2172 [accessed 2 January 2022].
13. Noble, W.S., 2006. "What is a support vector machine? Nature Biotechnology",24(12), pp.1565-1567.
14. Matveeva, N. (2021) "Artificial Neural Networks In Medical Diagnosis" System technologies 2, 33-41.
15. Anon. (2022) "Analysis of Neural Networks Based Heart Disease Prediction System". Ieeexplore.ieee.org. Https://ieeexplore.ieee.org/abstract/document/8431153 [accessed 9January 2022].
16. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). "Effective heart disease prediction using hybrid machine learning techniques". IEEE Access, 7, 81542-81554.
17. Bhatla N., & Jyoti, K. (2012). "An analysis of heart disease prediction using different data mining techniques". International Journal of Engineering, 1(8), 1-4.
18. Patel J.Tejal Upadhyay, D., & Patel, S. (2015)."Heart disease Prediction using machine learning and data mining technique." Heart Disease, 7(1), 129-137.
19. Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). "Heart disease prediction using machine learning techniques:" a survey. International Journal of Engineering & Technology, 7(2.8), 684687.
20. Sowjanya, K., & Krishna Mohan, G. (2020). "Predicting Heart disease using machine learning classification algorithms and along with tpot (Automl)". International Journal of Scientific and Technology Research, 9(4),3202–3210.