PREDICTING PRESENCE OF HEART DISEASES USING MACHINE LEARNING

Raushan Kumar[1], Chandini Singh[2], Tanu Kesharwani[3], Vijaya Choudhary[4*]

Abstract—
cardiac illness covers several cardiac issues. According to WHO statistics, cardiovascular illnesses kill 17.9 million people worldwide. This worrying rise in heart disease cases shows how important it is to find it early. The health of the patient is closely linked to early and accurate diagnosis, so this is a very important issue. To deal with this, programmes that use machine learning have been made to find and identify people with heart disease. In this project, we made a heart disease prediction system that uses a patient's medical background to figure out how likely it is that they will be diagnosed with heart disease. The method looks at things like chest pain, high blood pressure, heart attack, high cholesterol, and high blood pressure. Anaconda Python software has been used for machine learning algorithm implementation, experimental data processing, and visualisation. This project demonstrates how Python and machine learning can detect cardiac problems. This project used UCI data. It has 3000 cases, and each one has 14 different characteristics. These examples show the results of different tests done to predict how well heart disease can be diagnosed. The dataset has been split into two parts so that the success of the algorithms can be judged. Seventy percent of the data will be used for training, and the other thirty percent will be used for tests. The study uses SVM, KNN, LR, RF, and DT algorithms to enhance heart disease detection. The goal of these programmes is to be able to tell if a person has heart disease or not. With this study, we want to compare how well different models work and look at the results to improve how well heart disease can be found. Using machine learning techniques and Python code together could be a good way to deal with this world health problem. The accuracy is calculated and analysed using these values. It has 14 classification attributes. We found that random forest is performing well out of all five algorithms giving an accuracy of 98.05%, Decision tree is performing next to random forest giving an accuracy of 97.08%, Support vector machine is giving accuracy of 90.25%, K-nearest neighbour and logistics regression are giving accuracy of 81.82%.

Keywords— Heart disease, KNN algorithm, Decision Tree, Random Forest, Logistic Regression, SVM.

[1]raushan2002kumar@gmail.com
[2]chandini.20scse1010065@galgotiasuniversity.edu.in
[3]tanukesharwani570@gmail.com
[4*]vijaya.choudhary@galgotiasuniversity.edu.in

***Corresponding Author:** Vijaya Choudhary
*vijaya.choudhary@galgotiasuniversity.edu.in

DOI: - 10.48047/ecb/2023.12.si5a.0544

I. INTRODUCTION

cardiac illness covers several cardiac issues. According to WHO statistics, cardiovascular illnesses kill 17.9 million people worldwide. This worrying rise in heart disease cases shows how important it is to find it early. The health of the patient is closely linked to early and accurate diagnosis, so this is a very important issue. To deal with this, programmes that use machine learning have been made to find and identify people with heart disease.

In this project, we made a heart disease prediction system that uses a patient's medical background to figure out how likely it is that they will be diagnosed with heart disease. The method looks at things like chest pain, high blood pressure, heart attack, high cholesterol, and high blood pressure. Anaconda Python software has been used for machine learning algorithm implementation, experimental data processing, and visualisation.

This project demonstrates how Python and machine learning can detect cardiac problems. This project used UCI data. It has 3000 cases, and each one has 14 different characteristics. These examples show the results of different tests done to predict how well heart disease can be diagnosed.

The dataset has been split into two parts so that the success of the algorithms can be judged. Seventy percent of the data will be used for training, and the other thirty percent will be used for tests. The research aims to enhance heart disease detection utilising SVM, KNN, LR, RF, and DT approaches. The goal of these programmes is to be able to tell if a person has heart disease or not. With this study, we want to compare how well different models work and look at the results to improve how well heart disease can be found. Using machine learning techniques and Python code together could be a good way to deal with this world health problem.

The senior population is especially affected by this problem since many illnesses have similar symptoms. Consequences, including imminent death, may result from a delay in making an accurate diagnosis of an illness.

However, as time goes on, more and more information from studies and medical records becomes accessible. There are several publicly available data sets that include patient information, allowing for research into the efficacy of different computer systems in making correct diagnoses and detecting this illness early enough to save lives.

Notably, research from a number of prestigious institutions shows that cardiologists use a battery of diagnostic tools to accurately identify cardiac problems. These diagnostic procedures may include a number of tests, including but not limited to blood work and chest X-rays:

A type of electrocardiogram (ECG or EKG) is an EKG: An electrocardiogram is a test that measures the electrical activity in heart. It can find problems with the way the heart beats. You can get an ECG when you're at rest or when you're working out (called a "stress electrocardiogram").

Holter tracking : It is a way to keep track of how fast your heart beats. A Holter monitor is a portable ECG that you wear for 24 to 72 hours to record your heart rate. Holter tracking is used to find problems with the heart's beat that a normal ECG can't find.

Echocardiogram: Sound waves are used in this easy test to make detailed pictures of how your heart is built. It shows how your heart works to move blood through your body**.**

The importance of machine learning and AI in healthcare has been widely recognised in recent years. The source of this research is to apply machinery techniques to the quotation of cardiovascular disease diagnosis. This research zeroes focused on processing and computing tasks by means of the Python programming language and the Anaconda integrated development environment. Python was used for all classifier implementations, with the Pandas, NumPy, matplotlib, sci-kit learn (sklearn), and seaborn packages and libraries being particularly useful. Several machine learning algorithms were used to identify and categorise people with heart disease. An very useful strategy was developed to improve the precision of heart attack detection in people. Diseases may be diagnosed and outcomes identified and recognised using machine learning algorithms. These models can analyse genetic data in a comprehensive manner. Further, they allow for more in-depth study of medical data, leading to better forecasts, and they make it easier to train models for pandemic knowledge prediction.

Many research have been done to identify and detect heart disease diagnosis using various machine learning models. In order to identify patients at high risk for congestive heart failure, a machine-learning classifier has been created. This classifier was trained with an accuracy of 100% using a machine learning technique. Support vector machine, random forest, K-nearest neighbours

(KNN), decision tree, and logistic regression algorithms were among those compared in an effort to further improve performance.

II. LITERATURE SURVEY

Various Before the development of AI algorithms for diagnosing heart disease, many healthcare analytics research relied on statistical methods. However, because to its crucial relevance, advancement in this field has been considerably expedited via the coordinated efforts of engineers and clinicians. Here we look at the processes involved in cardiac illness diagnosis, and then we'll briefly address the use of AI in healthcare decision support systems using machine learning algorithms in other contexts.

We've assembled a timeline of research published in the last several years on how best to categorise heart disease. Asha Rajkumar et al. (2010) used Naive Bayes, Decision Tree, and KNN classification methods. Due to its faster accuracy computations, Naive Bayes emerged as the top performance in a comparison of these methods.

M. Srinivas, B. Raghavendra Kao, and C. Govardhan all used decision trees, Naive Bayes, and neural networks in the same year to identify heart disease based on 15 common risk factors found in the medical literature.

In 2012, T. J. Peter and K. Somasundaram examined the Naive Bayes classifier and Backpropagation Neural Network (BNN) for medical data mining classification. In supervised learning, the Naive Bayes classifier uses prior probabilities to derive an object's likelihood [2].In 2012, N. P. Sundre, D. P. Letha, and K. R. Chandrah calculated heart attack risk using Naive Bayes and a WAC [4]. S. A. Patkar and A. Parveen created a Naive Bayes algorithm-based intelligent web system in the same year to help physicians diagnose heart problems [5].

In 2013, M. Akhil Jabbar et al. tested Data Mining (DM) with a genetic algorithm to enhance classification performance, predicting cardiac disease with 95% accuracy [6]. S. Vijiyarani and S. Sudha used association rules to improve illness prediction in 2013 [7]. In the same year, S. U.

Amin, K. Agarwal, and R. Beg created a hybrid approach integrating genetic algorithms and neural networks to Using risk factors such age, familial obesity, elevated cholesterol levels, hypertension, drinking or smoking, and obesity, one can predict a risk of heart disease. [8]. Hlaudi Daniel Masethe et al. introduced a hybrid decision tree classification algorithm in 2014 with 98.14% accuracy [9]. In 2017, Neha Kunte et al. used the Simple Cart Classification method to accurately classify 11 medical practitioner qualities [10]. Septiani et al. used the kNN method to analyse 670 patients with 98% accuracy the same year [11].

Recent heart disease detection research have incorporated deep learning algorithms. Adyasha Rath et al.'s 2021 GAN-LSTM model has great accuracy, F1-score, and AUC [12]. KNN, DT, and RF reliably predicted cardiac disease with 100% accuracy in 2021, according to Mamun Ali et al. [13].

These papers demonstrate cardiac disease prediction methods and algorithms that enhance medical data analysis.

METHODOLOGY

This study utilizes several algorithms, namely random forest, support vector machine, K-nearest neighbour, logistic regression, and decision tree, to develop a model for detecting heart disease in a patient dataset provided by medical professionals. The primary objective of this research is to employ machine learning techniques to identify potential cases of heart disease and assess the performance of each algorithm in terms of accuracy. The overall methodology is presented in Figure 1, and each component will be discussed individually in this section. Prior to delving into the details, it is essential to introduce the dataset.

1. Data collection

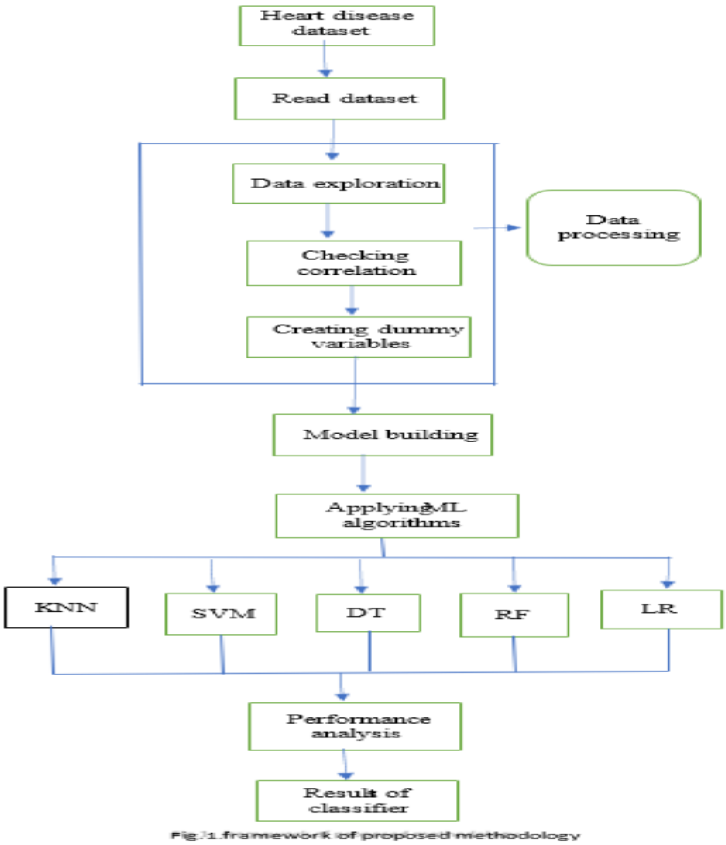


Figure 1:*Proposed method*

The heart disease detection method: Kaggle.com provided data. 11 characteristics are in the 70,000patient dataset. Any research project requires a dataset.

2.Data Pre-processing

Training, test, target, and feature data are segregated. Scaling data values to 0–1 before training Machine Learning models.

3.Correlation coefficient In Correlation coefficients are used in statistics to measure the closeness of a connection between two variables. They provide a numerical value to the linear connection between a dependent and an independent variable in statistical analysis. The letter 'r' is often used to denote the correlation coefficient. The correlation between each pair of columns in two datasets is calculated to assess how closely they are related. The analysis shows how strongly the columns in a dataset are related to one another. If the correlation values of two columns are equal, then one of the columns will be dropped to reduce data duplication.

$$r = \frac{m(\sum ab) - (\sum a)(\sum b)}{\sqrt{([m\sum a^2 - (\sum a)^2][n\sum b^2 - (\sum b)^2])}\ 'm'}$$

stands for 'quantity of information,' a for the total of the first variable value, b for the total of the second variable value, ab for the sum of the product of the first and second variable values, (a)2 for the sum of the square of the first value, and (b)2 for the sum of the square of the second value.

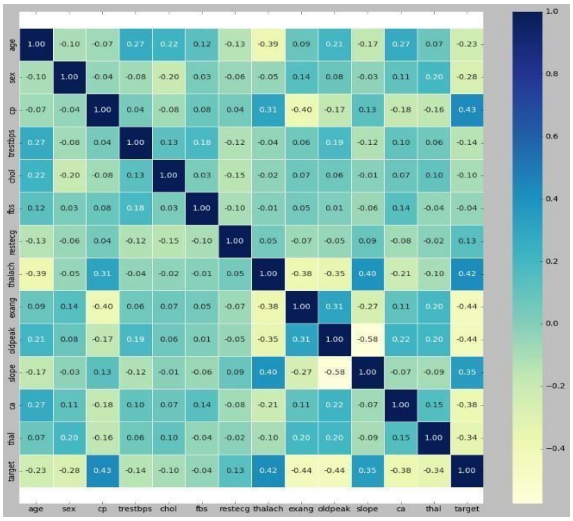


Figure 2:*. Correlation of each feature in the dataset using the heatmap*

The correlation matrix is a mathematical tool used to assess the strength and direction of a linear relationship between two variables. It provides a range of values from -1 to 1, indicating the degree of correlation between the variables. The

correlation matrix helps us understand how different coefficients are interconnected. It assumes that each value of a random variable is correlated. To easily identify and analyse these correlations, a heat map visualization of the correlation matrix can be used, which proves to be an effective method.

4. Applying algorithms

Compare five machine learning methods, including SVM, decision trees, random forest classifiers, K nearest neighbors, and logistic regression, to determine which parameters are most likely to cause disease. Every algorithm has pseudocode that can be used to create a programming language. In Python, there is an easy way to build all kinds of algorithms with simple and fast code, which makes it real easy. 5. Machine Learning Algorithms: The algorithms used in this study are very useful in checking the correct results of heart disease tests as well as controlling disease-causing factors.

i. K-Nearest Neighbour algorithm: KNN, also known as K-Nearest Neighbors, is a supervised classification algorithm used to determine the labels of test data based on similar data points in the training set. It is commonly used as one of the classification techniques. However, KNN has certain assumptions that need to be met for optimal performance. These assumptions include having a clean dataset with labeled instances and relevant features. One drawback of using KNN is that it can be computationally expensive, especially when dealing with large datasets. The algorithm needs to compare the test data with all the training instances, which can take a significant amount of time. In terms of accuracy, the KNN algorithm achieves a rate of 81.82% on the given dataset. This means that it correctly classifies 81.82% of the instances in the test set.

$$Euclidean\ d \quad (x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

$$Xs = \frac{X - min}{max - min}$$

ii. Random Forest Classifier: This classifier employs a random forest algorithm to classify data. The random forest classifier is a valuable tool in machine learning libraries. By utilizing this classifier, we can achieve improved accuracy and reduced training time. Beforehand, we need to create a model by partitioning variables into training and test sets. Then, we train the dependent variables and make predictions based on the divided data. The maximum accuracy predicted result using the random forest classifier is 98.07%.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

iii. Decision Trees Classifier:

Pre-processing is done in this algorithm by separating data into training and test data. Because the values are normalised before prediction, feature scaling is possible. Import a selection tree classifier to in shape the education units of established and independent variables is used to detect the accuracy or reaction for the check set. The accuracy gained with this algorithm is 97.08%.

$$Entropy = \sum_{i=1}^{c} -p_i * log_2(p_i)$$

iv. Support Vector Machine (SVM): SVM is also one of the category algorithms in system gaining knowledge of in which higher accuracy can be expected. In contrast of different algorithms, it's miles better for predicting accuracy in an expected manner. In our detection, detected accuracy is 90.25% by using SVM.

$$Q(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=2}^{N} \sum_{j=1}^{N} a_i a_j d_i d_j x_i^T x_j$$

Where $0 \le a_i \le C\ \forall$ i

v. Logistics Regression Logistic regression is a statistical and machine learning method used to classify data records in a dataset by analyzing input field values. It involves predicting the outcome of a dependent variable using one or more sets of independent factors. Logistic regression can be employed for both binary and multi-class classification tasks This algorithm achieves 81.82% accuracy rate.

$$P = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X1 + \beta 2 X2 + ---- \beta n Xn}}$$

6. Confusion matrix

The confusion matrix is a useful tool for identifying classification errors. It provides information about the number of correct and incorrect detections for each class. The correct detections are recorded in the corresponding row and column of the expected

class, while the incorrect detections are recorded in the expected row for one class and the detected column for another class. Essentially, the confusion matrix summarizes the actual and expected classifications made by a classification process. The performance is assessed based on the data in the matrix.

6.1. Calculation of Accuracy

The formula is used to calculate the accuracy.

$$Accuracy = \frac{AP + AN}{M + N}$$

where, M = AP + AN and N = ALP + AN.
Or AP + AN (TOTAL).

6.2. precision (positive predictive value)

Precision (PREC) is a classification technique for locating items in a given class that have been incorrectly labelled. The highest precision score is 1.0, while the lowest is 0.0.

$$precision = \frac{TP}{TP+FP}$$

6.3. Recall

Sensitivity (SN), also known as recall or true positive rate (TPR), is a metric used to assess the accuracy of positive predictions. It is calculated by dividing the total number of correctly predicted positive outcomes by the total number of actual positive instances. Sensitivity values range from 0.0 (lowest) to 1.0 (highest), representing the accuracy of positive predictions. A higher sensitivity score indicates a better ability to accurately identify positive cases.–

$$Recall = TP+FN$$

6.4. F1-Score

The F1-score is a weighted average of recall and precision value calculated.

F- Measure= $2\times Precision \times Recall$
$Precision+Recall$

III.EXPERIMENTAL RESULTS

1. Dataset

There are numerous data about illnesses here. It has 14 attributes. The nominally valued factors that are considered are the Patient Identification Id (replaced by placeholder values), Gender, Cardiogram, Age, Chest Anxiety, Blood Pressure, Heart Rate, Cholesterol, Smoking, Alcohol Use, and Blood Sugar Level. Each of the 3000 examples in the training data set contains 14 unique attributes. The dataset's instances display the results of several tests conducted to accurate heart illness prognosis. The results are inspected, and the classifiers' performance is assessed. The comparison results were obtained using 10 tenfold cross-validations. The dataset is divided into two halves based on the attributes, with 30% of the data used for testing and 70% for training.

1.1. Description of dataset

The dataset utilized in this project includes the following attributes:

- Patient identification number (id)
- Age of the patient in years (age)
- Sex of the patient (1 = male; 0 = female) (sex)
- Chest pain location, where 1 represents substernal and 0 represents other locations (painloc)
- Chest pain provoked by exertion, indicated by 1 for yes and 0 for no (painexer)
- Chest pain relieved after rest, denoted by 1 for yes and 0 for no (relrest)
- Type of chest pain categorized as follows: o Value 1: Typical angina (cp) o
- Value 2: Atypical angina o
- Value 3: Non-anginal pain o
- Value 4: Asymptomatic
- Resting blood pressure (trestbps)
- Serum cholestoral (chol)
- Family history of coronary artery disease, where 1 signifies yes and 0 signifies no (famhist)
- Resting electrocardiographic results:

  - Value 0: Normal (restecg)
  - Value 1: ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: Probable or definite left ventricular hypertrophy by Estes' criteria

- Month of exercise ECG reading (ekgmo)
- Duration of exercise test in minutes (thaldur)
- Maximum heart rate achieved (thalach)
- Resting heart rate (thalrest)
- Diagnosis of heart disease (angiographic disease status):

o Value 0: < 50% diameter narrowing (num)
o Value 1: > 50% diameter narrowing

- Additional vessel-related information is present in attributes 59 through 68.

These attributes have been extracted from the UCI database and are instrumental in predicting heart disease accurately.

2. Performance of algorithms

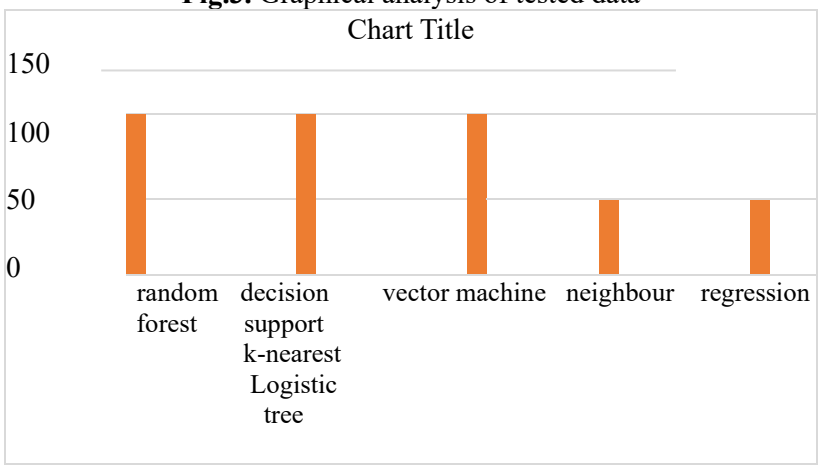
Table I shows the 14-attribute categorization secondary values and performance analysis.

Calculation, error, and accuracy are assessed. Random
Forest, Decision Tree, Support Vector Machine, KN earest Neighbour, and Logistic Regression were used. Random Forest fared best with 98.05% accuracy. Decision Tree has 97.08% accuracy. Support Vector Machine got 90.25% accuracy, K-Nearest Neighbour
81.82%, and Logistic Regression 81.82%. The Random Forest algorithm is accurate and fast.

Table - The performance of the classification algorithm

Sr no.	Model	Training accuracy %	Testing accuracy %
1	Random forest	100	98.05
2	Decision tree	100	97.07
3	Support vector machine	95.39	90.25
4	k-nearest neighbor	91.77	81.82
5	Logistic regression	89.53	81.82

Fig.3. Graphical analysis of tested data



V. CONCLUSION AND FUTURESCOPE

In India, like everywhere in the globe, heart disease is a leading killer. The early diagnosis of cardiac disease using cutting-edge technology like machine learning might have far-reaching positive societal effects. In high-risk people, problems may be mitigated or avoided altogether with prompt treatments like lifestyle changes made possible by early identification. With more and more individuals being diagnosed with heart disease every year, this is a significant medical advancement. Therefore, there is a critical need for accurate diagnosis and efficient treatment, and the application of suitable technology aid may considerably benefit both medical professionals and patients.

This study compares five major machine learning models for heart illness diagnosis: SVM, DT, RF, LR, and KNN. The dataset employed in this study includes 14 important variables linked to cardiac disease, allowing for a thorough assessment of the system as a whole. However, when taking into account all of the characteristics at once, the original system implementation showed inferior efficiency. Attribute selection methods were used to improve precision by focusing on a more manageable number of characteristics to analyse. The entire system's performance was enhanced by removing redundant features with strong correlation. The accuracy of these five machine learning techniques was compared, and a single prediction model was built as a consequence. The F1-score, accuracy, precision, recall, and confusion matrix were all used for illness prediction accuracy. The Random Forest classifier has the greatest accuracy of the algorithms we tested, at 98.05%. Machine learning may improve heart disease treatment choices. Heart disease prediction models exist. Many different approaches to treating heart disease may be taken once a diagnosis has been made. Machine learning may play a crucial role in selecting the best course of therapy by drawing on information from relevant databases.

In conclusion, using machine learning algorithms to early cardiovascular disease detection might transform the healthcare system. Better diagnosis, treatment options, and patient outcomes are all possible with the help of cutting-edge technology and the insights gleaned from massive volumes of data.

REFERENCES

1. Neha Kunte and Jay Nirmal "Detection of Heart Disease using Classification Algorithm" detection-of-heartdiseaseusing-classification-algorithm-IJERTCONV5IS011 68.pdf
2. Yasser Zeinali and Seyed Taghi Akhavan Niaki "Heart disease classification using signal processing and machine learning algorithms" Heart sound classification using signal processing and machine learning algorithms (1).pdf
3. Adyasha Rath and Debahuti Mishra "Heart disease detection using deep learning methods from imbalanced ECG samples" rath2021.pdf
4. Victor Chang and Vallabhanent Rupa Bhavani "An artificial intelligence model for heart disease detection using machine learning algorithms" 1-s2.0- S27724425220000 16 main.pdf
5. M.Akhil jabbar and B.L Deekshatulu "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm" j.protcy.2013.12.340 (1).pdf
6. Hlaudi Daniel Masethe and Mosima Anna Masethe "Prediction of Heart Disease using Classification Algorithms" WCECS2014_ pp809-812.pdf
7. Asha Rajkumar and Mrs. G.Sophia Reena "Diagnosis Of Heart Disease Using Datamining Algorithm" document.pdf
8. K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.
9. R. Chitra and V. Seenivasagam, "REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TEC HNIQUES," Journal on Soft Computing (ICTACT), vol. 3, no. 4, pp. 605– 609, 2013.
10. N. A. Sundar, P. P. Latha, and M. R. Chandra, "PERFORMANCE ANALYSIS OF ASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE," International Journal of Engineering Science & Advanced Technology, vol. 2, no. 3, pp. 470– 478, 2012.
11. S. A. Pattekari and A. Parveen, "PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES," International journal of Advanced Computer and Mathematical Sciences, vol. 3, no. 3, pp. 290–294, 2012. https://www.hindawi.com/journals/cin/2021/8387680
12. C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction.," IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society, vol. 10, no. 2, pp. 334–43, Apr. 2006.
13. Y. Xing, J. Wang, Z. Zhao, and A. Gao, "Combination
14. Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease," in 2007 International Conference on Convergence Information Technology (ICCIT 2007), 2007, pp. 868–872.
15. J.P. Li, A.U. Haq, S.U. Din, J. Khan, A. Khan, A. Saboor, Heart disease identification method using machine learning classification in E-healthcare, IEEE Access 8 (2020) 107562–107582.
16. A. Sengur, An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases, Expert Syst. Appl. 35 (1– 2) (2008) 214–222.
17. ˙ I. Babaoglu, O. Findik, E. Ülker, A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine, Expert Syst. Appl. 37 (4) (2010) 3177–3183. [13] C. Kwak, O.W. Kwon,
18. Cardiac disorder classification by heart sound signals using murmur likelihood and hidden Markov model state likelihood, IET Signal Process. 6 (4) (2012) 326– 334. [16] J. Nahar, T. Imam, K.S. Tickle, Y.P.P. Chen, Computational intelligence for heart disease diagnosis: a medical knowledge driven approach, Expert Syst. Appl. 40 (1) (2013) 96–104.
19. S. Shilaskar, A. Ghatol, Feature selection for medical diagnosis: evaluation for cardiovascular diseases, Expert Syst. Appl. 40 (10) (2013) 4146– 4153.
20. R. Tao, S. Zhang, X. Huang, M. Tao, J. Ma, S. Ma, C. Shen, Magnetocardiography based ischemic heart disease detection and localization using machine learning methods, IEEE Trans. Biomed. Eng. 66 (6) (2018) 1658–1667.
21. Logesh Kannan N and Gowsalya M" Heart Disease Detection Using Machine Learning" www.researchgate.net/publication/346432379_Heart_ Disease Detection_ Using_ Machine_ Learning Prediction of Heart Disease Using a Combination of M achine Learning and Deep Learninghttps://www.hindawi.com/journals/cin/2021/8387680/