iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** II **Month of publication:** February 2023

DOI: https://doi.org/10.22214/ijraset.2023.49304

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

Machine Learning Based Heart Disease Prediction System

Apoorva G O[1], Vishwas C N[2]
[1, 2]*Department of Computer Science and Engineering, SJMIT, Chitradurga*

Abstract: ***Heart attack disease is one of the leading causes of the death worldwide. In today's common modern life, deaths due to the heart disease had become one of major issues, that roughly one person lost his or her life per minute due to heart illness. Predicting the occurrence of disease at early stages is a major challenge nowadays. Machine learning when implemented in health care is capable of early and accurate detection of disease. In this work, the arising situations of Heart Disease illness are calculated. Datasets used have attributes of medical parameters. The datasets are been processed in python using ML Algorithm i.e., Random Forest Algorithm. This technique uses the past old patient records for getting prediction of new one at early stages preventing the loss of lives. In this work, reliable heart disease prediction system is implemented using strong Machine Learning algorithm which is the Random Forest algorithm, which read patient record data set in the form of CSV file. After accessing dataset the operation is performed and effective heart attack level is produced. Advantages of proposed system are High performance and accuracy rate and it is very flexible and high rates of success are achieved.***
Keywords: Random Forest (RF) and CSV

I. INTRODUCTION

Heart disease effects the functioning of the heart. World Health Organization had made a survey and made a conclusion that 10 million people are affected with heart disease and lost their lives. The problem that the Healthcare industry faces in today's life is early prediction of disease after a person is affected.

Records or data of medical history is very large and the data in real world might be incomplete and inconsistent. In past predicting the disease effectively and treatment to patients might not be possible for every patient at early stages under these circumstances.

Many scientists tried to build a model which is capable of predicting the heart disease in the early stage, but they are not able to build a perfect model. Every proposed system has disadvantages in its own way. In the existing system, Shen et al. had initially, proposed a system which is based on self-applied questionnaire.

In this system the user need to enter all the symptoms which he is suffering from, based on that the result is predicted. This study is based on the analysis data collected in SAQ. Chen et al. came up with an idea to predict heart disease. He used the technique of Vector Quantization which is one of the artificial intelligence techniques for classification and prediction purpose. Training of neural networks is performed using back propagation to evaluate the prediction system. In the testing phase approximately 80% accuracy is achieved on testing set. Practical use of data collected from previous records is time consuming. Low accuracy rate. So to overcome this we are implementing Random forest algorithm in order to achieve accurate results in less time. Machine learning is given a major priority in modern life in many applications and in healthcare sector. Prediction is one of area where machine learning plays a vital role, our topic is to predict heart disease by processing patient's dataset and a data of patients i.e., user of whom we need to predict the chances of occurrence of a heart disease. Heart disease can be detected using the symptoms like: high blood pressure, chest pain, hypertension, cardiac arrest, etc. There are many types of heart diseases with different types of symptoms. Like: 1) heart disease in blood vessels: chest pain, shortness of breath, pain in neck throat., 2)heart disease caused by abnormal heartbeats :slow heartbeat, discomfort, chest pain., etc. Most common symptoms are chest pain, shortness of breath, discomfort, chest pain., etc. Most common symptoms are chest pain, shortness of breath, fainting. Causes of heart disease are defects you're born with, high blood pressure, diabetes, smoking, drugs, alcohol. Sometimes in heart disease the infection also affects the inner membrane which is identified by symptoms like fever, fatigue, dry cough, skin rashes. Causes of heart infection are bacteria, viruses, parasites. Types of heart disease: Cardiac arrest, Hypertension, Coronary artery disease, Heart failure, Heart infection, congenital heart disease, slow heartbeat, Stroke type heart disease, angina pectoris. Nowadays there are too many automated techniques to detect heart disease like data mining, machine learning, deep learning, etc. So, in this paper we will briefly introduce machine learning techniques. In this we train the datasets using the machine learning repositories.

There are some risk factors on the basis of that the heart disease is predicted. Risk factors are: Age, Sex, Blood pressure, Cholesterol level, Family history of coronary illness, Diabetes, Smoking, Alcohol, Being overweight, Heart rate, Chest Pain.

A. Problem Statement

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

B. Objectives

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set. Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

C. Methodology

Figure 1 depicts the overall process of this work. Our aim is to build an application of heart disease prediction system using robust Machine Learning algorithm which is Random Forest algorithm. A CSV file is given as input. After the successful completion of operation the result is predicted and displayed.

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1) Collection of Dataset
2) Selection of attributes
3) Data Pre-Processing
4) Balancing of Data
5) Disease Prediction

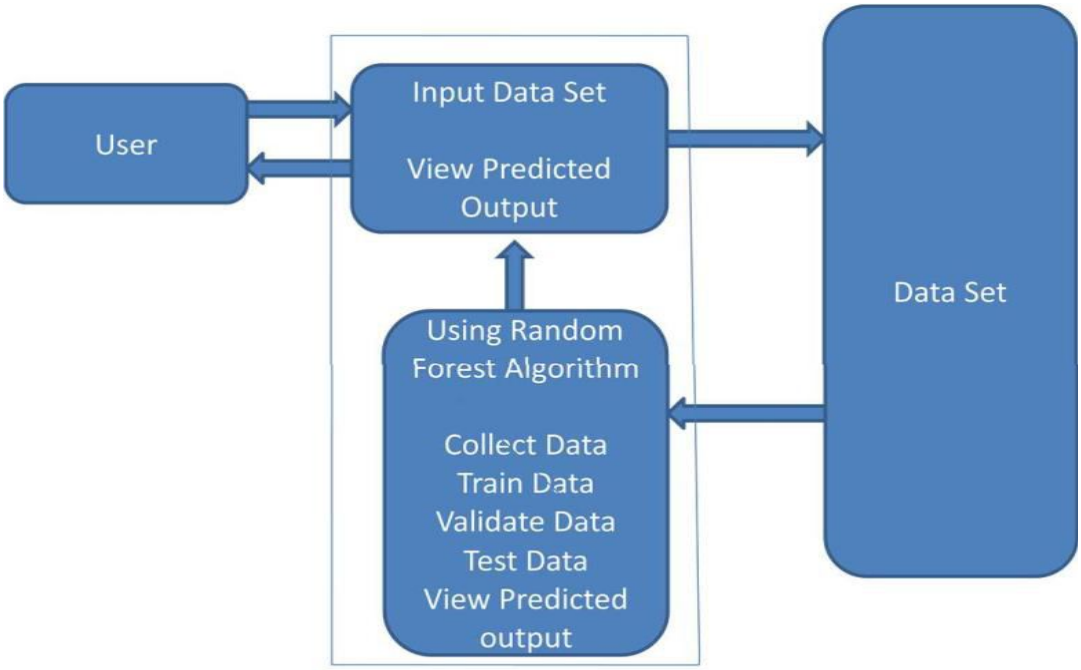


Fig. 1 Block Diagram of Proposed System

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

II. LITERATURE SURVEY

This section discusses the state-of-the-art methods for heart disease diagnosis using machine learning techniques that were accomplished by various effective research works.

R. Perumal et al. [18] developed a heart disease prediction model using the Cleveland dataset of 303 data instances through feature standardization and feature reduction using PCA, where they identified and utilized seven principal components to train the ML classifiers. They concluded that LR and SVM provided almost similar accuracy values (87% and 85%, respectively) compared to that of k-NN with 69%. C. B. C. Latha et al. [19] performed a comparative analysis to improve the predictive accuracy of heart disease risk using ensemble techniques on the Cleveland dataset of 303 observations. They appliedthe brute force method to obtain all possible attribute set combinations and trained the classifiers. They achieved a maximum increase in the accuracy of a weak classifier of 7.26% based on ensemble algorithm, and produced an accuracy of 85.48% using majority vote with NB, BN, RF, and MLP classifiers using an attribute set of nine attributes. D. Ananey-Obiri et al. [20] developed three classification models, namely, LR, DT, and Gaussian naïve Bayes (GNB), for heart disease prediction based on the Cleveland dataset. Feature reduction was performed using single value decomposition, which reduced the features from 13 to 4. They concluded that both LR and GNB had predictive scores of 82.75% and AUC of 0.87. It was suggested that other models, such as SVM, k-NN, and random forest, be included.

N. K. Kumar et al. [21] trained five machine learning classifiers, namely, LR, SVM, DT, RF, and KNN, using a UCI dataset with 303 records and 10 attributes to predict cardiovascular disease. The RF classifier achieved the highest accuracy of 85.71% with an ROC AUC of 0.8675 compared to the other classifiers. A. Gupta et al. [22] replaced the missing values based on the majority label and derived 28 features using the Pearson correlation coefficient from the Cleveland dataset and trained LR, KNN, SVM, DT, and RF classifiers using the factor analysis of mixed data (FAMD) method; the results based on a weight matrix RF achieved the best accuracy of 93.44%. M. Sultana et al. [23] explored KStar, J48, sequential minimal optimization (SMO), BN, and MLP classifiers using Weka on a standard heart disease dataset from the UCA repository with 270 records and 13 attributes; they achieved the highest accuracy of 84.07% with SMO.

S. Mohan et al. [24] developed an effective hybrid random forest with a linear model (HRFLM) to enhance the accuracy of heart disease prediction using the Cleveland dataset with 297 records and 13 features. They concluded that the RF and LM methods provided the best error rates.

S. Kodati et al. [25] developed a heart disease prediction system (HDPS) with the Cleveland dataset of 297 instances and 13 attributes using Orange and Weka data mining tools, where they evaluated the precision and recall metrics for the naïve Bayes, SMO, RF, and KNN classifiers.

A. Ed-daoudy et al. [26] researched the Cleveland dataset of 303 records and 14 attributes from UCI. They evaluated the performance of the four main classifiers, namely, SVM, DT, RF, and LR, using Apache Spark with its machine learning library MLlib.

I. Tougui et al. [27] compared the performances of LR, SVM, KNN, ANN, NB, and RF models to classify heart disease with the Cleveland dataset with 297 observations and 13 features using six data mining tools: Orange, Weka, RapidMiner, Knime, MATLAB, and Scikit-Learn.

V. Pavithra et al. [28] proposed a new hybrid feature selection technique with the combination of random forest, AdaBoost, and linear correlation (HRFLC) using the UCI dataset of 280 instances to predict heart disease. Eleven (11) features were selected using filter, wrapper, and embedded methods; an improvement of 2% was found for the accuracy of the hybrid model.

C. Gazeloglu et al. [29] projected 18 machine learning models and 3 feature selection techniques (correlation-based FS, chi-square, and fuzzy rough set) to find the best prediction combination for heart disease diagnosis using the Cleveland dataset of 303 instances and 13 variables.

N. Louridi et al. [30] proposed a solution to identify the presence/absence of heart disease by replacing missing values with the mean values during pre-processing. They trained three machine learning algorithms, namely, NB, SVM (linear and radial basis function), and KNN, by splitting the Cleveland dataset of 303 instances and 13 attributes into 50:50, 70:30, 75:25, and 80:20 training and testing ratios.

M. Kavitha et al. [31] implemented a novel hybrid model on the Cleveland heart dataset of 303 instances and 14 features with a 70:30 ratio for training and testing by applying DT, RF, and hybrid (DT + RF) algorithms.

B. A. Tama et al. [32] designed a stacked architecture to predict heart disease using RF, gradient boosting machine, and extreme gradient boosting with particle swarm optimization (PSO) feature selection using various heart disease datasets, including the Cleveland with 303 instances and 13 attributes.

From the experimental works, it is understood that data pre-processing and feature selection can substantially enhance the classification accuracy of machine learning algorithms. During pre-processing, most researchers [18,19,21,22,26,29–32] replaced the missing values, either by using the mean value or the majority mark of that attribute, to make sure the dataset was comprehensive. In some works [20,24,25,27], the missing valued instances were removed. Feature selection is a challenging task due to the large exploration space. It grows exponentially according to the number of features available in the dataset. To solve this issue, an effective comprehensive search technique is required during feature selection. Furthermore, some studies have employed ensemble models, which combine multiple basic learning algorithms to obtain a better prediction accuracy. However, the performance of these techniques can further be improved regarding accurately predicting disease.

III.SYSTEM DESIGN

System design thought as the application of theory of the systems for the development of the project. System design defines the architecture, data flow, use case, class, sequence and activity diagrams of the project development.

A. System Architecture

The below architecture diagram in figure 2 illustrates how the system is built and is the basic construction of the software method. Creations of such structures and documentation of these structures is the main responsible of software architecture.

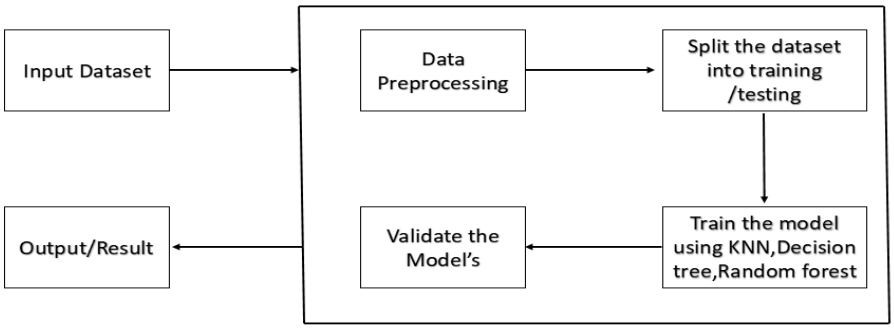


Fig. 2 Architecture Diagram of Proposed System

The working principle of the system is shown in fig.2

1) *Input Dataset:* The user enters the input.
2) *Data Pre-Processing:* Itcan refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process. Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre-processing of data is required for improving the accuracy of the model.
3) *Split Dataset Into Train and Test:* The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications. This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results.
4) *Train the Model:* Input which is compared with the data present in the existing data set by using the Random Forest Algorithm. It is an efficient ML algorithm that comes under supervised learning technique. It is be used for both Regression and Classification problems. To solve a complex problem, it uses a process of combining multiple classifiers, to increase the accuracy and performance of the model. "Random Forest is known as classifier that contains more number of decision trees on different subsets of the given dataset and considers the average to improve the predictive accuracy of that dataset."
5) *Validate:* Model validation refers to the process of confirming that the model actually achieves its intended purpose. In most situations, this will involve confirmation that the model is predictive under the conditions of its intended use.
6) *Output or Result:* Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Different classification techniques as stated will be applied on pre-processed data to predict the accuracy of heart disease. Accuracy measure compares the accuracy of different classifiers.

B. Flowchart

The first step is to collect data on patients, which typically includes demographic information as well as medical history. The data is then cleaned and pre-processed to prepare it for analysis.

The dataset is then split into training and testing sets. The training set is used to train the machine learning model, and the testing set is used to evaluate its performance. Various supervised learning algorithms can be used to train the model, such as logistic regression, decision trees, or neural networks. The accuracy of the model is then evaluated using the testing set.

The model can be further optimized by adjusting the hyper parameters using techniques such as cross-validation and grid search. Once the model is optimized, it can be deployed in a real-world setting to predict the likelihood of heart disease in new patients.

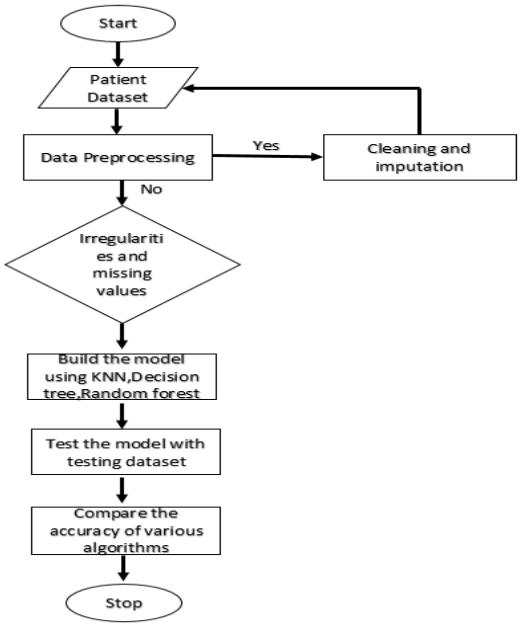


Fig. 3 Flowchart for Heart Prediction System

C. Sequence Diagram

A sequence diagram is a type of UML (Unified Modeling Language) diagram that depicts the interactions between objects or components in a system or process. It shows the order in which messages are exchanged between the objects, along with the time sequence of those messages. Below figure 4 shows the sequence diagram of crop and fertilizer recommendation.

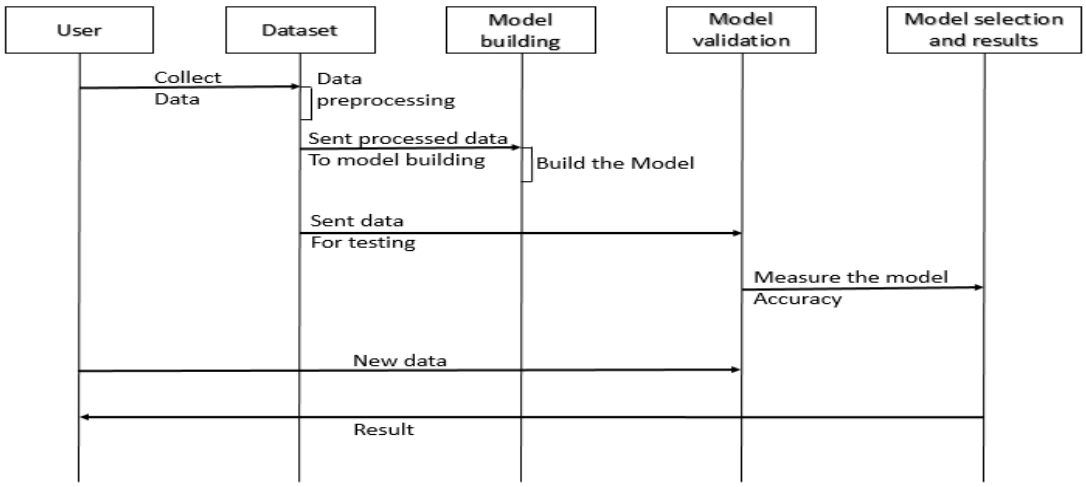


Fig. 4 Sequence Diagram of Crop Recommendation

IV. RESULTS

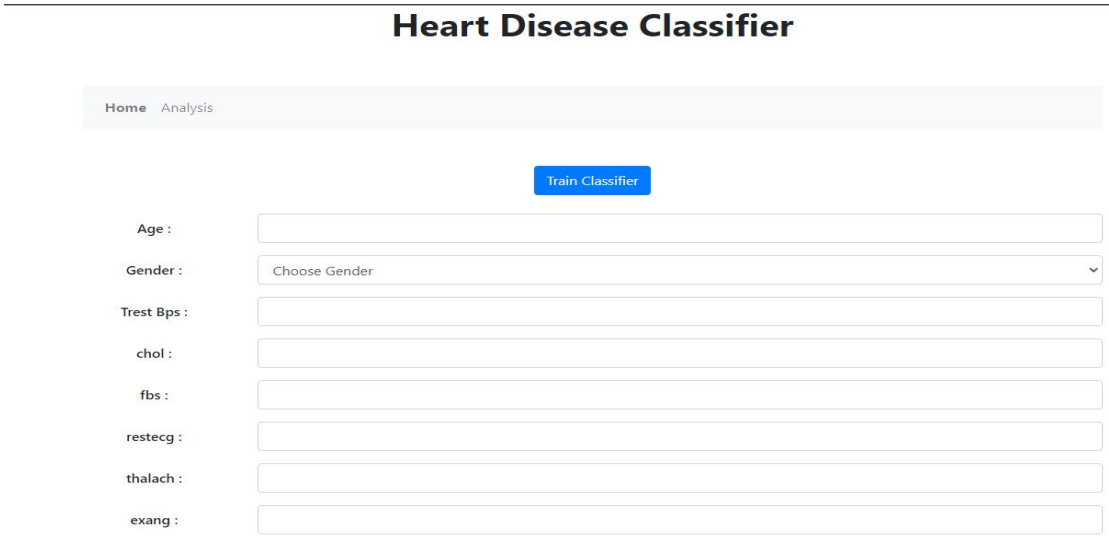


Fig. 5 Home Page

Fig 6 shows the home page of our proposed system, which includes various parameters text field to input the user data. It has also train classifier button to train the ML model.

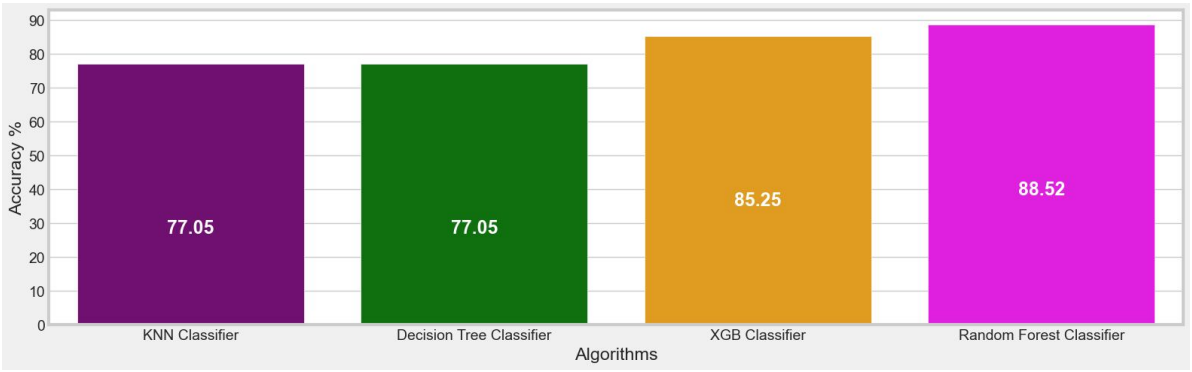


Fig. 6 Accuracy of different ML algorithms is compared

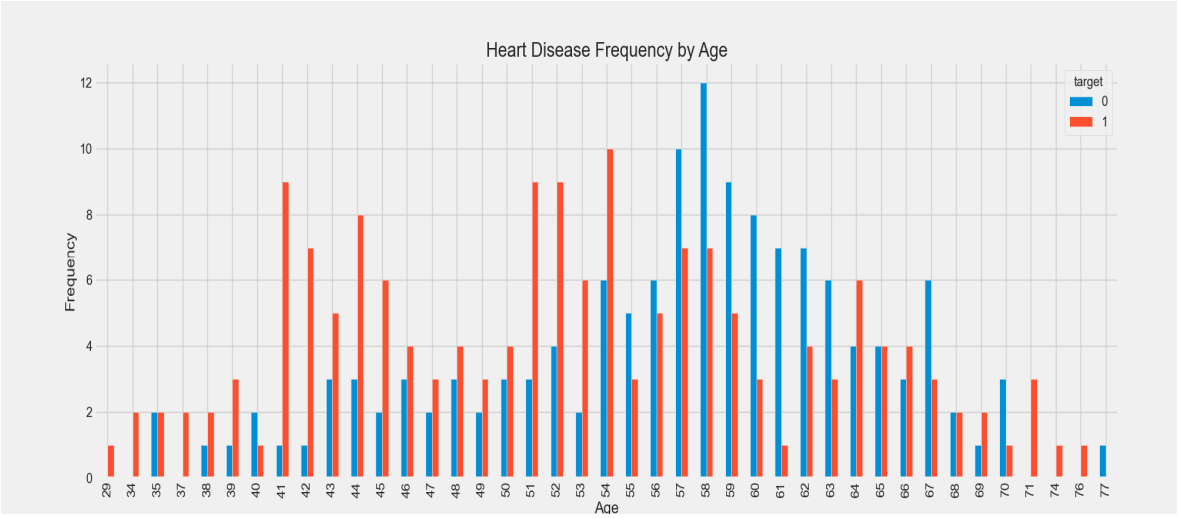


Fig. 7 Heart disease datasets represented based on age-frequency

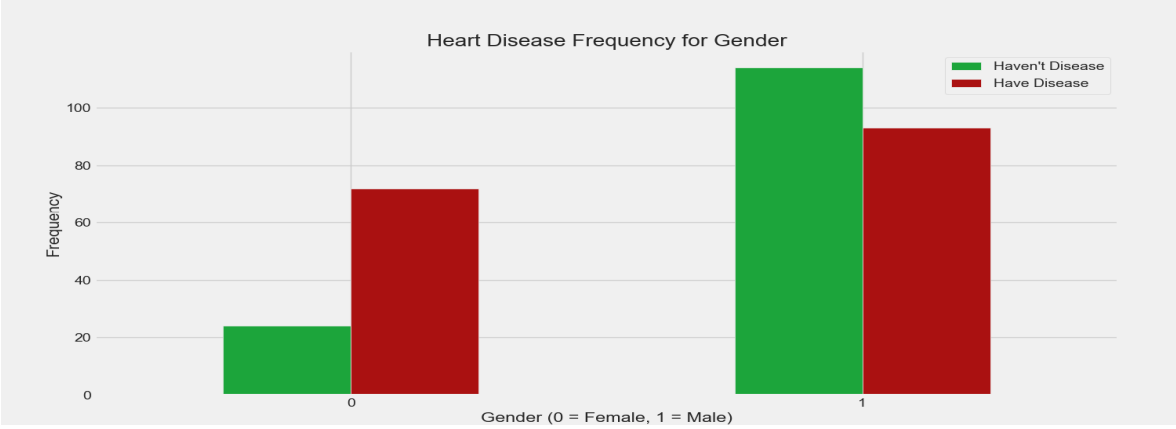


Fig. 8 Heart disease datasets represented based on Gender-frequency

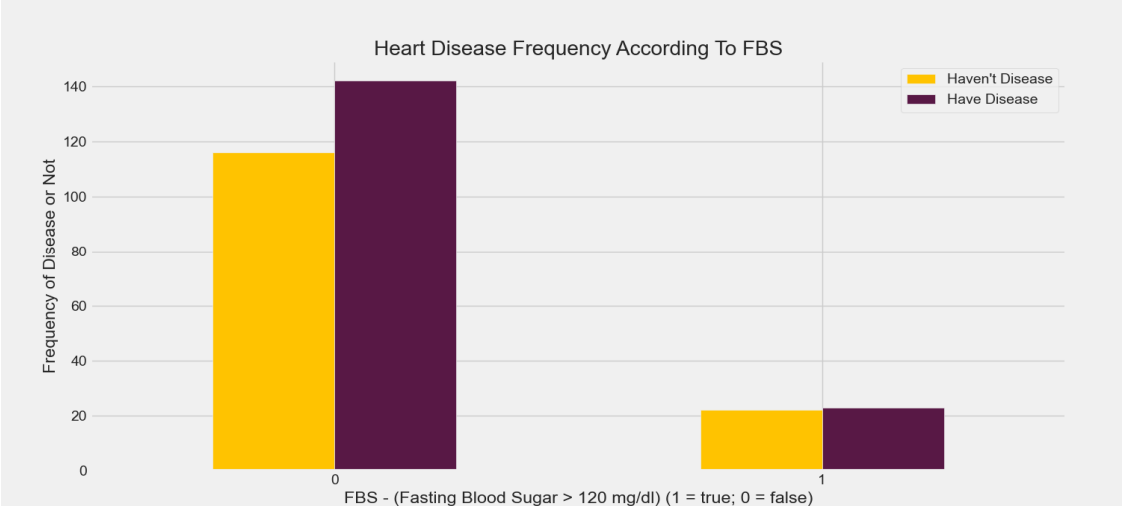


Fig. 9 Heart disease datasets represented based on FBS-frequency

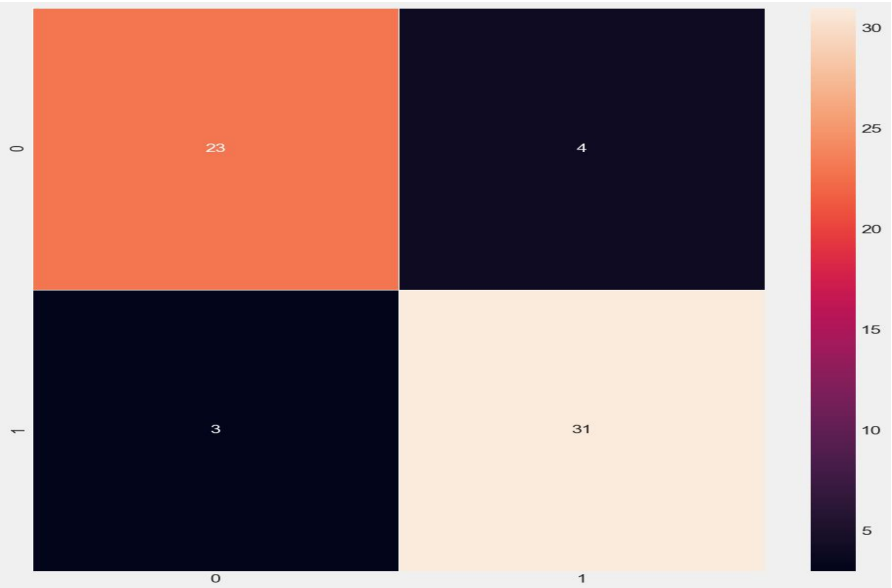


Fig. 10 Confusion Matrix of Random Forest Algorithm which we applied for prediction

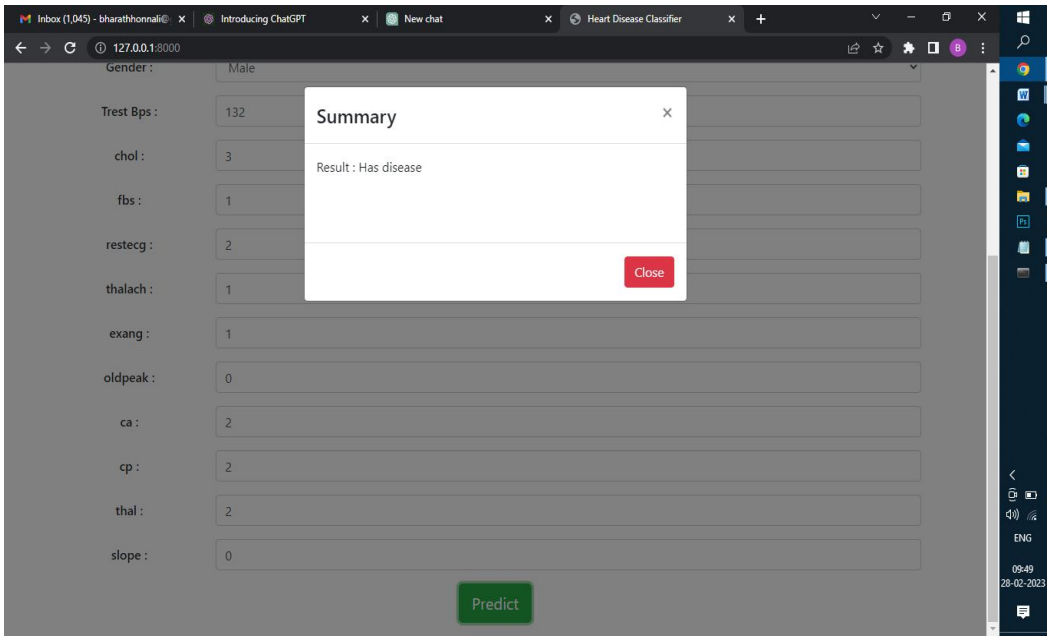


Fig. 11 Predicting result based on the User Input

V.CONCLUSION

Heart disease prediction is a major challenge in the present modern life. With this application if the patient/user is away from reach of doctor, he/she can make use of the application in prediction of disease just by entering the report values. And can proceed further whether to consult a doctor or not.

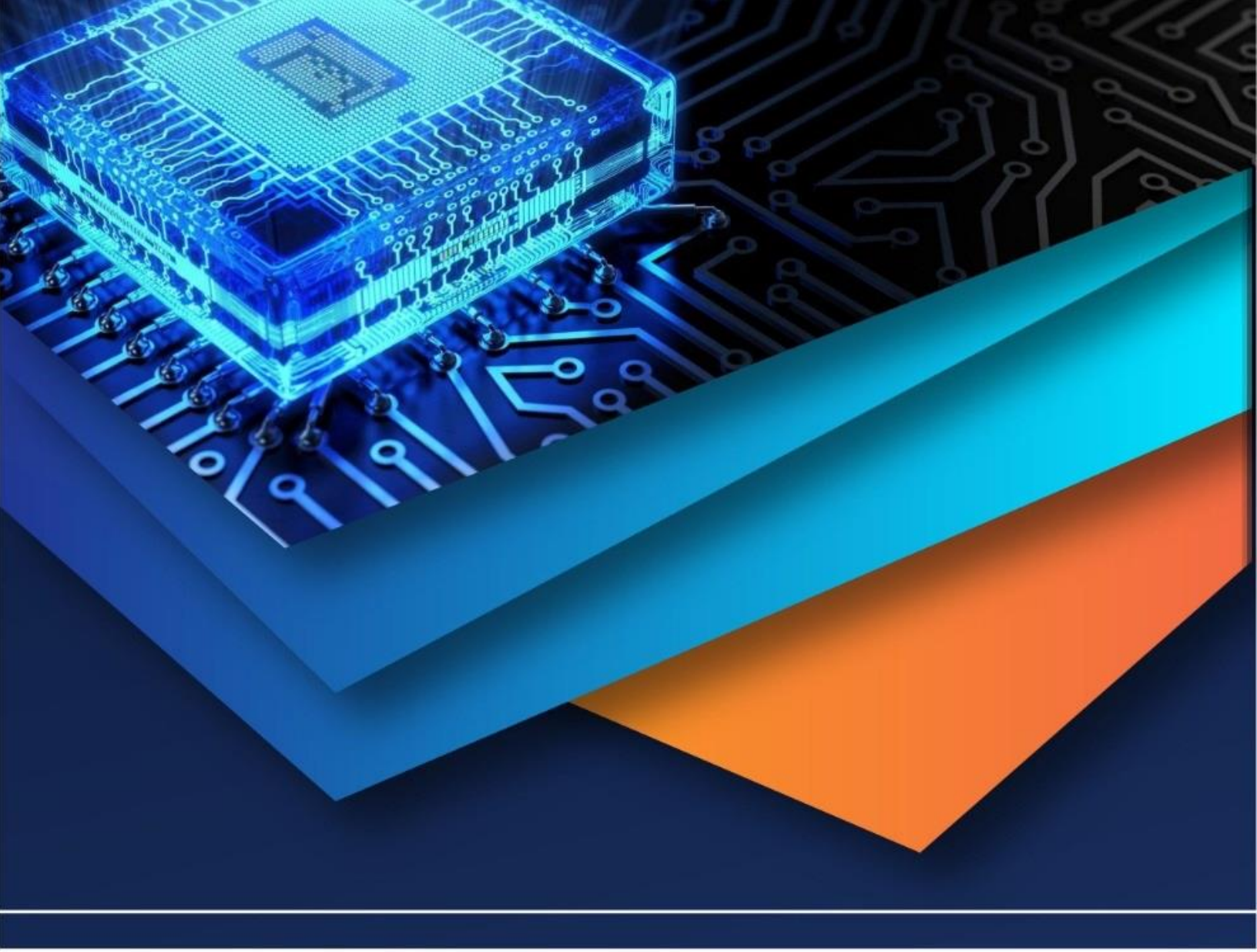
A. Future Scope

In future this application can extended by updating some features like, if the user is effected with heart disease all his family members will be notified with a message in early. And also the information should be passed to the nearest hospital. Another feature is there should be online doctor consultation with the nearest doctor available. In this regard, it is important to note that, ML applications using various efficient algorithms are utilized not only in disease prediction and diagnosis but also in the field of radiology, bioinformatics and medical imaging diagnosis etc.

REFERENCES

[1] Kaan Uyar and Ahmet İlhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks" in B.V ICTASC, Elsevier,pp
[2] Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir and Y. K. Sharma, "Heart Disease Prediction Using Data Mining Techniques", © IJRAT Special Issue National Conference “NCPC-2016”, pp. 104-106, 19 March 2016.
[3] Berry JD, Lloyd-Jones DM, Garside DB, et al. Framingham risk score and prediction of coronary heart disease death in
[4] young men. Am Heart J. 2007;154(1):80–6.
[5] Theresa Princy and R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", © IEEE ICCPCT, 2016. Kaur h Beant and Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", © IJRITCC, vol. 2, no. 10, pp. 3003-08, 2014.
[6] Kirmani, M.M., Ansarullah, S.I.: Prediction of heart disease using decision tree a data mining technique. IJCSN Int. J. Comput. Sci. Netw. 5(6), 885–892 (2016)
[7] Salam Ismaeel, Ali Miri et al., "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis", IEEE Canada International Humanitarian Technology
[8] Tahira Mahboob, Rida Irfan and Bazelah Ghaffar et al.”Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics” ©2017 IEEE
[9] Ammar Asjad Raja, Irfan-ul-Haq , Madiha Guftar Tamim Ahmed Khan “Intelligence syncope Disease Prediction Framework using DM-techniques” FTC 2016 –Future Technologies Conference 2016.
[10] M.A. Jabbar, B.L.Deekshatulu, and Priti Chandra, “ Intelligent heart disease prediction system using random forest and evolutionary approach”, Journal of Network and Innovative Computing, Vol. 4, pp.174-184, 2016.
[11] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” Machine Learning,(1997).
[12] Ayon Dey, Jyoti Singh, N. Singh “Analysis of supervised machine learning algorithms for heart disease prediction”
[13] Pasha, S.J.; Mohamed, E.S. Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms forEffective Disease Risk Prediction. IEEE Access 2020, 8, 184087–184108. [CrossRef]

[14] Swain, D.; Pani, S.K.; Swain, D. A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning. InProceedings of the 2018 International Conference on Advanced Computation and Telecommunication, ICACAT, Bhopal, India, 28–29 December 2018; pp. 1–6. [CrossRef]
[15] Weng, S.F.; Reps, J.M.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 2017, 12, e0174944. [CrossRef]
[16] Khan, Y.; Qamar, U.; Yousaf, N.; Khan, A. Machine Learning Techniques for Heart Disease Datasets: A Survey. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing, Zhuhai, China, 22–24 February 2019; pp. 27–35. [CrossRef]
[17] Goel, S.; Deep, A.; Srivastava, S.; Tripathi, A. Comparative Analysis of various Techniques for Heart Disease Prediction. In Proceedings of the 2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019, Mathura, India, 21–22 November 2019; pp. 88–94. [CrossRef]
[18] Perumal, R. Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. Int. J. Adv. Sci. Technol. 2020, 29, 4225–4234.
[19] Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked 2019, 16, 100203. [CrossRef]
[20] Ananey-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms. Int. J. Comput. Appl. 2020, 176, 17–21. [CrossRef]
[21] Kumar, N.K.; Sindhu, G.; Prashanthi, D.; Sulthana, A. Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 15–21. [CrossRef]
[22] Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. IEEE Access 2019, 8, 14659–14674. [CrossRef]
[23] Sultana, M.; Haider, A.; Uddin, M.S. Analysis of data mining techniques for heart disease prediction. In Proceedings of the 2016 3rd International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2016, Dhaka, Bangladesh, 22–24 September 2016; pp. 1–5. [CrossRef]
[24] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019, 7, 81542–81554. [CrossRef]
[25] Kodati, S.; Vivekanandam, R. Analysis of Heart Disease using in Data Mining Tools Orange andWeka Sri Satya Sai University Analysis of Heart Disease using in Data Mining Tools Orange and Weka. Glob. J. Comput. Sci. Technol. 2018, 18.
[26] Ed-Daoudy, A.; Maalmi, K. Performance evaluation of machine learning based big data processing framework for prediction of heart disease. In Proceedings of the International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco, 26–27 December 2019; pp. 1–5. [CrossRef]
[27] Tougui, I.; Jilbab, A.; El Mhamdi, J. Heart disease classification using data mining tools and machine learning techniques. Health Technol. 2020, 10, 1137–1144. [CrossRef]
[28] Pavithra, V.; Jayalakshmi, V. Hybrid feature selection technique for prediction of cardiovascular diseases. Mater. Today Proc. 2021, 22, 660–670. [CrossRef]
[29] Gazelo˘ glu, C. Prediction of heart disease by classifying with feature selection and machine learning methods. Prog. Nutr. 2020, 22, 660–670. [CrossRef]
[30] Louridi, N.; Amar, M.; El Ouahidi, B. Identification of Cardiovascular Diseases Using Machine Learning. In Proceedings of the 7th Mediterranean Congress of Telecommunications 2019, CMT 2019, Fez, Morocco, 24–25 October 2019; pp. 1–6. [CrossRef]
[31] Kavitha, M.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S. Heart Disease Prediction using Hybrid machine Learning Model. In Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, Coimbatore, India, 20–22 January 2021; pp. 1329–1333. [CrossRef]
[32] Tama, B.A.; Im, S.; Lee, S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. BioMed Res. Int. 2020, 2020. [CrossRef]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY