International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 5 ǁ May 2023 ǁ PP. 450-457

Heart Disease Detection Using Machine Learning

Riya Dubey, Dhruv Agnihotri

Abstract:- *If the correct prediction is made of heart disease, people can avoid life-threatening situations, but if the wrong prediction is made, it can prove to be fatal. Every day the cases of coronary heart illnesses are growing at a rapid rate and it's very important and regarding predict one of these diseases. The main objective of the project is to get better accuracy to detect the heart-disease using algorithms in which the target output counts whether a person has heart disease or not. In this report, different machine-learning algorithms are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. Using different types of parameters in the dataset we can predict cardiac disease. This is beneficial in the field of medical science as it could reduce the load of medical practitioners and treat the patient more efficiently. Treatment can be done according to the severity of the detected disease. The term Heart disease covers all diseases and disorders related to the heart and blood vessels. The diagnosis and treatment for this sum up to a huge amount. This model would be able to develop a more accurate system for the detection of heart disease using classification algorithms.*
Index Terms *-Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, Machine Learning*

Date of Submission: 04-05-2023 Date of acceptance: 15-05-2023

I. INTRODUCTION

The number of deaths because of vessel diseases increased by forty first between 1990 and 2013, the ascent from 12.3 million deaths to seventeen.3 million deaths globally. In addition thereto, 1/2 the deaths within the u. s. and other developed countries square measure because of constant issues. Therefore, early detection of heart diseases is needed to reduce health complications.

To control the death rates, the preventive measures square measure steered by the World Health Organization (WHO) when the prediction of rates of hefailureured patients. The rates square measure increasing yearly and preventive measures got to be taken.

To control the death rates, the preventive measures square measure steered by the World Health Organization (WHO) when the prediction of rates of hefailureured patients. The rates square measure increasing yearly and thhe preventive measures got to be taken.

Machine learning has been widely utilized in the fashionable attention sector for designation and predicting the presence of disease exploitation knowledge models. Logistic regression is one such comparatively used machine learning algorithm for studies involving risk assessment of complex diseases.

II. RELATED WORK

Heart failure unwellness has emerged and has been rising in the concert of the foremost deathful diseases for decades around the world. The death rates in our unit have increase over the years. China had the biggest death rates last year (2020) followed by Russia, India,, America and land. Some others Japan, France, and South American country had low death rates. Among men, the death rate was underlying nine.6 million and in feminine, the death rate was underlying eight.9 million. Among the individuals old-time between thirty and seventy years, the rates were concerning half-dozen totally different distinctive datasets are accustomed to predict the center failure. The supervised and unsupervised innovations have raised the accuracy of the nosology of models. because of the increase in heart death rates, 17.9 million individuals area unit dying. Hence, to predict the causes of the center failure, the machine learning models area unit engineered on connected parameters. To predict the death rate of heart failufailureents, the data mining quest area unit was used. The crucial attack on unwellness could be a threatening and commonest unwellness. there's a desire to predict the prevalence of this unwellness-supported combination of risk factors. Hence, totally different techniques have to be enforced and compared to supported commonplace metrics.

In Medical Field, the center malady prediction at the Associate in Nursing early stage is extremely necessary to avoid wasting the patient's life. Hard copy there ought to be a tool out there to predict the center malady. This tool is provided are doctors to discover the presence of a malady. Hence, performance analysis is completed to supply this tool.

There are numerous totally different machine learning algorithms, among those, one algorithmic program, KNN, is employed to predict the center malady at the associate early stage.

Heart failure prediction accuracy has been improved by exploiting UCI dataset. several techniques of machine

learning are wont to predict the probabilities of heart disease. To predict the center unwellness, the accuracy of various machine learning algorithms is calculated. The United Nations agency (World Health Organization) collects the from varied health centers to predict the count of heart failure patients. It uses several needed techniques to count the death rates and brings on some preventive measures to be taken to forestall the reason for heart disease. The EHMS (Electronic – Hospital Management System) that is additionally known as HIS (Hospital info System) was designed to manage the complete hospital administration. HMS/HIS are strictly liable for managing the main points relating to the patients. The medical records of the patients are kept in information and the risk of databases is secured exploitation the computer database.

III. METHODS

This paper uses the "heart_failure_clinical_records_dataset.csv" dataset that has reviews on heart failure patients. Dataset consists of 377,650 views. The records have knowledge|the info|the information} format for data within the dataset as shown below.

Column_ID	**Description**
Age	Age
Anemia	Decrease of red blood cells or hemoglobin (Boolean)
creatinine_phosphokinase	Level of the CPK enzyme In the blood (mcg/L)
Diabetes	If the patient has diabetes (Boolean)
ejection_fraction	Percentage of blood leavin--g the heart at each contraction
high_blood_pressure	If the patient has hypertension (Boolean)
Platelets	Platelets in the blood (kiloplatelets/mL)
serum_creatinine	Level of serum creatinine in the blood (mg/dL)
serum_sodium	Level of serum sodium in the blood (mEq/L)
Sex	Woman or man (binary)
Smoking	If the patient smokes or not (Boolean)
Time	Follow-up period (days)

DEATH_EVENT If the patient deceased during the follow-up period (Boolean) fig.1.
The projected system has the method diagram as shown below in Fig. 1. the subsequent area unit the steps distributed.

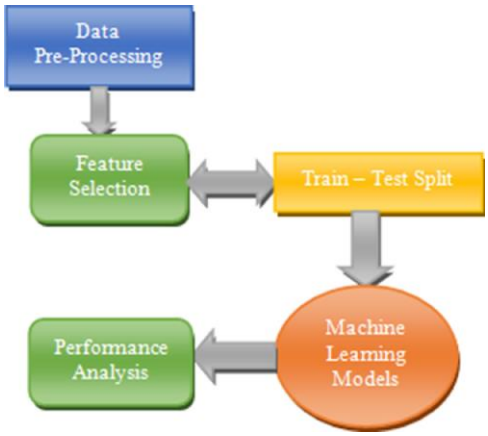


Fig. 1. Proposed system process diagram

IV. Data Pre-Processing

It is a method of modifying the specified columns during a knowledge set (if the information set contains raw data) for analyzing the data. as an example, the dataset utilized in this paper had few string formats that represent labels, genders, etc. knowledge Pre-processing was applied to retrieve numerical knowledge, string knowledge, categorical knowledge fittingly. during this paper, the information preprocessing is done to get rid of the inessential values within the columns like , missing values, false values, and null values. Our dataset could also be noisy typically, therefore to get rid of these inessential and missing columns, this step is doled out.

V. Feature Selection

Feature choice may be a method of choosing the mandatory variables to extend the accuracy. {the choice|the choice} of variables throughout the feature selection is manual or automatic. during this paper, the mandatory options square measure hand-picked to boost the share of accuracy. This step involves choosing the relevant options and discarding the orthogonal options. during this paper, the options relating to the center failure square measure hand-picked, and access options square measure removed.

VI. Train- Test Split

Train–check Split could be a technique to divide the given dataset into subsets and train additional. this method is often accustomed to rate the performance of machine learning algorithms. during this paper, the dataset is split to suit the models. the information is split into split check subsets by taking the dimensions of check.

VII. Machine Learning

Machine learning is widely used in nearly many fields at intervals in the world likewise the health care sector. Machine learning is AN application of computing (AI) that offers systems the pliability to automatically learn and improve from experience while not being expressly programmed. Further, machine learning at its simplest is the application of using algorithms to dissect data, learn from it, then build a determination or prediction concerning one factor at intervals across the globe. Their unit of measurement is a pair of major categories of problems usually resolved by machine learning i.e. regression and classification. Mainly, the regression algorithms unit of measurement used for numeric data and classification problems embrace binary and multicategory problems.

Machine learning algorithms square measure any divided into 2 classes supervised learning and unattended learning. Basically, supervised learning is performed by mistreatment previous data in output values whereas unattended learning doesn't have predefined labels thus the goal of this can be to infer the natural structures at intervals in the dataset. Therefore, the choice of machine learning rule has to be compelled to be fastidiously evaluated.

7.1 LOGESTIC REGRESSION

Logistic regression is one of the foremost style Machine Learning algorithms, that comes below the supervised Learning technique. It's used for predicting the specific variable employing a given set of freelance variables. Logistic regression predicts the output of a categorical variable. so the result should be a categorical or distinct worth. It will be either affirmative or No, 0 or 1, true or False, etc. however rather than giving the precise worth as zero and one, it provides the probabilistic values that lie between zero and one.

Logistic Regression is way the same as the regression toward the mean except that however they're used. regression toward the mean is employed for finding Regression issues, whereas provision regression is employed for finding the classification issues. In provision regression, rather than fitting a regression curve, we tend to match associate "S" formed provision operates, that predicts 2 most values (0 or one). The supply regression forms 3 types as below.

a) Binary supply regression (two attainable outcomes in an exceedingly DV)
b) Multinomial supply regression (three or additional categories in DV while not ordering)
c) Ordinal supply regression (three or additional categories in DV with ordering).

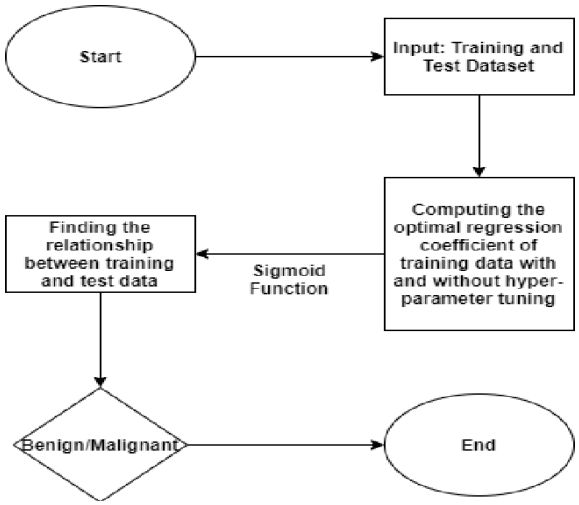


Fig.2 Flowchart for logistic regression

7.2 K- nearest neighbor algorithm

The k-nearest neighbor's algorithmic program, an addition called KNN or k-NN, may be a non-parametric, supervised learning classifier, that uses proximity to create classifications or predictions concerning the grouping of a private datum. whereas it is used for either regression or classification issues, it's usually used as a classification algorithmic program, operating off our belief that similar points are found close to each other.

For classification issues, a category label is allotted on the idea of a majority vote—i.e. the label that's most often diagrammatic around a given datum is employed. whereas this can be technically thought about as "plurality voting", the term, "majority vote" is additionally normally utilized in the literature. the excellence between these terminologies is that "majority voting" technically needs a majority of larger than five hun--dredth,which primarily works on their square measure solely 2 classes. after you have multiple classes—e.g. four classes, you don't essentially want five hundred of the vote to create a conclusion for a couple of classes; you may assign a category.

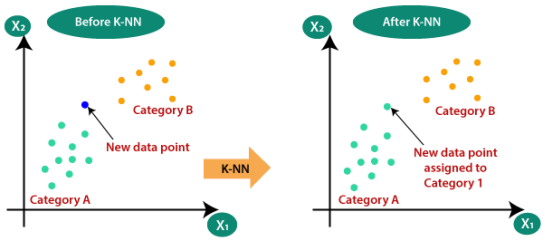


Fig 3. K-nearest algorithm

7.3 RANDOM FOREST CLASSIFICATION

The random forest rule is created from a group of call trees, and every tree within the ensemble is comprised of an information sample drawn from coaching set with replacement, referred to as the bootstrap sample. Of that coaching sample, a common fraction of it's put aside as check knowledge, called the out-of-bag (OOB) sample, which we'll come to later. Randomness is then injected into the dataset through feature material, reducing correlation among call trees and adding diversity to the dataset counting on the sort of downside, the deter-

mination of the prediction can vary. For an average of a regression task and a classification task, we are going to average the individual decision trees in the ensemble to yield a final, expected outcome.

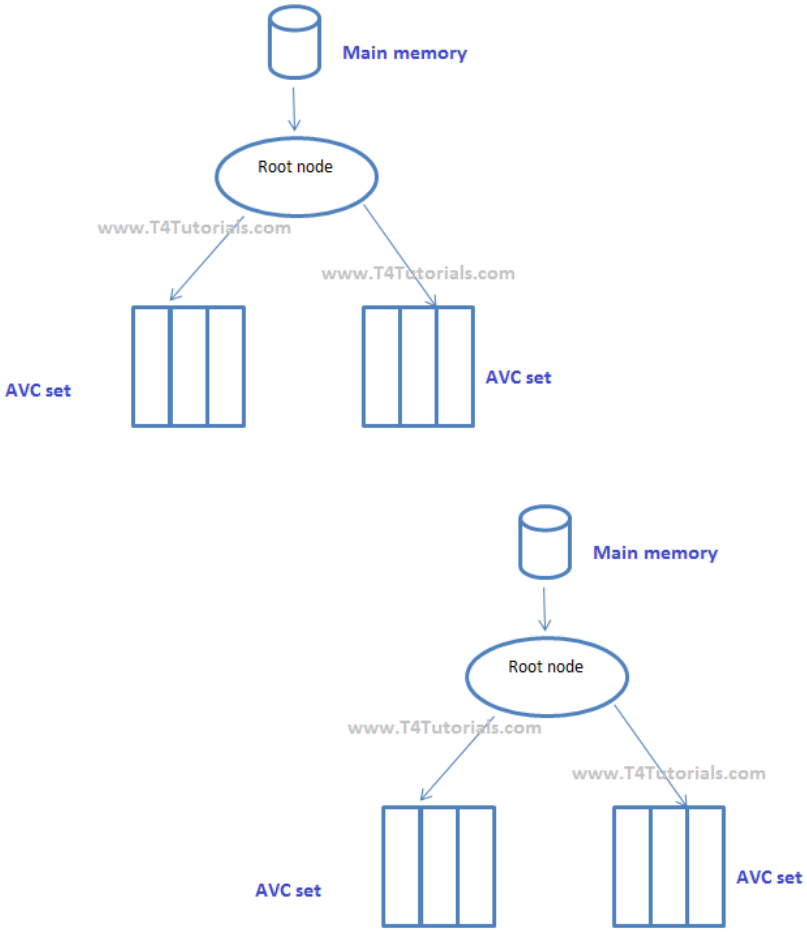


Fig 4 flowchart for rainforest classification

7.4DECSION TREE CLASSIFICATION

Decision Trees (DTs) square measure a non-parametric supervised learning methodology used for classification and regression. The goal is to form a model that predicts the worth of a target variable by learning easy call rules inferred from the info options. A tree is seen as a piecewise constant approximation.

For instance, within the example below, call trees learn from information to approximate a wave with a collection of if-then-else call rules. The deeper the tree, the a lot of advanced the choice rules and also the fitter the model.

Some blessings of call trees are:
Simple to grasp and to interpret. Trees are often envisioned.
Requires very little knowledge preparation. alternative techniques typically need knowledge of social control, dummy variables ought to be created and blank values to be removed. Note but that this module doesn't support missing values.
The cost of victimisation the tree (i.e., predicting data) is index within the range of knowledge points wont to train the tree.

Able to handle each numerical and categorical knowledge. However, the scikit-learn implementation doesn't support categorical variables for currently. alternative techniques area unit typically specialised in analyzing datasets that have only 1 style of variable. See algorithms for a lot of information.

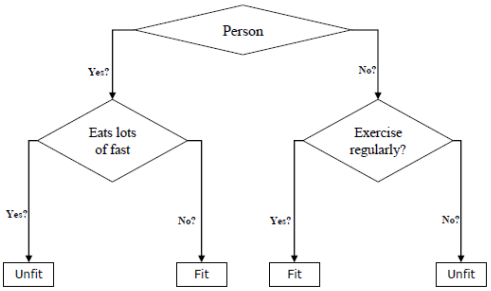


Fig 5. Decision tree classification

7.5MULTI-LAYER PERCEPTION

Multi-layer perception is additionally referred to as MLP. it's totally connected dense layers, that remodel any input dimension to the specified dimension. A multi-layer perception may be a neural network that has multiple layers. to make a neural network we tend to mix neurons along so the outputs of some neurons square measure the inputs of alternative neurons.

A gentle introduction to neural networks and TensorFlow is found here:

Neural Networks Introduction to TensorFlow.A multi-layer perceptron has one input layer and for every input, there's one neuron(or node), it's one output layer with one node for every output and it will have any variety of hidden layers and every hidden layer will have any variety of nodes.

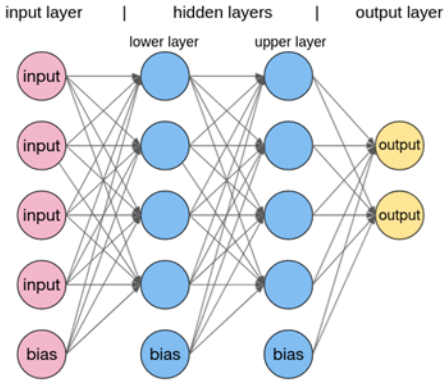


Fig 6. Multi-layer Perception

VIII. CONCLUSION AND FUTURE WORK

With the rising number of deaths due to heart disease, it is becoming increasingly important to build a system that can effectively and accurately forecast heart disease. The project's goal was to find the most efficient machine learning algorithm for detecting heart diseases. Using a typical dataset from kaggle, this study analyses the accuracy score of Decision Tree, Logistic Regression, Random Forest, MLP, and KNN algorithms for predicting heart disease. According to the findings of this study, the Random Forest algorithm is the most efficient algorithm for predicting heart disease, with a score of 100 percent accuracy. In the future, the study might be improved by creating a web application based on the Random Forest method and employing a larger dataset than the one used in this analysis, which would help to deliver better results and aid health professionals in successfully and efficiently forecasting heart disease.

With the rising number of deaths due to heart disease, it is becoming increasingly important to build a system that can effectively and accu-rately forecast heart disease. The project's goal was to find the most efficient machine learning algorithm for detecting heart diseases. Using a typical dataset from kaggle, this study analyses the accuracy score of Decision Tree, Logistic Regression, Random Forest, MLP, and KNN algorithms for predicting heart disease. According to the find-ings of this study, the Random Forest algorithm is the most efficient algorithm for predicting heart disease, with a score of 100 percent accuracy. In the future, the study might be improved by creating a web application based on the Random Forest method and employing a larger dataset than the one used in this analysis, which would help to deliver better results and aid health professionals in successfully and efficiently forecasting heart disease.

REFERENCES

[1]. A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[2]. A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ‘‘Using PSO algorithm for producing best rules in diagnosis of heart disease,’’ in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306– 311.

[3]. N. Al-milli, ‘‘Backpropogation neural network for prediction of heart disease,’’ J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

[4]. C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ‘‘Analysis of neural networks based heart disease prediction system,’’ in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[5]. P. K. Anooj, ‘‘Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,’’ J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 27–40, Jan. 2012. doi: 10.1016/j.jksuci.2011.09.002.

[6]. L. Baccour, ‘‘Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets,’’ Expert Syst. Appl., vol. 99, pp. 115–125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.

[7]. C.-A. Cheng and H.-W. Chiu, ‘‘An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide 81552 VOLUME 7, 2019 S. Mohan et al.: Effective Heart Disease Prediction Using Hybrid ML Techniques database,’’ in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566–2569.

[8]. H. A. Esfahani and M. Ghazanfari, ‘‘Cardiovascular disease detection using a new ensemble classifier,’’ in Proc. IEEE 4th Int. Conf. Knowl.- Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.

[9]. F. Dammak, L. Baccour, and A. M. Alimi, ‘‘The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,’’ in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), vol. 9, Aug. 2015, pp. 1–8.

[10]. R. Das, I. Turkoglu, and A. Sengur, ‘‘Effective diagnosis of heart disease through neural networks ensembles,’’ Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, May 2009. doi: 10.1016/j.eswa.2008.09.013.

[11]. M. Durairaj and V. Revathi, ‘‘Prediction of heart disease using back propagation MLP algorithm,’’ Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235–239, 2015.

[12]. M. Gandhi and S. N. Singh, ‘‘Predictions in heart disease using techniques of data mining,’’ in Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl. Manage. (ABLAZE), Feb. 2015, pp. 520–525.

[13]. A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ‘‘Prediction of heart disease using machine learning,’’ in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278. [14] B. S. S. Rathnayakc and G. U. Ganegoda, ‘‘Heart diseases prediction with data mining and neural network techniques,’’ in Proc. 3rd Int. Conf. Converg. Technol. (I2CT), Apr. 2018, pp. 1–6.

[14]. N. K. S. Banu and S. Swamy, ‘‘Prediction of heart disease at early stage using data mining and big data analytics: A survey,’’ in Proc. Int. Conf. Elect., Electron., Commun., Comput. Optim. Techn. (ICEECCOT), Dec. 2016, pp. 256–261.

[15]. J. P. Kelwade and S. S. Salankar, ‘‘Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series,’’ in Proc. IEEE 1st Int. Conf. Control, Meas. Instrum. (CMI), Jan. 2016, pp. 454–458.

[16]. V. Krishnaiah, G. Narsimha, and N. Subhash, ‘‘Heart disease prediction system using data mining techniques and intelligent fuzzy approach: A review,’’ Int. J. Comput. Appl., vol. 136, no. 2, pp. 43–51, 2016.

[17]. P. S. Kumar, D. Anand, V. U. Kumar, D. Bhattacharyya, and T.-H. Kim, ‘‘A computational intelligence method for effective diagnosis of heart disease using genetic algorithm,’’ Int. J. Bio-Sci. Bio-Technol., vol. 8, no. 2, pp. 363–372, 2016. [

[18]. M. J. Liberatore and R. L. Nydick, ‘‘The analytic hierarchy process in medical and health care decision making: A literature review,’’ Eur. J. Oper. Res., vol. 189, no. 1, pp. 194–207, 2008.

[19]. T. Mahboob, R. Irfan, and B. Ghaffar, ‘‘Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics,’’ in Proc. Internet Technol. Appl. (ITA), Sep. 2017, pp. 110–115. [21] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, ‘‘Computational intelligence for heart disease diagnosis: A medical knowledge driven approach,’’ Expert Syst. Appl., vol. 40, no. 1, pp. 96–104, 2013. doi: 10.1016/j.eswa.2012.07.032.

[20]. J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, ‘‘Association rule mining to detect factors which contribute to heart disease in males and females,’’ Expert Syst. Appl., vol. 40, no. 4, pp. 1086–1093, 2013. doi: 10.1016/j.eswa.2012.08.028.

[21]. S. N. Rao, P. Shenoy M, M. Gopalakrishnan, and A. Kiran B, ‘‘Applicability of the Cleveland clinic scoring system for the risk prediction of acute kidney injury after cardiac surgery in a South Asian cohort,’’ Indian Heart J., vol. 70, no. 4, pp. 533–537, 2018. doi: 10.1016/j.ihj.2017.11.022.

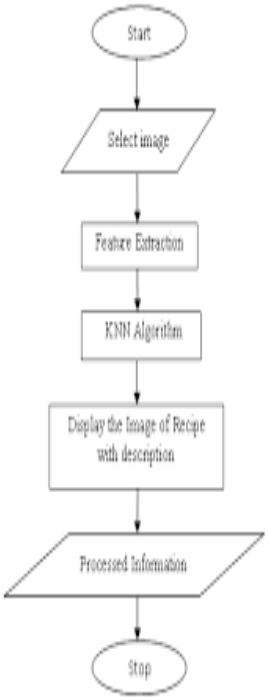
[22]. T. Karayılan and Ö. Kılıç, ‘‘Prediction of heart disease using neural network,’’ in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Antalya, Turkey, Oct. 2017, pp. 719–723.

[23]. J. Thomas and R. T. Princy, ‘‘Human heart disease prediction system using data mining techniques,’’ in Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT), Mar. 2016, pp. 1–5.

[24]. C. Raju, ‘‘Mining techniques,’’ in Proc. Conf. Emerg. Devices Smart Syst. (ICEDSS), Mar. 2016, pp. 253–255.

[25]. D. K. Ravish, K. J. Shanthi, N. R. Shenoy, and S. Nisargh, ‘‘Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural networks,’’ in Proc. Int. Conf. Contemp. Comput. Inform. (IC3I), Nov. 2014, pp. 1–6.

[26]. F. Sabahi, ‘‘Bimodal fuzzy analytic hierarchy process (BFAHP) for coronary heart disease risk assessment,’’ J. Biomed. Informat., vol. 83, pp. 204–216, Jul. 2018. doi: 10.1016/j.jbi.2018.03.016. [29] M. S. Amin, Y. K. Chiam, K. D. Varathan, ‘‘Identification of significant features and data mining techniques in predicting heart disease.


Start
Select image
Feature Extraction
KNN Algorithm
Display the Image of Recipe with description
Processed Information
Stop