International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195247

Heart Disease Prediction System Using K-Nearest Neighbor Classification Technique

Sowbarnica V. S*[1], Vismaya V[1], Vidhyapoonthalir M[1], Dr. S. Bhuvana[2]

[1]Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

[2]Associate Professor, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

The heart is an operating system of the human body. If it does not function properly it will affect other parts also. Heart disease problem describes a range of conditions that affect the heart. The existing system uses Support Vector Machine (SVM), it takes more time to train the database. propose a system for heart disease prediction. The Proposed system includes the various phases such as preprocessing , feature Extraction and classification.The input data are Pre-Processed using Min-Max scalar and Normalization,Feature extraction by PSO algorithm, Classification of data using K-Nearest Neighbour. The experimental results show that the proposed system obtained better Accuracy than existing methods.

Keywords : *Heart disease,pre-processing, PSO(Particle Swarm Optimization),KNN(K-Nearest Neighbour).*

I. INTRODUCTION

There is no scarcity of records regarding medical problems of patients regarding heart strokes. However the potential they have- to help us predict similar possibilities in apparently healthy adults are going unnoticed. For instance: According to the Indian Heart Association, 50% of heart strokes occurs under 50 years of age and 25% of all heart strokes occurs under 40 years in India. Urban population is thrice as endangered to heart attacks as rural population. [9].This thus propose to collect pertinent data concern all elements related to the field of study, train the data according to the proposed algorithm of machine learning(K-nearest neighbour) and find how strong is there a possibility for a patient to undertaking a heart disease. Most of the hospitals admitted in heart disease patient, this disease mostly affected in male because smoking habits[2,5].The below sections will discusses about the system design of Particle swarm optimization and K-nearest neighbour classifier how it works for predicting the heart disease in humans. And also expounds its workflow and structure.

The rest of the paper is organized as . The Literature review discussed in section 2 . The proposed model described in section 3.The experimental results are shown in section 4 and finally the conclusion of the proposed method discussed in section 5.

II. RELATED WORKS

Previously several methods have been proposed for heart beat prediction. In [1], Ali. Adeleetal. performed a work on the diagnosis of heart disease by designing a fuzzy expert system. The designed system is based on the V.A. Medical Center, Longs Beach and Cleveland Clinic Foundation database. The system has 13 input fields and one output field

where some input fields are blood pressure, sex, chest pain type, cholesterol, heart rate, resting electrocardiography (ECG), and thallium scans.

In [7], Tahmida Tabassum, discussed a work where Electrocardiogram (ECG) gives useful information about morphological details of heart which is used to find various cardiac diseases. It deals with a method of detecting cardiac diseases using support vector machine (SVM) is proposed. In the proposed model disease are modelled using the time domain features of ECG signal which are extracted using BIOPAC software. Raw ECG signal contains these useful features which can be used to predict cardiac arrhythmia. The various ECG parameters like heart rate, QRS complex, PR interval, ST segment elevation, signal are used for analysis. Based on these parameters of ECG's, different heart disease like atrial fibrillation, sinus tachycardia, myocardial infarction are detected.

In [3] Gomathy, B. (2014) performed regarding the healthcare industry collecting large amounts of health related information which cannot be mined to find unknown information for efficient evaluation. Heart disease is a term for defining a large amount of health conditions that are related to the heart. This medicinal condition directly control all the parts of the heart. Different data mining techniques like association rule mining, classification, clustering are used to predict the heart disease in health industry .The heart disease database is preprocessed to make the mining process even more efficient.

In [4] Kavitha, R (2016) stated ,in the classification of the heart disease data set a high dimensional data set is used in the pre processing stage of data mining technique. This raw dataset consist of redundant and inconsistent data thereby maximizes the search space and storage of the data. To achieve the classification accuracy we need to eliminate the redundant and the irrelevant data present. The dimensionality reduction technique is used to compress the high dimensional data to lower dimensional data with few constraints.

In [6] Ravish, D. K,(2014) performed the below

.Heart Attacks are the major cause of death in the world now-a-days, particularly in India. The need to predict this is a major thing for improving the country's healthcare sector. Accurate and precise prediction of the heart disease majorly depends on Electrocardiogram (ECG) data. These data's must be fed to a nonlinear disease prediction design. This nonlinear heart monitoring module must be able to detect arrhythmias such as tachycardia, bradycardia, myocardial infarction, atrial, ventricular fibrillation.

2.1. MOTIVATION

The existing system using Support Vector Machine(SVM)[3], it propose a system for heart disease prediction. It was not provide accurate results and taks more time to train the database images[5]. The heart beat parameter in ECG signal is noticed and mean heart rate, standard deviation and frequency domains (e.g., LF and HF powers) are derived and also faces some issues like less efficiency, takes more time to predict[2,5].By contrast, the proposed method with K-Nearest Neighbour is applicable for an important step towards an accurate, reliable and early detection system for heart problem.

III. PROPOSED WORK

The goal of the proposed system is to provide the system with an efficient heart disease prediction. Figure 1 shows the overall process of proposed model.The proposed system consists of the following phases:Pre-Processing of the input data with Min-Max scalar and Normalization, Feature extraction by PSO algorithm, Classification of data by K-Nearest

Neighbour mainly applicable when the dimensionality of the inputs is high.

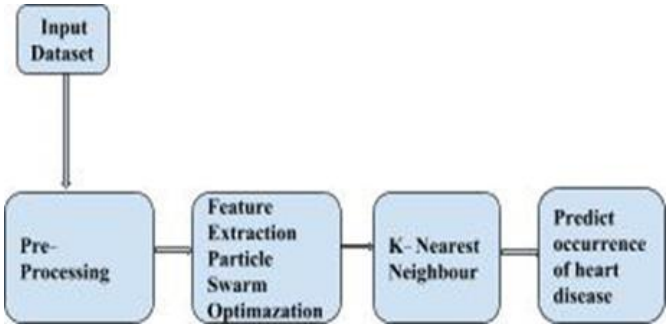


Figure 1. Overall process of proposed system

Reading the dataset file and then it is pre processed by using the min-max scalar and normalization technique to prepare them to the next process Feature extraction. The Particle swarm optimizer (PSO) is then done to find the nearest points and finally in the K-Nearest Neighbour (KNN) algorithm the k value is set to 3 because it is an essential to be set in odd range.

3.1. Pre-Processing

Pre-processing is nothing but the alignment of the data which is the most significant role in data-mining. The pre-processing technique is used to examine the dataset and it clears the missing data. The training phase in the Data Mining during Knowledge Discovery will be difficult if the data contains irrelevant or redundant information or more noisy and unreliable data.

The Pre-processing technique is summarized into following steps:

- The dataset is first collected and 70% of the data is split up as train data and 30% of the data is split up as test data.
- Once the pre-processing is done ,there will be so many missing data's .The empty values in the dataset in should be allocated as NA values which is done by the Min-max scalar algorithm.
- To normalise the data's to scalar range, min max algorithm is applied.

3.2. Minmax scalar

Transforms features by scaling each feature to a given data range. This estimator scales and translates each featured heart disease data set individually such that in the given range on the training set. An alternative approach to Z-score normalization is the so-called Min-Max scaling (often also simply called "normalization" is a common cause for ambiguities). In this approach, the data is scaled to a particular fixed range - usually 0 to 1. The cost of having this bounded limit - in contrast to standardization - is that will end up with smaller standard deviations, which suppress the effect of outliers.

A Min-Max scaling is typically done by the following equation:

$$Xsc=(X–Xmin)/ (Xmax–Xmin.) \quad (1)$$

1. minimum: 0 set as default. Lower bound after the transformation.
2. maximum: 1 set as default. Upper bound after the transformation.

3.3. PSO ALGORITHM

Particle swarm optimization a population based optimization technique. The system is initialized with a population of random solutions and searches for optimum value by updating the generations. However, unlike others, PSO has no evolution operators like crossover and mutation. In PSO, the

potential solutions, called particles, fly through the problem space by following the current particles.

The advantages of PSO are that PSO is easy to implement. PSO has been successfully applied in many parts: function optimization, artificial neural network training, fuzzy system control, and other areas where GA can be implemented.

In every iteration, each particle is updated with two "best" values. The first one is the best solution (fitness) it has achieved till now. This value is popularly known aspbest. Another "best" value that is then tracked by the particle swarm optimizer is the best value, obtained so far by any particle in populations. This best value is a global best and known as gbest. When a particle takes part of the population as its topological neighbors, the best value is a local best and is known as lbest.

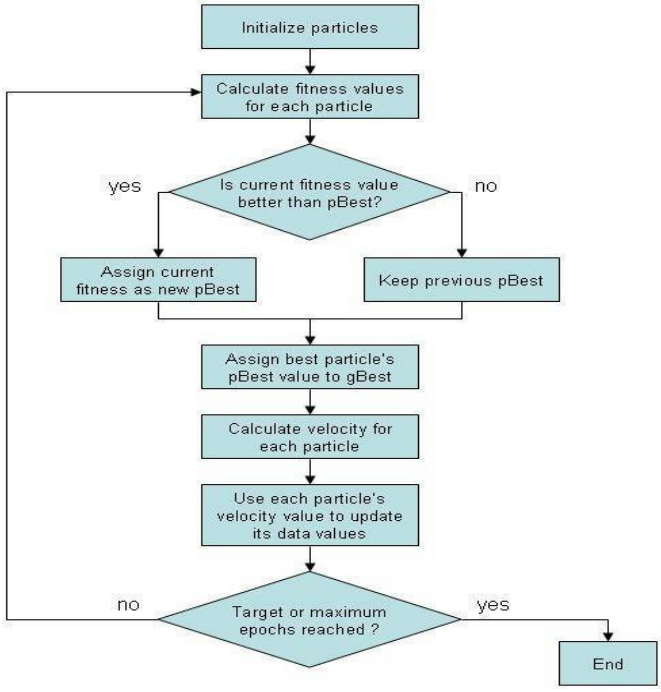


Figure 2. PSO optimization steps

In Figure 2 initially particles were mapped and the fitness value for each particle is found and it is initially fixes some value as a pBest and checks the founded fitness value is better than the pBest if yes, assigns current fitness as new pBest else keeps the pBest value same. Then after applying the case it assigns best particles pBest value to gBest. After the assigning process it calculate the velocity for each particle and updates the data. Finally it checks the target reached if yes,it ends the process else starts the process from the beginning.

3.4. K-Nearest Neighbour(KNN) Classifier

KNN is a non-parametric supervised learning model in which it is used to classify the data's to a given category with the help of training set. Predictions are made for a new instance (x) by searching through the entire training dataset for the K similar cases (neighbors) and summarizing the output variable for those K cases. In classification this is the mode for class value.Its purpose is to use a dataset in which the data points were separated into several classes to predict the classification of a new sample data.

Classification steps :-

1. Training phase: a model is built from the training instances.
   - ✓ classification finds relationships between predictors and targets.
   - ✓ relationships are summarised in models
2. Testing phase: tests model on a test sample whose class labels are known but not used as training the model
3. Usage phase: use the models for classification on new data's whose class labels are unknown

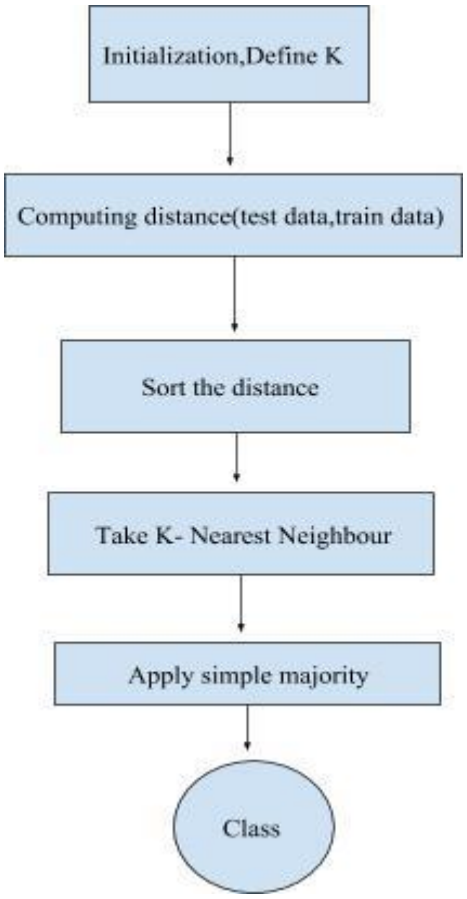


Figure 3. KNN classification steps

In the Figure 3 of KNN diagram, initially the K value must set to be in odd range and the K value will find the distance of the nearer datas of the person. It will consider all the near values and finally by determining the majority it will predict the occurence of heart disease in human.

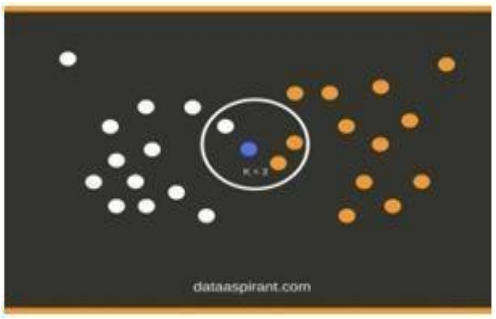
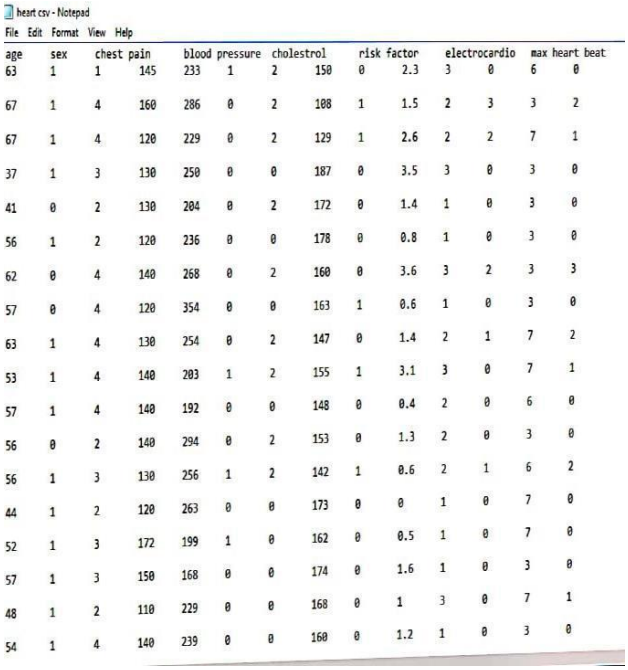


Figure 4. Working of K- Nearest neighbor algorithm

In Figure 4 shows the different target classes as white and orange circles. It contains totally 26 training samples. it predict the target class in blue circle. By setting k value as 3, The system calculated the similarity between input data and training samples using Euclidean distance metrics

IV. RESULTS AND DISCUSSION

The proposed heart disease prediction system is implemented using python 2.7.15 shell (32 bit) console.



heart csv - Notepad

File Edit Format View Help

age	sex	chest pain		blood pressure		cholestrol		risk factor		electrocardio		max heart beat	
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0

Figure 5. Kaggle database images

The heart disease dataset is collected from kaggle.com is shown in Figure 5. It consists of Person age, sex, heart beat count, liquor consumption, heart beat rate before doing exercise and after doing exercise, BP rate and smoking consumption.

There are 13 attributes used for predicting the person is incurred with the heart disease or not.The csv values are added to the program by using Imputer and numpy which is a formal method of using csv into python program.

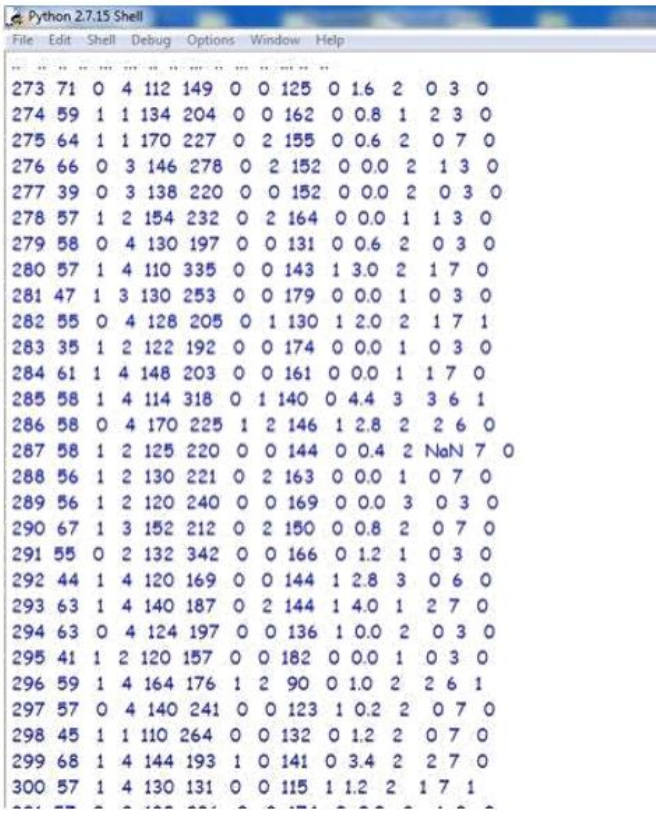


Figure 6. Experimental results After min-max scalar

The result after applying Min-Max scalar is shown in Figure 6 and first imported numpy for importing the csv file. The numpy array is then created and the data is imported and element is a different training point and each element is a feature. Then fits and finds min value and max value and then this finds the empty space in the dataset and replace the data set with the NA value Finally the NA value is replaced with the minimum and the maximum values.

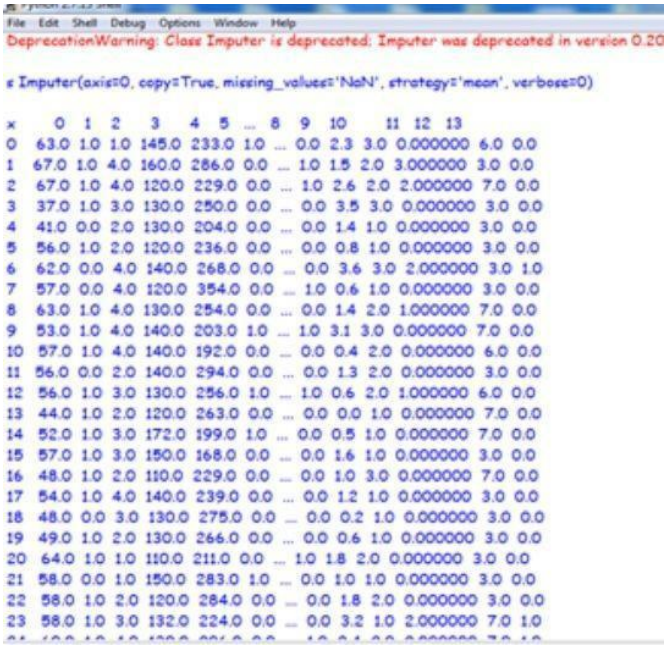


Figure 7. Resuls after applying PSO

The results after applying PSO algorithm is shown in Figure 7.

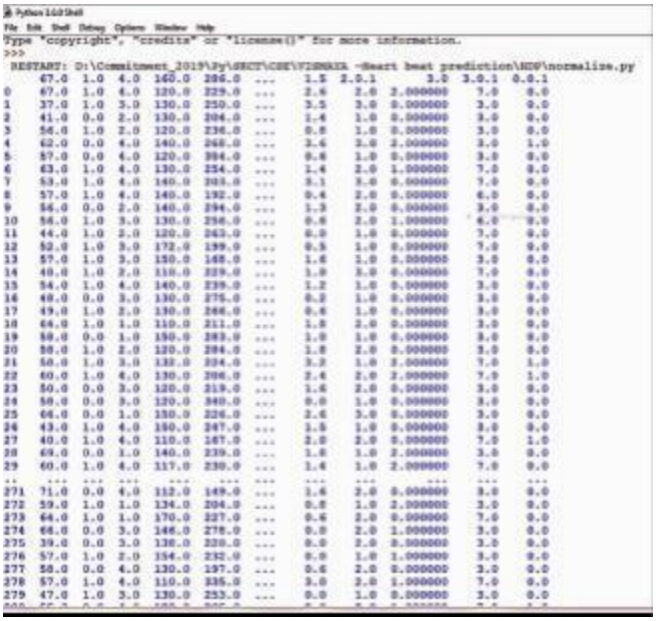


Figure 8. Output of KNN classifier

The results after applying all the stages of Heart desease has been shown in Figure 8.

Initially the K value should be set in odd range here the K value is used as 3.The persons whose attributes nearer to this value will be showed as 1 and the person who are not having heart problem will be indicated as 0.

V. CONCLUSION

The Heart Disease Prediction System using K-Nearest Neighbour Classifier provides its users with a prediction result that gives the state of a user leading to heart problem. The algorithm gives the nearby reliable outputs based on the input provided. If the number of people using the system get increased, then the awareness about their current heart status will be known and the rate of people dying due to heart diseases will get reduce eventually. In future, the numbers of attributes could be reduced and accuracy would be increased using some other algorithms. Furthermore, the experimental results show that the system The proposed system significantly improve the heart disease prediction rate using KNN algorithm than SVM method.

VI. REFERENCES

[1]. A. Adeli and M. Neshat, "A fuzzy expert system for heart disease diagnosis," in Proceedings of International Multi Conference of Engineers.

[2]. Alkeshuosh, A. H., Moghadam M. Z., MansooriI. A., &Abdar, M. (2017). Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease. 2017 International

[3]. Conference on Computer and Applications (ICCA).doi:10.1109/comapp.2017.807978.

[4]. Ammar Aldallal, Amina Abdul Aziz Al-Moosa,.(2018) Using Data Mining Techniques to Predict Diabetes and Heart Diseases,.

[5]. Banu, M. A. N., &Gomathy, B. (2014). Disease Forecasting System Using Data Mining Methods. 2014 International Conference on Intelligent Computing Applications.doi:10.1109/icica.2014.36.

[6]. Cincy Raju, Philipsy E, Siji Chacko., Padma Suresh., Deepa Rajan S.(2018, March). A Survey on Predicting Heart Disease using Data Mining Techniques.

[7]. Jagdeep Singh, Amit Kamra, Harbhag Singh. (2016). Prediction of Heart Diseases Using Associative Classification.

[8]. Kavitha, R.,&Kannan, E. (2016). An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS).doi:10.1109/icetets.2016.7603000.

[9]. Monika Gandhi , Shailendra Narayan Singh. (2015).Predictions in Heart Disease Using Techniques of Data Mining.

[10]. Philipsy E, Siji Chacko, 4 Padma Suresh, 5 Deepa Rajan S.(2018, March) . A Survey on Predicting Heart Disease using Data Mining Techniques.

[11]. Prasanna Lakshmi, K., Reddy, C.R.K. (2015). Fast Rule-Based Heart Disease Prediction using Associative Classification Mining.

[12]. Rajathi,S.,Radhamani,G. (2016,March) Prediction and analysis of Rheumatic heart disease using KNN classification with ACO.In Data Mining and Advanced Computing (SAPIENCE),international conference on(pp. 6873).

[13]. Ravish, D. K., Shanthi, K. J., Shenoy, N. R., & Nisargh, S.(2014). Heart function monitoring, prediction and prevention of Heart Attacks: Using Artificial Neural Networks.2014 International Conference on Contemporary Computing and Informatics (IC3I).doi:10.1109/ic3i.2014.7019580.

[14]. Suvarna,C.,Sali,A.,&Salmani,S.(2017).Efficient heart disease prediction system using optimization technique.2017 International Conference on Computing Methodologies and Communication(ICCMC).doi:10.1109/iccmc.2017.8282.

[15]. Tabassum,T., & Islam,M. (2016). An approach of cardiac disease prediction by analyzing ECG signal. 2016 3rd International Conference on Electrical Engineering and

[16]. Information Communication Technology (ICEEICT).doi:10.1109/ceeict.2016.7873093.

[17]. Zhang,Y ., Liu, F., Zhao,Z., Li., D., Zhou,X., & Wang, J (2012,June). Studies on application of Support Vectore Machine in diagnosis of coronary heart disease . In Electromagnetic Field Problems and Applications (ICEF), 2012 Sixth International Conference on(pp.1-4).

Cite this article as :

Sowbarnica V. S, Vismaya V, Vidhyapoonthalir M, Dr. S. Bhuvana, "Heart Disease Prediction System Using K- Nearest Neighbour Classification Technique", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 229-236, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195247
Journal URL : http://ijsrcseit.com/CSEIT195247