Morality Prediction Model in Cardiovascular Disease with Significant Feature Selection and Hybrid KNN Classification Technique

C.Sowmiya, P. Sumitra

***Abstract:** Nowadays morality rate is increasing globally due to heart disease. It is one of the leading health risk facing men today. So early detection of heart disease assist the patients to maintain a healthy life style. Several techniques are used in the medical field to detect or diagnose disease in view of patient family health history and some other aspects. However, developing a system to predict the heart diseases without any medical tests is still challenging. Machine learning (ML) approaches is suitable and effective in providing decision and prediction from enormous health care data. Several previous researches provide an overall view in ML methods for disease prediction but the accuracy of prediction is still needed to be improved. In this study, a novel framework is presented that intent at removing the unwanted features with Bacterial Colony Optimization algorithm and applies the Hybrid KNN algorithm with great accuracy in identifying the heart disease. This prediction model is developed with UCI Cleveland dataset with several known classification approaches. An enhanced model is presented with 99.83% accuracy in heart disease prediction. The presented study is compared with other classification approaches.*

***Keywords:**Data mining, Machine learning, Feature selection, Heart disease, Hybrid KNN.*

I. INTRODUCTION

Based on the data provided by world health organization (WHO), yearly 17.9 million humans are died because of cardiovascular diseases (CVDs) [1-2]. In India mortality occurs between the 30–59 years age-group people which are twice than that in the US [3]. Therefore the early prediction of this health issue may save human life and helps to reduce the death rate. Data mining (DM) provides a new way to cardiovascular disease prediction. Several DM approaches are utilized to found and extract required data from the medical dataset with least user inputs [4]. Researchers seek several ways to utilize the data mining in medical data to obtain an accurate prediction of cardiovascular diseases [5-8].

In order to use the machine learning algorithm in health care data it is important to remove the unwanted features which degrades the prediction accuracy [9, 10]. Therefore, a suitable attribute selection approach is required to obtain a great result in prediction. Selecting the appropriate feature selection approach is still a challenging issue among the researchers [11]. However, it is not easy to identify the proper technique and select the significant features. Hence the proposed system introduces the Bacterial Colony Optimization algorithm in finding the finest combination of essential features among the 13 attributes that work well with the classification algorithms. The Knn classification algorithm is enhanced in the present study to improve the prediction accuracy. Along with the Euclidean distance parameters such as class center, distance, Item Weight to Class and Item's strength are added to evaluate the disease risk level. The experiment is performed on the UCI Cleveland dataset generally used by the ML researchers. Six important features were selected and applied five classification algorithms to develop the prediction system. The capability of the proposed approach is evaluated by comparing the maximum accuracy attained by the HKNN technique against the maximum accuracy attained in the former research.

The remaining section of this research is arranged as: section II describes the heart related research implemented by ML approaches. Section III and IV discuss the problem statement and dataset description respectively. Section V presents the classification techniques applied in the proposed system. Section VI describes the experimental setup and result obtained in the study. Eventually section VII provides the conclusion.

Related works

The existing work which is closely related to the proposed study is reviewed in this section. The following research work is implemented on the Cleveland dataset.

Saxena et al [12] present a heart disease prediction framework by the classification rule based on decision tree structure. This approach predicts the risk level of patients based on their health parameters and assists **non**-expert doctors to make correct decision about the patient heart disease risk level with the accuracy of 86.3 % in testing phase and 87.3 % in training phase. P.K. Anooj [13] introduced a weighted fuzzy rule-based prediction system. The researchers create two important factors such as automatic fuzzy rules creation and the weighted procedure which provides a powerful fuzzy system for heart disease prediction. The author implemented this system on Cleveland, Hungarian and Switzerland dataset with the accuracy of 0.6, 0.4 and 0.5 percentage respectively.

T. Santhanam and E.P. Ephzibah [14] developed a prediction model based on Principal Components Analysis (PCA) as a feature selection method and feed forward neural network as a classifier with an accuracy of 95.2%.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

C.Sowmiya *, Ph.D Research Scholar, PG and Research Department of Computer Science and Applications Vivekananda College of Arts and Sciences for Women (Autonomous), Elayampalayam, Tiruchengode-637205, Tamilnadu, India. sowmiyac83@gmail.com

P. Sumitra, Assistant Professor in PG and Research Department of Computer Science and Applications, Vivekanandha College of Arts and Sciences for Women, Elayampalayam, Tiruchengode(TK), Namakkal(DT), TamilNadu, India.

Authors compared this system with regression model that obtain 92.0% accuracy.V. Krishnaiah et al [15] present a model for heart disease prediction by removing the redundancy of the data using Fuzzy K-NN classifier with the accuracy of 91%. A combination of genetic algorithm and recurrent fuzzy neural networks based classifier is applied on Cleveland data and achieved 97.78% accuracy [14].

Seyedamin Pouriyeh et al [16] in his study a 10 fold cross validation is applied on the heart disease data using classification algorithms with boosting technique where Support Vector Machine provides the better accuracy of 84.15 %. Tiwaskar, S.A et al [17] applied the Convolution Neural Network on Cleveland dataset to provides an efficient prediction model. The author makes a statistical study on 13 attributes and filters the 6 highly correlated attribute (cp, exang, oldpeak, slope, ca, thal) and obtain the highest accuracy of 93%. Shan Xu et al [18] combined the Correlation-based Feature Selection approach with Best-First-Search to reduce the unwanted features in dataset. The authors applies various machine learning approaches like RF, SVM, C4.5, Naïve Bayes, RBF and Ada boost and shows that Random forest obtain a better accuracy with 91.6% than others.

The aim of the proposed study is to compare almost all the machine learning approaches which have been applied in the previous study to provide a most suitable heart disease prediction model with good accuracy.

II. PROBLEMS IDENTIFIED

The Discovery of early stage of heart disease prevents the human from death. But in most of the case it is identified in final stage or after death. Therefore the early detection is significant and performed by the health care organizations. Data mining is the suitable technique to identify the disease that helps the doctor to save the patient by proper diagnosis. There are several method are available in data mining which can be applicable for disease prediction. Further very less features are utilized in the prior research. The problem in heart prediction model is to select the appropriate algorithm that provides better accuracy with less time.

III. DATASET DESCRIPTION

The present study utilized the Cleveland database from UCI repository for heart disease prediction. In UCI four heart related database (Cleveland, Hungary, Switzerland, and the VA Long Beach) is available. But Cleveland database is selected for this research because several former researchers used machine learning algorithm on these database. Dr. Robert Detrano collected these data from three health care industries with 303 patient's records. 6 patients data contains missing values so 297 patient data is taken in this study. The dataset now contains only 14 attributes whereas it has 76 attributes while collecting it. But the researches use only 14 attributes for experiments. Hence 14 attributes with 297 patient's data is applied in the present experiment.

IV. CLASSIFICATION TECHNIQUES

A. Support Vector Machine

SVM is a supervised machine learning algorithm mainly used for classification issues [19]. Let the training samples having dataset $= \{p_i, q_i\}$;$i = 1,2,\cdots,n$ where $p_i \in R^n$ denotes the i^{th} vector and $q_i \in R^n$ represent the target item. The linear SVM finds the optimal hyper plane of the form $f(x) = d^t p + b$ where d is a dimensional coefficient vector and b is a offset. This is done by solving the subsequent optimization problem:

$$min_{w,b,\varepsilon_i} \frac{1}{2} d^2 + g \sum_{i=1}^{n} \varepsilon_i \qquad (1)$$

$$s.t \begin{cases} q_i(dp_i + b) \geq 1 - \varepsilon_i \\ \varepsilon_i \geq 0, i = 1,\cdots,k \end{cases} \qquad (2)$$

Where g is the penalty variable ε_i and $i = 1,\cdots,k$ are the set of slack variables. Through this variable SVM reduce the classification errors.

B. K-Nearest Neighbors

KNN is a non-parametric ML approach used for heart disease classification. The KNN depends on the similar observations occur in closeness [20]. The prediction result is provided based on the Euclidean distance $dis(x_i, x_j)$ among the samples and the k value.

$$dis(x_i, x_j) = \sqrt{(x_1 - x_2)^2 + \cdots + (x_k - x_k)^2} \qquad (3)$$

C. Artificial Neural Network

ANN is a powerful ML approach that has various computing features known as neurons that provides a interconnected system. [21].

These neurons are related with a link. Each link has a weight. Wt_i represent the weights for input x_i where $X = \{x_1, x_2, x_3, \cdots, x_n\}$, represent the k input applied to the neurons. bv_i is the bias value, y_i is the output of the neuron network. Each neuron has an activation function $(f(.))$ to define the output y_i. The sigmoid is activation function used in ANN computed by the below equation

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (4)$$

The steps proceed in ANN is listed below

$$nt_i = \sum_{j=1}^{k} wt_{i,j} * x_i + bv_i \qquad (5)$$

$$y_i = f_i(nt_i) \qquad (6)$$

$$SA_i = \frac{\partial f_i}{\partial nt_i} \qquad (7)$$

$$ter_i = ao_i - po_i \qquad (8)$$

Where: i is the index of neurons, SA_i is the slope of activation function$f$, ter_i is the training error at output i, ao_i is the actual output and po_i is the predicted output.

D. Random forest

RF is an efficient classification technique performed based on tree structure [22]. It is a collection of several algorithms thus develops a jungle with huge number of trees. RF provides the decision from a random model of the training set. The loop is repeated with various random samples and provides the final result based on majority voting. For a given data, $X = \{x_1, x_2, x_3, \cdots, x_n\}$with responses, $Y = \{x_1, x_2, x_3, \cdots, x_n\}$ which repeats the bagging from $b = 1$ to B.

V. PROPOSED METHODOLOGY

Netbeans IDE was used in this study to perform the experiment of proposed model because it provides a robust background for creating a analytic workflows. Figure 1 shows Architecture of the proposed prediction model. During the implementation the Cleveland dataset was imported into Netbeans. Initially feature engineering is performed to select various combinations of features and then applies the machine learning approach to create a prediction model. These processes are continued for all combination of the features. The loop is repeated on each subset with six features and applied classification model to it. The proposed prediction model performance is stored on each classification model and provides the output after the executing the whole process.

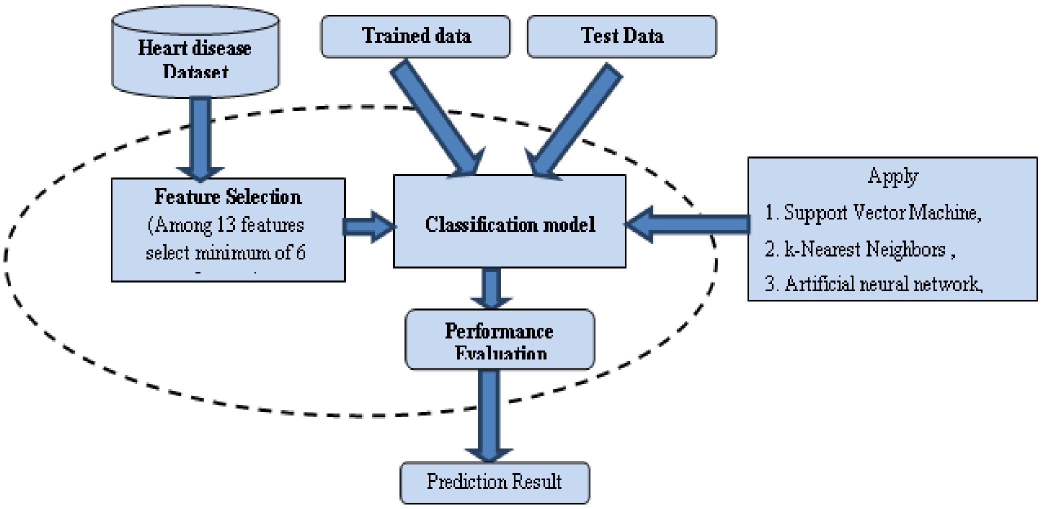

Figure 1. Architecture of the proposed prediction model

A. Feature Selection

Bacterial Colony Optimization (BCO)

BCO is an efficient optimization approach introduced by Niu et al., recently [23]. In BCO several optimization processes is reduced that makes implementation easier when compared to the bacterial foraging optimization (BFO) [24].

The chemotaxis is the most important operation performed in BCO. Chemotaxis contains two main approach such as running and tumbling approach. The running approach is responsible for generating a new point $\left(\theta_k(e)\right)$ based on the previous point $\left(\theta_k(e-1\right.$, personal best position $pbestk$, and globe best $gbest$. The tumbling approach is performed when running approach provide invalid result. It improves the variety and grouping. The term (tum_k) is random parameter utilized in tumbling operation to provide a better position.

***Running approach*:**

$$\theta_k(E) = \theta_k(E-1) + S_k \times \big(Gbest - \theta kE{-}1{+}1{-}Sk{\times}pbestk{-}\theta kE{-}1 \qquad (9)$$

***Tumbling approach*:**

$$\theta_k(E) = \theta_k(E-1) + S_k \times \big(Gbest - \theta_k(E-1)\big) + (1 - S_k) \times \big(pbest_k - \theta_k(E-1)\big) + ch \times tum_k \quad (10)$$

where ch denotes the chemotaxis process size. The constraint for randomness is mentioned by

$$tum_k = {}^{\Delta(k)}\Big/_{\sqrt{\Delta^e(k)\Delta(k)}} \qquad (11)$$

In the above equation $\Delta(k)$ denotes the direction angle of the k^{th} feature whose elements lies between -1 to 1. Variables S_1 and S_2 are two constants and their values are generated between 0 to 1.

The BCO was developed by the researchers to solve the optimization issue. But in our research BCO is utilized to select the combination of features for heart disease prediction model. In basic BCO population are changed based on sizes so feature selection is difficult task with high dimensional attributes. The operation such as reproduction, removal and distribution cannot be performed for feature selection. Hence it is modified by applying the multi-dimensional population concept instead of above mentioned operation as described in [24]. A better result is achieved through this feature selection method that provides the combination of six features as showed in table1.

B. Hybrid K-Nearest Neighbors

KNN is the efficient method in machine learning performed based on a numerical data. Multi-class issues can be solved by this approach which is a main advantage of Knn. In training phase Knn takes the decision based on the training items. The decision rule of KNN method determines firstly the *K* nearest items in the feature space; afterwards, it assigns the unclassified item to the class of the majority vote of those *K* items [25]. The implementation of Knn is depends on two main operators: *K* value and distance, which represents the neighbors count and distance between two items respectively. Euclidean distance is the popular function to calculate the distance among test item and training item which is represented in below equation.

$$dis(P,Q) = \sqrt{\sum_{x=1}^{n}(P_i - Q_i)} \qquad (12)$$

Where $dis(P,Q)$denotes the Euclidean distance among items *P* and *Q* of features (1, 2,…, v), P_i symbolizes the test item features, *P*, Q_i denotes the features of a particular training item *Q*, and *n* denotes the total count of features.

The disadvantages is that the result is depends upon the *K* value and the execution time for measuring the interval between test item and entire class item. Hence the Knn is modified by assigning weights for each and every item that reduces the execution time effectively similarly K value is automatically assumed by the system. Consider the two classes A and B in a two-dimensional space. The proposed HKNN is executed based on six major steps. The basic Knn is enhanced with some new parameters to properly decide heart disease risk level.

Inside a class each item depends on the distance among this sample and the center point which is calculated by

$$CN_z = \frac{1}{a}\sum_{j=1}^{a} H_{j,a} \qquad (13)$$

Where CN_z is the center of class *z* having *a* items expressed as $M = \{H_1, H_2, \dots, H_a\}$. Thirdly, compute the distance among each of the class items $H_j$and the class center CN_z, that is computed by

$$D(CN_z, H_j) = \sqrt{\sum_{i=1}^{y}\left(CN_{z,i} - H_{j,i}\right)^2} \qquad (14)$$

Where $D(CN_z, H_j)$ denotes the Euclidean Distance between class center CN_z and item $H_{j,i}$, and *m* is the features. And then calculate Item Weight to Class (IWtC) which is the distance from the class item to the center of the class. This value assigned to each class item to represent its importance and denoted by *w* (*i*, *C*). IWtC is calculated by the below equation.

$$w(H_j, CN_z) = \frac{1}{D(CN_z, H_j)} \qquad (15)$$

Where $w(H_j, CN_z)$ represents item $H_j$weight to class *X*. Fourthly, compute the distance $D_z$from class center $CN_z$to a test item *E* based on equation 16

$$Dis_z = \sqrt{\sum_{i=1}^{m}(I_i, CN_z)^2} \qquad (16)$$

Where $Dis_z$is the Euclidean distance between the class center $CN_z$and an incoming test item *I* and m is the features count. Therefore distance among *I* and those K is computed by equation 17.

$$D(I, H_j) = \sqrt{\sum_{i=1}^{m}(I_i, H_{j,i})^2} \qquad (17)$$

Where $D(I, H_j)$ are the Euclidean distance among a test item *I* and an item *Hj*, and m is the total number of feature. And then for each class, calculate $K_H$and $K_I$item's strength to test item *I*. Item's strength (*IH*) is computed according to Eq. (18)

$$IH(I, H_j) = \frac{1}{D(I, H_j)} * w(H_j, CN_z) \qquad (18)$$

Eventually, when the summation of IH is high that classes are selected by the hybrid Knn. Through this method the heart disease prediction model is developed to determine their risk level.

```
Input
Input dataset with m features with a set $F = \{f_1, f_2, f_3, \cdots, f_n\}$
Input Training dataset $Td = \{t_1, t_2, t_3, \cdots, t_n\}$
Output
Input data will classified as heart disease presented or not
Begin
For each sample $t_i \in D$ do
        $t_i$ in the dataset of m dimension
Next
For each target class $A \in tc$ do
        Find the center $CN_z$ of $A$ for $t_j \in tA = \{t_{1a}, t_{2a}, t_{3a}, \cdots, t_{na}\}$ as in eqn 16
                For each item $t_{ja} \in A \forall j \in \lfloor 1 \rightarrow total\ items\ in\ Aj$ do
                        Calculate distance among class $t_j$ and class center $CN_z$ using eqn 15
                        Calculate IWtc of item $t_j$ to class $A$ using eqn 16
Calculate the distance among $I$ and $CN_z$ using eqn 17
Calculate $IH$ of the item $A_J$ to test item I as eqn 18
$SS(A) += IH$
If $SS(A) += MAX(SS)$ Then
        The target class = A
End If
End
```

Experimental Result and Discussion

The performance evaluation of the study is discussed in this section. The proposed system is developed in java language. The Net Beans IDE is utilized for front end design. MYSQL is used for database access. The result of the proposed work is presented in Table 2 that shows significant features have improved the accuracy of classification process. The best heart disease prediction system is developed by utilizing the 6 important attributes as shown in table 1.

Table 1. Dataset and the selected attributes

Dataset	Selected attributes
Cleveland	Age, cp, fbs, restecg, exang, oldpeak

The comparison of various algorithms is performed on Cleveland dataset and their performance measures are listed in table 2.

Table 2. Comparison of classification accuracy of models

Models	Accuracy (%)	precision	Recall	F-measure
SVM	98.24	95.6	96.3	98.24
KNN	97.12	93.8	95.5	97.12
ANN	97.3	94.3	96.1	97.3
RF	98.41	95	97.7	98.41
HybridK-NN	99.83	96.1	97.4	99.83

The proposed approach performance is showed in figure 1. The accuracy of the H-knn is 99.83 whereas the accuracy of the SVM is 98.24.

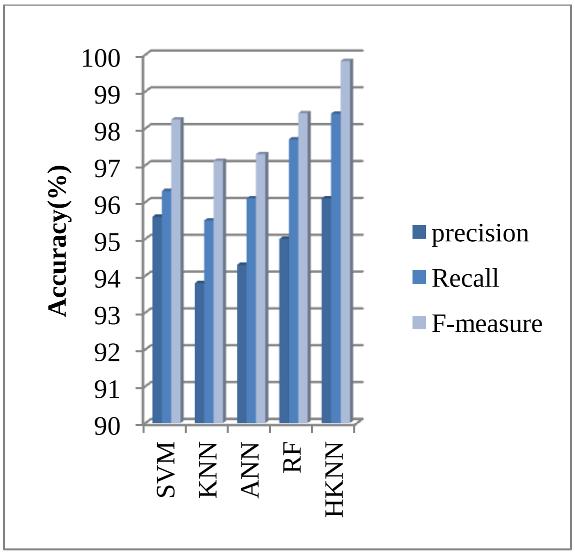

Figure 2. Performance comparison of the proposed model

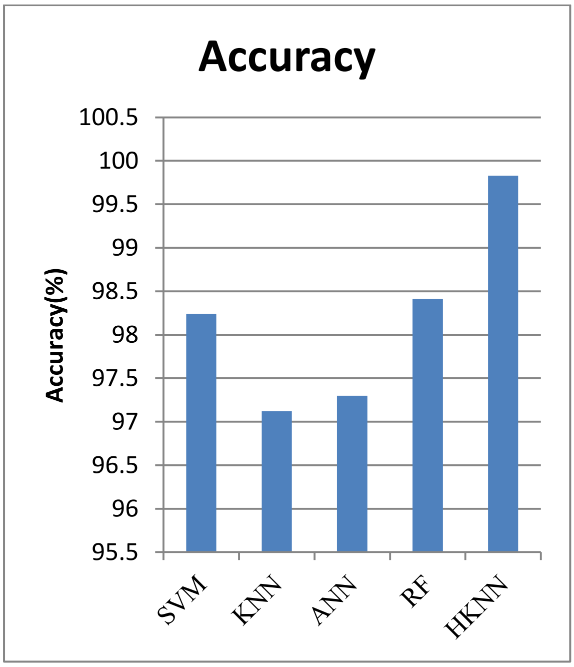

Figure 3. accuracy comparison of the proposed model

VI. CONCLUSION

An hybrid classification approach is introduced to found the heart disease in early stage. The Bacterial Colony Optimization with multi-level population is utilized to remove the unwanted features from the Cleveland dataset. Through this approach the 6 attributes are finalized for the classification process from 13 attributes. The Hybrid classification method provides a better result with 99.83 % accuracy compared to other techniques. The proposed model predicts the heart disease in an efficient manner and assists the non-specialist physicians to decide a right choice on the heart patient risk level.

REFERENCES

1. WHO. Cardiovascular Diseases (CVDs). [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases/
2. Mackay J, Mensah G. Atlas of heart disease and stroke. Nonserial Publication; 2004.
3. WHO. Cardiovascular Diseases (CVDs. South-East Asia http://www.searo.who.int/topics/cardiovascular_diseases/en/
4. Srinivas, K., Rani, B.K. and Govrdhan, A., 2010. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), pp.250-255.
5. Vijayakrishnan, Rajakrishnan, Steven R. Steinhubl, Kenney Ng, Jimeng Sun, Roy J. Byrd, Zahra Daar, Brent A. Williams, Christopher Defilippi, Shahram Ebadollahi, and Walter F. Stewart. "Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record." *Journal of cardiac failure* 20, no. 7 (2014): 459-464.
6. Ilayaraja, M. and Meyyappan, T., 2015. Efficient data mining method to predict the risk of heart diseases through frequent itemsets. *Procedia Computer Science*, *70*, pp.586-592.
7. Tayefi, Maryam, Mohammad Tajfard, Sara Saffar, Parichehr Hanachi, Ali Reza Amirabadizadeh, Habibollah Esmaeily, Ali Taghipour, Gordon A. Ferns, Mohsen Moohebati, and Majid Ghayour-Mobarhan. "hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm." *Computer methods and programs in biomedicine* 141 (2017): 105-109.
8. Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H. and Jahan, H., 2018. PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques. *Procedia computer science*, *132*, pp.1068-1076.
9. Schmidt, S.E., Holst-Hansen, C., Hansen, J., Toft, E. and Struijk, J.J., 2015. Acoustic features for the identification of coronary artery disease. *IEEE Transactions on Biomedical Engineering*, *62*(11), pp.2611-2619.
10. Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. A. H. 2016. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 145-150. IEEE.
11. Shouman, M., Turner, T., Stocker, R., 2013. Integrating clustering with different data mining techniques in the diagnosis of heart disease. J. Comput. Sci. Eng. 20(1).
12. Saxena, K. and Sharma, R., 2016. Efficient heart disease prediction system. *Procedia Computer Science*, *85*, pp.962-969.
13. Anooj, P.K., 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, *24*(1), pp.27-40.
14. Santhanam, T. and Ephzibah, E.P., 2013. Heart disease classification using PCA and feed forward neural networks. In *Mining Intelligence and Knowledge Exploration* (pp. 90-99). Springer, Cham.
15. Krishnaiah, V., Narsimha, G. and Chandra, N.S., 2015. Heart disease prediction system using data mining technique by fuzzy K-NN approach. In *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1* (pp. 371-384). Springer, Cham.
16. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H. and Gutierrez, J., 2017, July. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In *2017 IEEE Symposium on Computers and Communications (ISCC)* (pp. 204-207). IEEE.

17. Tiwaskar, S.A., Gosavi, R., Dubey, R., Jadhav, S. and Iyer, K., 2018, August. Comparison of Prediction Models for Heart Failure Risk: A Clinical Perspective. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.
18. Xu, S., Zhang, Z., Wang, D., Hu, J., Duan, X. and Zhu, T., 2017, March. Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(* (pp. 228-232). IEEE.
19. Y. Zhu, J. Wu, Y. Fang, Study on application of SVM in prediction of coronary heart
disease, J. Biomed. Eng. 30 (6) (2013) 1180–1185.
20. Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications, 3(5), 605-610.
21. Bishop, C. M. (1996). Neural networks: a pattern recognition perspective.
22. 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
23. B. Niu, H. Wang, Bacterial colony optimization, Discrete Dyn. Nat. Soc., 2012 (2012), pp. 343-361.
24. K.M. Passino, Biomimicry of bacterial foraging for distributed optimization and control, IEEE Contr. Syst. Mag., 22 (2002), pp. 52-67.
25. Wang, H., Tan, L. and Niu, B., 2019. Feature selection for classification of microarray gene expression cancers using Bacterial Colony Optimization with multi-dimensional population. Swarm and Evolutionary Computation, 48, pp.172-181.
26. Saleh, A.I., Shehata, S.A., Labib, L.M., 2017. A fuzzy-based classification strategy (FBCS) based on brain–computer interface. Soft comput. 1–25.

AUTHORS PROFILE

Sowmiya. C received her M.Phil Degree in Computer Science from Selvam Arts and Science College, (Autonomous) Namakkal, in the year 2013 and received her M.C.A Degree in Computer Applications from Sengunthar Arts and Science College, Tiruchengode, in the year 2012. She is doing her Full Time Ph.D (Data Mining) in PG and Research Department of Computer Science and Applications Vivekananda College of Arts and Sciences for Women (Autonomous), Elayampalayam, Tiruchengode-637205, Tamil Nadu, India. She published three papers in International conference and one papers in National Conference. she published five UGC Approved Journal and one IEEE Journal paper. Her research areas include Data Mining.

Sumitra. P received her Ph. D Degree in Computer Science from Mother Teresa Women's University, Kodaikannal, TamilNadu in the year 2013. She is presently working as an Assistant Professor in PG and Research Department of Computer Science and Applications, Vivekanandha College of Arts and Sciences for Women, Elayampalayam, Tiruchengode(TK), Namakkal(DT), TamilNadu, India. She is a life member of The Indian Science Congress Association. She is currently guiding 7 Ph.D Research Scholar and 1 M.Phil Research Scholar. Her research interests are in Image Processing, Soft Computing and Data Mining.