International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Real Coded Binary Artificial Bee Colony (RC-BABC) Based Feature Selection and Relieff Based Feature Extraction Techniques for Heart Disease Prediction

Venkateswarlu Tata[1], Dr.R. Bhavani[2], Dr.S.V.N. Srinivasu[3], Dr.R. Priya[4]
[1]Research Scholler, Department of Computer Science and Engineering, FEAT, Annamalai University, Chidambaram, [1]venkat543@gmail.com.
[2]Professor& Head, Department of Computer Science and Engineering, FEAT, Annamalai University, Chidambaram, [2]bhavaniaucse@gmail.com.
[3]Professor, Department of Computer Science and Engineering, Narasaraopeta Engineering College, Narasaraopeta, [3]drsvnsrinivasu@gmail.com.
[4]Professor, Department of Computer Science and Engineering, FEAT, Annamalai University, Chidambaram, [4]prykndn@yahoo.com.

Abstract—Diagnosing heart disease is really a challenging task for which several intelligent diagnostic systems were developed for enhancing the performance of diagnosing heart disease. However, in these systems, low accuracy of predicting heart disease is still a challenging task. To provide better accuracy in predicting heart risks, a novel feature selection approach is proposed which employs Real Coded Binary Artificial Bee Colony (RC-BABC) optimization algorithm with adaptive size for feature elimination. This method has the advantages of reducing algorithmic computational time, improving prediction accuracy, enhanced data quality, and saves resources in successive data collection phases. Once the features are selected, the important feature extraction phase uses ReliefF based feature extraction method to extract the features from the heart disease data set. The scores of features are computed by estimating a comparison of feature values and class values neighbor samples. The proposed Real Coded Binary Artificial Bee Colony (RC-BABC) optimization algorithm is compared with three well known methods namely an artificial neural network (ANN), K-means clustering approach and Classification and Regression Algorithm (C&RT) with measures like accuracy, precision, recall and F1-score. The proposed method achieved 96.77% of accuracy,98.8% of recall, 97.8% of precision and 98.34% of F1-score.

Keywords-- Heart disease, preprocessing, feature extraction, feature selection, prediction, optimization.

I. INTRODUCTION

In the healthcare industry, data science is considered as an important part and healthcare professionals appreciates its capacity of rapidly providing useful information and perceptions [1]. Generally, data pertaining to healthcare are electronic medical records of patients. In healthcare, data is commonly used to construct a decision supporting systems using data together with artificial intelligence and domain knowledge [2]. These information helps the professionals in detecting risks or critical errors and sends alerts accordingly [3]. Statistics, database systems, machine learning, pattern recognition, and artificial intelligence are all used in data mining [4].The process of data mining involves extracting information from large amounts of original data by extracting hidden patterns and trends [5]. The process of extracting formerly unknown, inherent, and likely functional information from data has been described as not trivial. As part of the process of detecting information from the database, it is one of the tasks. By mining data, information can be found and displayed in a manner that is easy to understand. Data collection and analysis are routinely carried out in this process [6].In an investigation, data mining is most helpful due to the large amount of non-trifling information that can be found. Humans and computers collaborate on this project. By matching the expertise of human authorities with the exploration capabilities of computers, superior results can be achieved [7]. Prediction and description are two major objectives of data mining. The model includes several variables to identify unidentified or potential values of other variables of interest in the dataset. Further, description models are used to evaluate patterns according to human-derived information. A major function of data mining is the detection of diseases. For widely disseminated raw medical data, these patterns are exploited for clinical diagnosis. A controlled method should be used to collect these data. It is possible to combine these pieces of information into a hospital information system [8]. The purpose of data mining is to identify patterns in data that can be exploited. Clinical diagnosis of heart disease may be improved by using

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

these tools [9], providing an effective way to retrieve information obscured by data.

An improved disease prediction model is presented in this paper, which includes three phases: preprocessing, optimization of feature extraction, and classification. To begin with, the given dataset is preprocessed via data transformation and rules are generated. Following that, the optimal features are selected using the newly introduced feature selection and feature extraction techniques. Lastly, these optimal features are sent to the classifier, thereby enabling a more accurate prediction.

This work aims to carry out several experiments which supports to improve the effectiveness of these predictive classifiers for diagnosing the heart disease using feature selection and optimization algorithm in terms of accuracy, diagnosis performance and enhance the computational time. The complex non-linear problems have to be dealt with flexibility and adaptability using the best optimization technique.

This paper consists of section 1 portraying the outline of heart disease, function of feature extraction and optimization in heart disease prediction, section 2 depicts the current strategies for heart disease prediction with its limitations. Section 3 gives feature extraction and feature selection mode. Section 4 gives productive experimental analysis and then section 5 ends up with concluding points and future work.

II. RELATED WORKS

In[11], To select features from the Cleveland dataset, we used a genetic algorithm. Seven subsets of features were obtained by applying four machine learning (ML) approaches namely Support Vector Machine(SVM), J48, multilayer perceptron and K-Nearest Neighbor (KNN) for predicting heart-disease. A tenfold cross-validation method was used for the evaluation of the model and results were compared with the models constructed using actual feature set and the one selected by using common feature selection approaches. In [12], machine learning technique was used to access the risk of the heart-disease. The derived dataset Clinical Practice Research Datalink in the UK was used for testing standard classification approaches namely Random forests (RF), logistic regressions (LR), neural networks (NN), gradient boosting machines and the baseline model of the American Heart Association and American College of Cardiology (ACC/AHA).75% of the data from the dataset was used for training and 25% for testing. All the above methods produced better results than ACC/AHA model. In [13], four different algorithms for classifying data namely KNN, Naive Bayes (NB), decision tree (DT) and bagging were tested on Cleveland heart-disease dataset. To obtain the most useful features, the authors used features based on domain knowledge instead of using feature selection algorithm. It was observed that the approach improved the accuracy when NB and KNN techniques while the accuracy obtained was less for DT and bagging techniques. In [14], heart-disease classification framework was created which involved feature extraction technique using Principal Component Analysis (PCA). Data dimensionality was reduced and the benefits of this was stated. Moreover, prediction accuracy was increased by the classifier with reduction in computational cost. In [15], PCA was applied to a dataset formed with people of Punjab, in India, over three generations to determine the heart-disease risk features. The features of this dataset include Circumference of the waist, body mass index, blood pressure, pulse, and weight. In [16], constructed a heart disease prediction model which involved feature selection with a Chi-squared feature evaluator integrated with RF algorithm which was tested on stat log heart-disease database. In [17], PCA in combination with feature selection was involved for extracting PCA datasets from various feature sets for which Cleveland heart-disease dataset was used. Totally, six various datasets were obtained including the original one having 14 features. Then, a feed-forward neural network (FFNN) and regression approaches were applied on every dataset mentioned above for building a new prediction model. Among six PCA datasets, only one produced 95.2% of accuracy with FFNN. In [18], an association rule has been implemented for diagnosing heart disease. Here, an improved pruning and prediction approach depending on the arithmetical measures was deployed to produce an efficient association rule for improving the effectiveness of the classifiers.

III. PROPOSED METHODOLOGY

In this work, Figure 1 shows the proposed methodology. The data collection from the data source, then it can be preprocessed by using noise removal and removing duplicate values. Once the features are selected by using Real Coded binary Artificial Bee Colony (RC-BABC) approach. Features are extracted by ReliefF method, then classify the features based on KNN classifier and finally it is evaluated by the parameters.

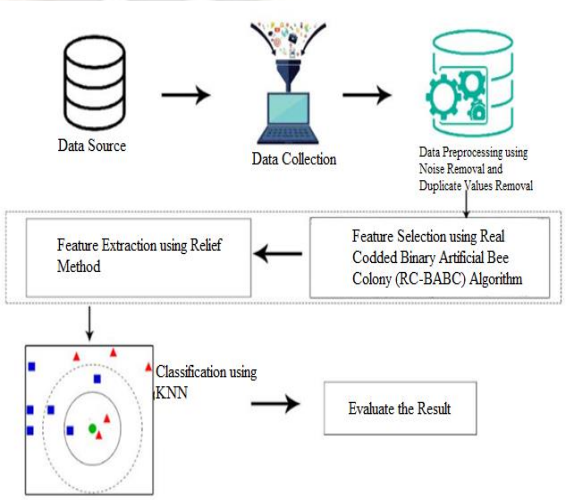

Figure 1: Proposed architecture for heart disease classification

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

A. *Dataset collection*

We are using the Heart Disease Dataset at UCI Machine Learning Repository, which has 74 independent features and a label called coronary angiography. The presence of heart disease is indicated by values 1 to 4 in the original dataset. To evaluate the developed model, historical data were supplied and the practitioners examined them physically. As a part of the protocol, three non-invasive tests were performed namely exercise electrocardiogram, coronary calcium fluoroscopy, and exercise thallium scintigraphy. The results of coronary angiogram were interpreted by the cardiologist with no knowledge of results obtained from non-invasive tests. The attributes area follows,

1. **Age:** provides the patient age.
2. **Sex:** give the patients gender information. Male is represented by 1 and female by 0.
3. **Type of Chest-pain:** Provides the chest-pain type of the patient who have experienced it Few types namely Anginal pain, atypical angina, nonanginal pain, and asymptotic angina are indicated as 1 to 4 respectively
4. **Resting Blood Pressure**: value of resting blood pressure is given with unit as mmHg
5. **Serum Cholestrol**: provides thecholesterol in serum with unit as mg/dl
6. **The fasting blood sugar is:** The value of the fasting blood sugar of the patient is compared 120mg/dl. If greater then indicated as 1 otherwise it is 0
7. **Resting ECG :** results of resting electrocardiographic is displayed. Normal, abnormal ST-T waves, and left ventricular hypertrophy types are indicated as 0, 1 and 2 respectively
8. **Max heart rate:** provides the maximum heart rate obtained from a patient.
9. **Exercise induced angina:** 1 indicates true and 0 as false
10. **Exercise inducing Depression at rest:** displays an integer or a float value.
11. **Peak exercise segment:** the values 1 to 3 represent upsloping, flat and down sloping respectively
12. **Major vessels colored by fluoroscopy**: displays an integer or a float value.
13. **Thal:** provides thalassemia value where 3, 6 and 7 indicates normal, fixed defect and reversible defect respectively
14. **Diagnosing heart disease:** Indicates if a person is affected by heart disease: 0 indicates the absence while 1 to 4 indicates the occurrence of heart disease.

B. *Preprocessing data*

Preprocessing improves the quality of data, address inconsistencies, handle missing values, reduce noise, and enhance overall reliability and effectiveness. To ensure a successful pre-processing of our data, we conducted preliminary data investigations with EDA.

1. Data preprocessing requires the identification and handling of missing values; otherwise, incorrect conclusions and inferences may be drawn from the data. On the 303 rows, there are 6 rows that contain null values, Four variables for 'ca' and two for 'thal'. A null value can be dropped or imputed. If the dataset contains enough samples, we should use the former, ensuring that there is no bias after deleting the data. Since there were relatively few null values, the mean was imputed by applying the latter. In order to replace the missing value, we use the mean of a particular feature, such as ca, that contains a missing value. Data loss can be effectively negated by adding variance to the dataset. In this way, it provides more accurate results than the first method (dropping).
2. Encoding categorical data: Categorical data is information that is categorized. Among the categorical variables included This dataset contains 'sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', and 'thal'. ML models are based on mathematical equations. In this way, We can intuitively see that categorical data will cause problems in the equation since the equations only deal with numbers. Therefore, Numerical values were converted.
3. Assigning training and testing sets to the dataset: In this step of data preprocessing, there are two parts to the dataset. First, the training dataset is used to fit or train the model. In Test dataset is the second subset, the fit model is validated. Data sets are generally divided into 70:30 or 80:20 ratios, with 70% or 80% used for training, and 30% or 20% for testing or evaluating the ML model after it has been trained or fitted. The splitting ratio will vary depending on the data set's size and shape.
4. Scaling features: In data preprocessing, scaling features is one of the key techniques used to standardize an independent variable in a dataset. As a result of feature scaling, we can compare independent variables on equal footing. Heart disease dataset includes variables 'age', 'restbp', 'chol', 'thalach', and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

'oldpeak' with different scales. A calculation of Any two values from 'restbp' and 'chol', however, will produce incorrect results if the 'chol' values dominate the 'restbp' values. Perform feature scaling to solve this problem.ML algorithms that utilize gradient descent as an optimization technique, such as logistic regression, MLP neural networks, and others, require scaling of data. The range of features affects distance algorithms, such as SVMs and KNNs. In order to determine their similarity, they use distance between data points. Scaling features is commonly done using normalization and standardization.

C. *Feature selection*

This is a process in which the feature subset is chosen from the original features where certain selection criteria Real Coded Artificial Bee Colony (RC-ABC) algorithm are followed. During this process, the best attributes is selected from the given data particularly in diagnosing medical data for which Heuristic methods are employed. Determining the highest value of the objection function is the optimization problem. Clustering method is considered as an ideal method to detect the difference among feature. The model developed here for feature selection namely Real Coded Artificial Bee Colony (RC-ABC) algorithm is summarized which adapts clustering method.

1. Clustering with optimization method

Clustering is a process of grouping multi-dimensional data based on some similarities. Here, distance measure is used estimate the similarities between the set of samples. The issue here is to place each object into any cluster *K*with given *N* objects and to minimize the Euclidean distances between the cluster center and the objects in the cluster which is given by:

$$J(w,z)=\sum_{i=1}^{n}\sum_{j=1}^{k} wij(xi - zj)^2 \quad (1)$$

where, xi (i = 1, . . . , N) is the ith sample, and zj (j = 1, . . . , N) denoting the center of jth sample is computed using:

$$Zj=\frac{1}{n}\sum_{j=1}^{N} wij.\, xi \quad (2)$$

where, n is the samples of jth cluster, and wij is the relationship of j cluster and xi sample having a value of 0 or 1. If sample i (xi) is from cluster$j$, then wij is 1, or else 0.

Therefore, by minimizing the objective function, optimization determines the center of cluster. Distance between the center of a training cluster and every sample within it (pi CL known(xj) i) is summed up where the samples of Euclidean space with n-dimension are minimized.

$$Fi=\frac{1}{dtrain}\sum_{j=1}^{dtrain} d(xj, pi\ \text{CLknown}(xj)\ i) \quad (3)$$

Here, dtrain represents the training samples. In the cost function, the number is normalized to a value ranging from 0.0 to 1.0. The value of pCL known(xj) $i$denotes the class center which belongs to the used training samples. Here, the Real Coded Artificial Bee Colony (RC-ABC) approach was the optimization techniques used for clustering.

2. Real Coded Binary ABC (RC-BABC) for optimized feature selection

- Initialize population- initially M= [X1, X2, ..., Xm] T population with m solutions or positions of food source is generated at random in the solution space with multidimension where m is the population size. Every solution Xi =Pi1,Pi2….Pij…PiD (i ranging from 1 to m and j from 1 to D, the number of parameters to be optimized) is represented by a D-dimensional vector. Here, D = N (number of attributes) for the optimization problem. In every solution vector, the element indicates the actual output of attribute units which are uniformly distributed between minimum and maximum generation bounds. With the lowest feasible features, the best accuracy is searched for feature selection which follows forward search scheme in the proposed model.
- For food sources, a feature subset is submitted to the classifier and for fitness of these food sources, accuracy is used:
- Fitness function: For every food source in respect with the bee employed, the fitness value is evaluated.

$$\text{Fitness= A[1-\%cost]+B[1-\% error]} \quad (4)$$

$$\text{Error} = \sum_{i=1}^{n} \text{pi} - \text{pl} - \text{pd} \quad (5)$$

$$\%\ \text{error} = \frac{stringerror - minerror}{maxerror - \min error} \quad (6)$$

here A and B indicates positive weight coefficients, String error is the error of the individual string which meets power balance constraint, Min error and Max error denote the minimum and maximum constraint error respectively in the population

- Constraint handling-The position when modified has to be verified for power balance constraint and generating violations in limits forunit's capacity. When power balance constraint is not meet, error is summed up with any unit chosen in random. This operation is performed until the constraint is met. In the meantime, the power outputs are checked if any violation exists in the limit for unit capacity. If any violation found, extreme limits are set. Moreover, the operating zone constraint which is prohibited is checked, when any power output of the unit lies in this prohibited zone, then set them either to upper or lower bound of that particular zone.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Xi=[1if Ri<MR (or) Xi otherwise] (7)

3. Algorithm-RC-BABC

Input:

- Control parameters
- System data
- Maximum cycle number (MCN)

Output:

- The best solution found and its corresponding objective function value
  1. The control parameters are initialized and the system data are read.
  2. The time interval t is set as 1
  3. Initial population is generated
  4. Fitness of the population is evaluated
  5. Cycle is set as 1.
  6. For every bee which is employed
  - new solution is produced and the constraints are checked
  - fitness value is computed
  - the greedy selection process is applied.
  7. Probability values Priis computed for the solutions
  8. For every bee that is unemployed or onlooker
  - a solution based on Pri is selected,
  - new solution is produced and the constraints are checked,
  - fitness value is computed
  - the greedy selection process is applied.
  9. Bees exploit abandon sources
  10. The best solution obtained are recorded.
  11. Increase the value of cycle by 1.
  12. Iterate steps 6 to 11until cycle equals MCN.
  13. Increase the value of tby 1.
  14.Iterate steps 3 to 13 until t equals 24. /
  15. Return the best solution found and its corresponding objective function value.

- The value of 24 is used as a stopping criterion for the algorithm. The algorithm stops when the value of t exceeds 24. Therefore, step 14 should be updated to: "If t is less than 24, go to step 3."

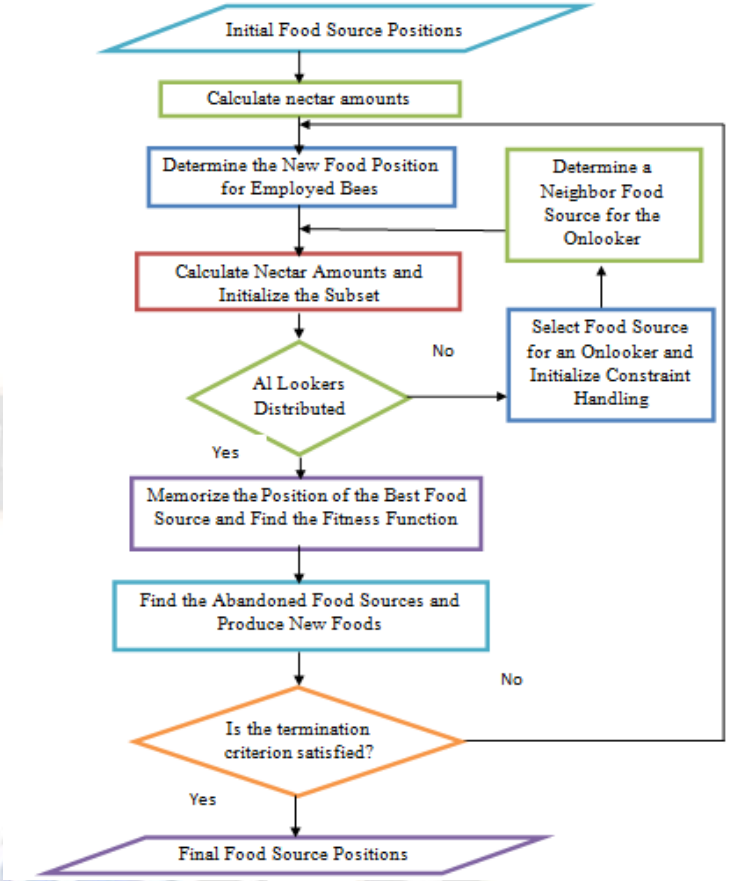

Figure-2 Flow chart for Real Coded Binary Ant Bee Colony (RC-BABC) algorithm

D. *Extraction of features*

During this process, a new small set of variables is created aiming to obtain significant information available in original variables for prediction. These variables are obtained after transformation is applied on original variables. The original variables are projected by transformed variables to a new variable space. Here, the distinct group of outcomes provide separation in a better way than the original variable space. The ReliefF method is used to extract features, and the feature scores are calculated by comparing adjacent feature and class values. As a result, useful attribute, it is anticipated that the closest distances in a class will be closer to each other than in next classes. Therefore, this attribute has a weight of calculated as follows:

In this case, W to be defined as

W-diff(xij,near_hitij)2-diff(xij,near_missij)2-diff(xij,near_missij)2/m (8)

m is the size of the sample selected randomly from the training subset diff (xij, near_hitij), diff(xij,near_missij) provides the difference between attribute values of arbitrarily chosen j distance and nearest training sample near_hitij of While diff(xij,near_missij) is the same class represents the nearest

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

training sample for another class. Values of near_missij must be closer to one another for the attribute to be useful. The closest attributes to the class labels are selected if the differences are not useful. Additionally, the filtering method reduces redundancy among all selected attributes. As described below, mutual information I(x,y) is applied to every attribute and vector of class label vector. An attribute's similarity to its vector of class labels is calculated.

$$I(x,y)=\sum_{I,J}^{n} p(x,y)\log\frac{p(xi,yj)}{p(xi)p(yj)} \quad (9)$$

p(xi) and p(yj) are A marginal feature probability function is a joint probability distribution. When two variables are completely independent, I(x,y) equals zero. Maximum redundancy is provided by

$$I(x,y)\ \max = \frac{1}{s}\sum_{k=0}^{n} I(h,i) \quad (10)$$

The minimum redundancy is given as

$$\text{Min } I(x,y) = \frac{1}{s}\sum_{k=0}^{n} I(i,j) \quad (11)$$

Where, h= {h1,h2….hk}

Feature score is calculated by determining reconstruction error once low-dimensional class-wise embedding is completed. Reconstruction error Property which varies among classes is used. For every dataset, reconstruction error is calculated as the difference between original and reconstructed data. Particularly, reconstruction errors for features are determined by summing up the reconstruction errors of every feature. At last, to obtain the final feature score, feature-wise reconstruction error of every data and the data error for every class are utilized.

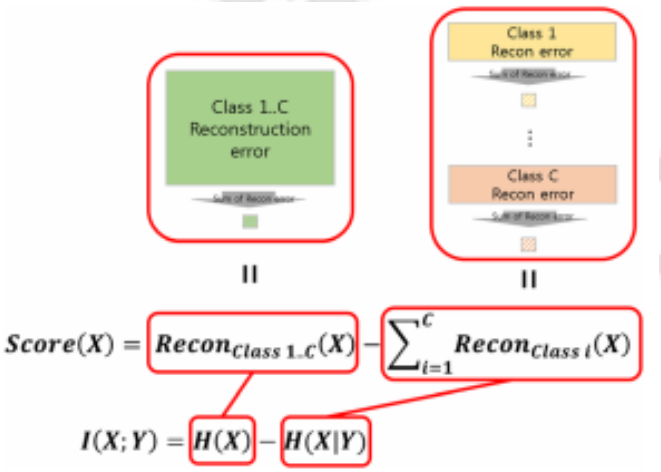

Figure-3 Determination of feature score in ReliefF method

ReconClass 1.C(X) indicates the reconstruction error once low-dimensional data X is embedded; $\sum_{i=0}^{c}$ ReconClass 1. (X) represents the sum of the all the applicable terms of every class separately. The first part is computed which used no label information and is related to H(X). The next one is a reconstruction error with label information which relates to the conditional entropy. The following is obtained when this is related to the score equation as given in figure 3,

$$\text{Score (x)=Recon class 1.C(X)-} \sum_{1}^{c} \text{Reconi(x)}$$

$$= \text{Recon 1.C(Fj)-}\sum_{1}^{c}\sum_{1}^{F} \text{Reconi(Fi)}$$

$$=\text{Recon1.C(Fj)-}\sum_{1}^{F}\sum_{1}^{C} \text{Reconi(Fi)} \quad (12)$$

Total score of X is calculated by summing up the scores obtained for every feature. Thus this is termed to be the feature score which differs from that of the label.

E. *Classification using KNN*

Symmetrical uncertainty was utilized as a proper measure for ranking the attributes. With these ranks, the attributes with least ranking was selected which are provided as input for KNN algorithm for classifying heart disease. Symmetrical uncertainty compensates information gain

$$SU(X, Y) = 2[IG(X/Y)/H(X)\ H(Y)] \quad (13)$$

Similarity among the features is determined as

$$(X,Y) = \sqrt{\sum_{i=1}^{n} f(xi, yi)} \quad (14)$$

here n represents the total number of attributes. This model contains 2 phases. In the first phase, predominant features are selected while in the second datasets are classified and the accuracy is measured. Steps involved in classification are as follows:

Step 1: Every predominant feature is identified and other features are removed. (These are ranked and the one with the least rank is selected from the dataset.)

Step 2: KNN is applied on the dataset.

Step 3: The accuracy of the classifier is determined.

Step 4: Using KNN, the dataset is classified as Acc (Accuracy without SU)

Step 5: Using KNN, the dataset is classified as Acc SU (accuracy with SU)

Step 6: When (Acc) falls below (Acc (SU), the feature with lowest SU is removed

Step 7: Iterate steps 4 to 6 until Acc (SU) < (Acc)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

F. *Performance analysis*

The experimental result is carried out in MATLAB software and the parameters used for analysis are accuracy, f1-score, recall and precision. These values obtained for these parameters are compared against three standard methods namely Artificial Neural Network (ANN), K-means clustering algorithm and Classification And Regression Algorithm (C&RT)

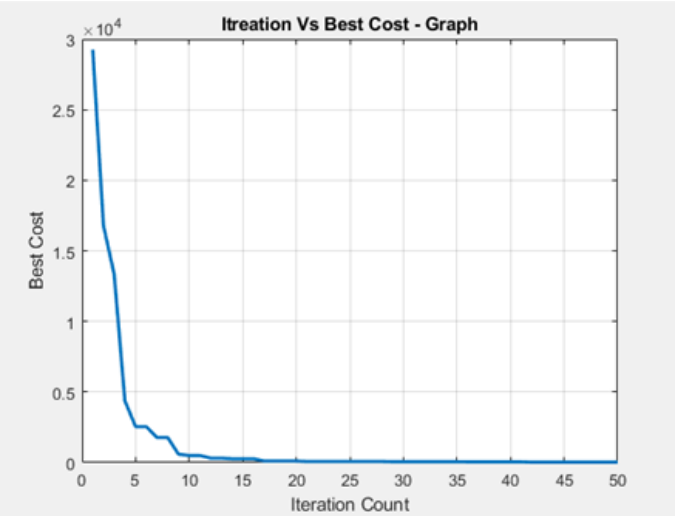

Figure-4 Analysis of iteration

Figure-4 shows the analysis of iteration during optimization. X-axis indicates the iteration count and Y axis shows the best cost. According to this graph, when the iteration count is 10, the best cost is achieved.

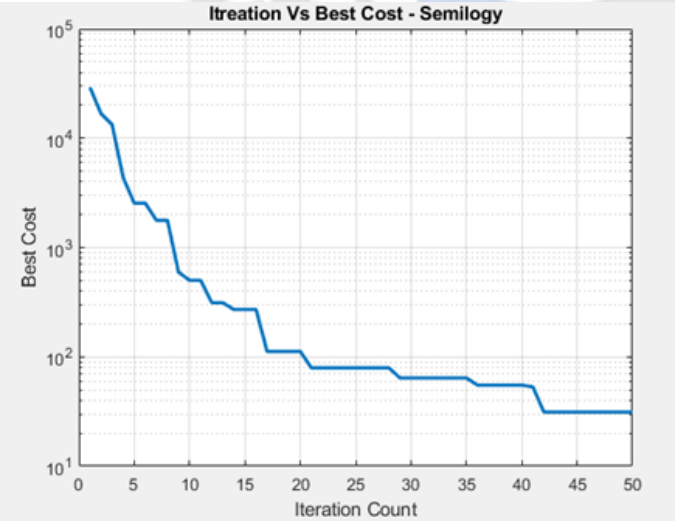

Figure-5 Analysis of iteration in semi log mode

Figure-5 shows the analysis of iteration in semilog mode during optimization. X-axis indicates the iteration count and Y axis shows the best cost. According to this graph, when the iteration count is 50, the best cost is achieved.

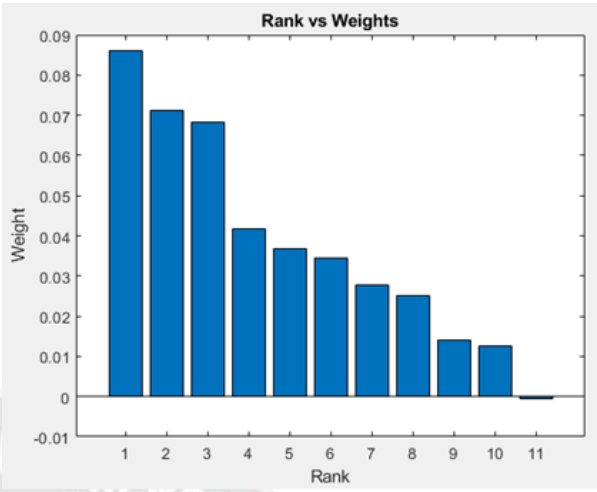

Figure-6 Analysis of rank and weight

Figure-6 shows the analysis of rank and weight during ReliefF method, where x-axis indicates the rank and y axis shows the weight. According to this graph, when the rank is increased,the weight is also increased which shows better feature extraction

Accuracy facilitates prediction for the proposed model. A true positive (TP) and a true negative (TN) represent the absence and presence of an attack as predicted by the classifier. A false positive (FP) and false negative (FN) are false predictions made by the proposed model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (15)$$

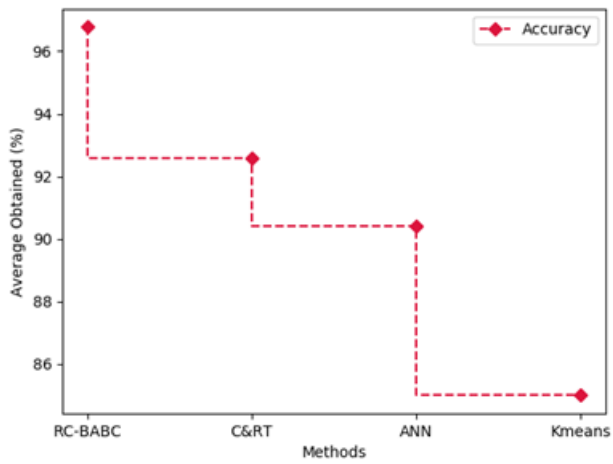

Figure 7: Comparison of accuracy

Figure 7 compares the accuracy of existing C&RT, ANN, K-Means and proposed RC-BABC. The X axis represents the various methods whereas the Y axis represents the obtained values in percentage. According to the existing method, 92.6%, 90.4%, and 85% are achieved, while the proposed method achieves 4% more than C&RT, 6% more than ANN, and 11% more than K means.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

F1-Score is used to determine prediction performance. Precision and recall are weighted averages. An F1-score of 1 indicates the best case scenario, while a score of 0 indicates the worst case scenario. The F1-Score is calculated as follows:

$$F1\ Score = \frac{2*P*R}{P+R} \qquad (16)$$

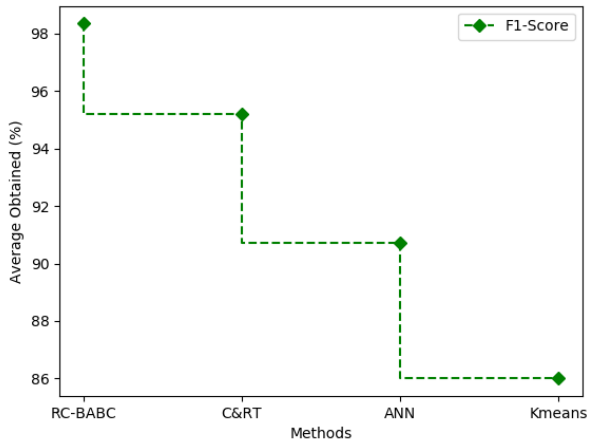

Figure 8: Comparison of F1 score

Figure 8 compares F1 score between existing C&RT, ANN, K-Means and proposed RC-BABC whereas X axis represents the various methods while Y axis shows the obtained values in percentage. Compared with existing methods, the proposed method achieves 95.2%, 90.7%, and 86%, while C&RT achieves 3.2% better, ANN achieves 8.3% better, and K means achieve 12% better.

The precision of the classification model indicates how well it predicts attacks. This is the measure which gives predicted positive result of the classifier if disease is present and is estimated as:

$$Precision\ (P) = \frac{TP}{TP+FP} \qquad (17)$$

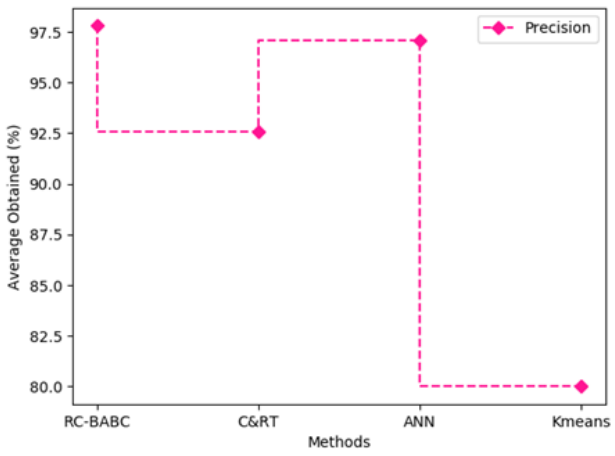

Figure 9: Precision comparison

As shown in figure 9, we compare the precision of existing C&RT, ANN, K-Means, and the proposed RC-BABC for which the X axis shows the various methods and the Y axis shows the obtained values in percentages. The proposed method achieves 5.2% better than C&RT, 0.7% better than ANN and 17.8% better than K means when compared to the existing method, which achieves 92.6%, 97.1% and 80% respectively.

Recall refers to the detection ability to correctly reject disease parts in dataset. Mathematically, this can also be given as follows,

$$\text{Recall (R)} = \frac{TN}{TN+FP} \qquad (18)$$

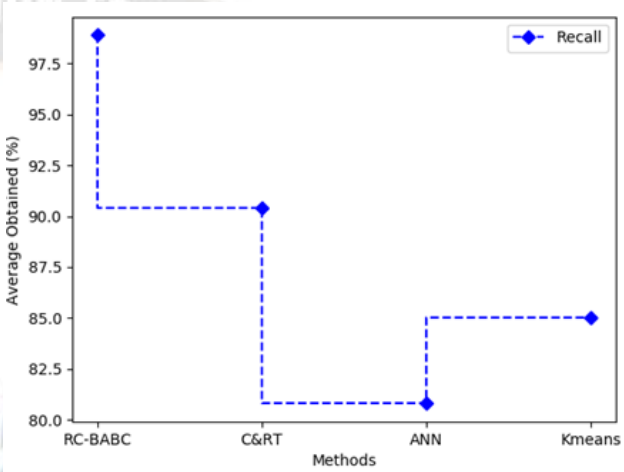

Figure 10: Comparison of recall

In Figure 10, recall is compared between existing C&RT, ANN, K-Means, and the proposed RC-BABC, where the X axis represents the various methods and the Y axis shows the obtained values in percentages. Based on comparison, existing methods achieve 90.4%, 80.8%, and 85% while the proposed method achieves 8.4% over C&RT, 18.2% over ANN, and 13% over K means.

Table-1 shows the overall comparison of existing C&RT, ANN, K-Means and proposed RC-BABC approaches for the performance measures considered for evaluating the proposed model.

Table 1- A comparison of existing and proposed methods

Method	Accuracy	Recall	Precision	F1-score
RC-BABC	96.77	98.8	97.8	98.34
C&RT	92.6	90.4	92.6	95.2
ANN	90.4	80.8	97.1	90.7
K Means	85	85	80	86

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

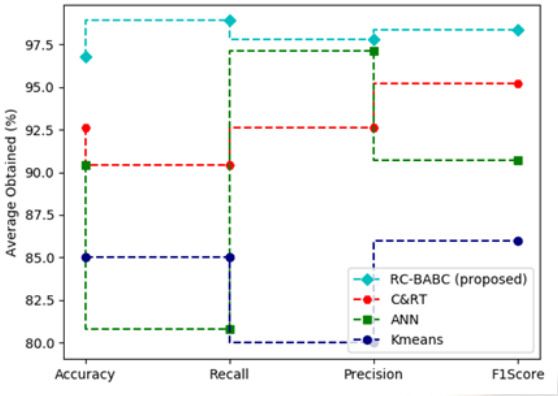

Figure 11 Overall comparative analysis

Figure 11 illustrates the overall comparison of existing C&RT, ANN, K-Means and proposed RC-BABC, where X axis shows the parameters to be analysed and Y axis shows the average values calculated. In comparison, the proposed method achieves 96.77% accuracy, 98.8% recall, 97.8% precision, and 98.34% F1score.

IV. CONCLUSION

This work has put in an effort in designing a two stage diagnosing system which improves the accuracy of predictingheart risks and failure. Two systems developedwere feature extraction and feature selection. Both these systems used the same classification method, and achieved a classification accuracy of 96.77% with the proposed RC-BABC based feature selection system and 98.8% of recall, 97.8% of precision and 98.34% of F1-score.Moreover, this proposed diagnosing system performed better than other methods used for comparison namely C&RT, ANN and K-Meansclustering. Thus, this proposed system provided a good assistance for the physicians in making accurate decision duringdiagnosis. The future work concentrates on including machine learning method in feature selection for better prediction.

REFERENCE

[1] Cresswell K, Majeed A, Bates DW, et al. Computerised decision support systems for healthcare professionals: An interpretative review. J Innov Health Inform 2013; 20(2): 115–128.

[2] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In: *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 31 March 2008, pp. 108--115.

[3] Thabtah F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics Health Social Care 2019; 44(3): 278–297.

[4] EvanthiaE.Tripoliti, TheofilosG.Papadopoulos, Georgia S.Karanasiou, Katerina K.Naka and DimitriosI.Fotiadis, "Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques," Computational and Structural Biotechnology Journal, vol. 15, pp. 26-47, 2017.

[5] Wuyang Dai, Theodora S.Brisimi, William G.Adams, Theofanie Mela, Venkatesh Saligrama and IoannisCh.Paschalidis, "Prediction of hospitalization due to heart diseases by supervised learning methods," International Journal of Medical Informatics, vol. 84, no.3, pp. 189-197, 2015.

[6] JaberAlwidian, Bassam H.Hammo and Nadim Obeid, "WCBA: Weighted classification based on association rules algorithm for breast cancer disease," Applied Soft Computing, vol. 62, pp. 536-549, 2018.

[7] MahinVazifehdan, Mohammad Hossein Moattar and Mehrdad Jalali, "A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction," Journal of King Saud University - Computer and Information Sciences, 2018

[8] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE Access, vol. 5, pp. 8869-8879, 2017.

[9] Ilayaraja M and Meyyappan T, "Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Itemsets," 4th International Conference on Eco-friendly Computing and Communication Systems, ICECCS, vol. 70, pp. 586-592, 2015.

[10] Jabbar MA, "Prediction of heart disease using k-nearest neighbor and particle swarm optimization", Biomedical Research, vol. 28, no. 9, pp. 4154-4158, 2017.

[11] Gokulnath CB, Shantharajah SP. An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing 2018.

[12] Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Plos One 2017; 12(4). [PMC free article] [PubMed] [Google Scholar]

[13] Khateeb N and, Usman M. Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique. In: Proc Int Conf Big Data Internet Things - BDIOT2017 2017.

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

[15] Badaruddoza, Kumar R, Kaur M. Principal component analysis of cardiovascular risk traits in three generations cohort among Indian Punjabi population. J Advanced Res 2015; 6(5): 739–746. [PMC free article] [PubMed] [Google Scholar]

[16] Jabbar MA, Deekshatulu BL, Chandra P. Prediction of heart disease using random forest and feature subset selection. Advances Intelligent Syst Comp Innovations Bio-Inspired Comp App 2015; 187–196.

[17] Santhanam T, Ephzibah EP. Heart Disease Classification Using PCA and Feed Forward Neural Networks. Mining Intell Knowledge Exploration Lecture Notes Comp Sci 2013; 90–99.

[18] J. Zhang et al., "Coupling a Fast Fourier Transformation With a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment," IEEE Access, vol. 5, pp. 10674-10685, 2017.

[19] Anand, D., Tata, V., Samriya, J.K., Kumar, M. (2023). A Review on Deep Learning-Enabled Healthcare Prediction Technique: An Emerging Digital Governance Approach. Soft Computing: Theories and Applications. Lecture Notes in Networks and Systems, vol 627. Springer, Singapore. https://doi.org/10.1007/978-981-19-9858-4_22