Hybridized Firefly And Differential Evolution Optimization Algorithm Based Feature Selection For Disease Prediction

VR. Nagarajan, Dr. D. Vimal Kumar

ABSTRACT: Early detection and characterization of the disease are considered to be critical factors in the management and control of diseases. Firefly Optimization Algorithm with Homogeneous Fuzzy Rule-Based Classification System (FOA-Homo-FRBCS) and FOA with Heterogeneous FRBCS (FOA-Hetro-FRBCS) were proposed for early prediction of lung cancer, leukemia and heart diseases in a distributed environment. In FOA-Homo-FRBCS and FOA-Hetro-FRBCS method, the clinical dataset was converted into fuzzy sets. The converted data was split into a number of partitions and each partition was processed in mappers. The most significant features of clinical data were selected by using Firefly Optimization Algorithm (FOA). In FOA-Homo-FRBCS, the selected features were explored by FRBCS where only Random Forest (RF) was used to generate rules for prediction of diseases. In FOA-Hetro-FRBCS, the selected features were explored by FRBCS where RF, Bayesian Tree, and NeuroTree were used to generate the rules for prediction of diseases. Sometimes, FOA based feature selection has a slow convergence problem. So in this paper, Hybridized Firefly and Differential evolution Optimization Algorithm (HFDOA) is proposed for feature selection. It integrates the Firefly Algorithm (FA) with Differential Evolution (DE) algorithm by combining the attraction mechanisms of FA with the mixing ability of DE to increase the speed of convergence and the diversity of the population. Moreover, Auto tuned hybridized Firefly and Differential search evolution Optimization Algorithm (AFDOA) is proposed to automatically tune the randomness parameters of HFDOA. AFDOA makes the search space becomes narrow for feature selection. It effectively reduces the time consumption for feature selection and also increases the prediction accuracy. The selected features by AFDOA are used in Hetro-FRBCS to predict the diseases. This process is named as AFDOA-Hetro-FRBCS.

Keywords: Firefly Optimization, Homogeneous Fuzzy Rule-Based Classification System, Heterogeneous Fuzzy Rule-Based Classification System, Hybridized Firefly and Differential evolution Optimization Algorithm, Auto Tuned Hybridized Firefly and Differential search evolution Optimization Algorithm.

1. INTRODUCTION

Early detection of disease can improve the survival rate of people by properly treating the people. The application of data mining [1] with big data brings a new dimension to predict various diseases like lung cancer, leukemia and heart disease. Data mining with big data can handle a huge volume of clinical data. Data mining with big data is used to identify and extract useful information from huge volume of a clinical dataset. Nowadays, researchers explored various ways to implement data mining in clinical data to achieve an accurate prediction of diseases. The clinical dataset consists of unwanted or irrelevant features which can reduce the performance of data mining techniques. So, a proper feature selection [2] is required to achieve high disease prediction accuracy. A modified differential evolution (MDE) [3] was introduced for optimal feature selection. The MDE provides faster convergence speed but this faster convergence yields premature convergence. It can be handled by employing a large population but it consumes more time to estimate the fitness function. So, Firefly Optimization Algorithm (FOA) [4] was introduced for feature selection and it was implemented in a distributed environment by using the MapReduce framework. It reduced the time consumption for feature selection. The selected features were used in Naïve Bayes, C4.5 and Random forest to predict the lung cancer, leukemia and heart disease.

- *VR. Nagarajan, Dr. D. Vimal Kumar*
- *Research Scholar, Nehru Arts and Science College, and Assistant Professor, PG & Research Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India, vrnag74@gmail.com*
- *Associate Professor, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India, vimal1519@yahoo.co.in*

This work was extended by proposing Firefly Optimization Algorithm with Homogeneous Fuzzy Rule-Based Classification System (FOA-Homo-FRBCS) and FOA with Heterogeneous FRBCS (FOA-Hetro-FRBCS) [5] for efficient disease prediction. In this paper, feature selection in FOA-Hetro-FRBCS is improved by proposing a Hybridized Firefly and Differential evolution Optimization Algorithm (HFDOA) method. It increases the convergence speed and enhances the prediction accuracy by combing the firefly and differential evolution algorithm. The improper selection of $\alpha, r1, r2$ and $r3$ randomness parameters in HFDOA can increase the search space for feature selection. To control the randomness in HFDOA, Auto tuned hybridized Firefly and Differential search evolution Optimization Algorithm (AFDOA) is proposed. It automatically tunes the randomness parameters and selects the features using HFDOA. The auto-tuned randomization parameters make the search space becomes narrow for feature selection. Thus the time consumption for feature selection is reduced by AFDOA. The selected features are given as input to Hetro-FRBCS for rule generation to predict the lung cancer, leukemia and heart disease with high prediction accuracy.

2. LITERATURE SURVEY

A modified differential evolution (MDE) algorithm [3] was introduced as an optimal feature selection technique for the prediction of heart disease. In MDE, DE/rand/2-wt/exp strategy was used. DE processed the conventional differential evolution process and then rand randomly selected vector to be perturbed. The 2-wt strategy selected four-vectors and weighted differences are included to form the mutant vector and the exp used exponential crossover which was performed on particular variables in one loop until it was within the crossover rate. The selected features were used in fuzzy analytical hierarchy process with feed-

forward neural network to predict the heart disease. The error minimization would be carried out in the prediction of heart disease. A feature selection strategy called Genetic Algorithm based on Random Forest (GARF) [6] was proposed for oesophageal cancer prediction. Initially, a Spearman's correlation analysis was performed to group the correlated features and then selected the most discriminative features by using GARF. GARF was developed based on GA associated with a new multi-parametric fitness function taking into account an RF misclassification rate, Area Under Curve (AUC) measurement and sparsity constraint. However, RF requires more space to store the data. A hybridized feature selection and extraction approach [7] was proposed to enhance the cancer prediction. For cancer prediction, the hybridized approach used a DNA methylation degree in probes and promoters regions. The hybridized approach utilized the F-score filter feature selection method which reduced the dimensionality of DNA methylation data. Then, features set such as mean methylation density, Fast Fourier Transform (FFT), the number of the methylation density peaks and the difference between the maximum and minimum methylation density peak were created by feature extraction model. Finally, these features were preceded by Support Vector Machine (SVM), Random Forest (RF) and Naïve Bayes (NB) to predict the cancer types. In some cases, the prediction accuracy of this approach still needs improvement. A new Multiswarm Heterogeneous Binary Particle Swarm Optimization (MHBPSO) algorithm [8] was proposed for improved feature selection in liver and kidney disease diagnosis. MHBPSO consisted of BPSO, Boolean PSO (BoPSO), Self Adjusted Hierarchical BoPSO (SAHBoPSO) and Catfish SAHBoPSO (CSAHBoPSO). MHBPSO performed a heterogeneous search on the entire solution space using BPSO, BoPSO, SAHBoPSO and CSAHBoPSO algorithms. Based on the fitness value, the leader was selected and the remaining particles followed the leaders to select the optimal features for liver and kidney disease prediction. However, the computational complexity of MHBPSO algorithm is high. Genetic Algorithm (GA) based feature selection with Adaptive Neuro-Fuzzy Inference System (ANFIS) [9] was proposed for breast cancer prediction. It was a combination of GA and ANFIS. The GA was used to select the most relevant features from a pool of features while the ANFIS method was used to classify the data as existence and non-existence of breast cancer. The most relevant features were used in ANFIS. It was a rule-based system to predict breast cancer. However, ANFIS has strong computational complexity restrictions. A frequent feature selection method [10] was proposed for feature selection of efficient heart disease prediction. The prediction performance was enhanced by using relevant non-linear integral and fuzzy measure in the prediction model. The non-addivity of the fuzzy measure reflected the importance of the feature attributes as well as their interactions. The medical profiles such as blood sugar, age, sex, and blood pressure were used to predict the likelihood of patients getting heart disease. However, the accuracy of this method is low. An expert system [11] was introduced based on stacked Support Vector Machine (SVM) for efficient diagnosis of heart disease. Two SVMs models were used in the expert system. In the first SVM model, the irrelevant features in the dataset were eliminated through shrink their coefficients to zero. The second SVM model was used as a predictive model to predict heart disease based on the selected features. Hybrid Grid Search Algorithm (HGSA) was proposed to simultaneously optimize the two SVM models which enhanced the heart disease prediction accuracy. However, the proper selection of kernel function in SVM is more challenging one.

3. PROPOSED METHODOLOGY

In this section, Hybridized Firefly and Differential evolution Optimization Algorithm (HFDOA) method and Auto tuned hybridized Firefly and Differential search evolution Optimization Algorithm (AFDOA) method for prediction of lung cancer, leukemia, and heart diseases are described in detail. Initially, lung cancer, leukemia and heart disease dataset are collected and convert the data into fuzzy membership values. The fuzzified data are split into a number of segments. Each segment corresponds to one map task. Each map task selects the optimal features using HFDOA and combines each mapper result at reducer. Then, the selected features are processed in Hetro-FRBCS to predict the diseases. This process is named as HFDOA-Hetro-FRBCS. The randomness parameters of HFDOA are auto tuned and optimal features are selected by AFDOA. The selected features are given as input to Hetro-FRBCS for disease prediction. This process is named as AFDOA-Hetro-FRBCS.

3.1 Hybridized Firefly and Differential evolution Optimization Algorithm for feature selection

HFDOA is introduced for efficient feature selection which combines the advantage of FA and DE algorithm. The conventional FA based feature selection can subdivide the features into subgroups automatically in terms of attractiveness mechanism through the variation of light intensity. One of the FA deviations can escape from the local minima due to long-distance mobility by Levy flight. From this, it is known that FA is good at exploration and diversification. The feature selection by using FA is briefly explained in [4]. The DE [12] based feature selection uses a vectorized mutation operator. In DE, the features of lung cancer, leukemia, and heart disease dataset are encoded in the form of strings called chromosomes. A collection of strings is called a population represented as P. It is a collection of NP number of d-dimensional feature vectors $\mathrm{x_{i,G}} = [\mathrm{x_{1,i,G}}, \mathrm{x_{2,i,G}}, \ldots \mathrm{x_{D,i,G}}], \mathrm{i} = 1,2, \ldots \mathrm{NP}$ for each generation G and D denotes the number of features. The objective function is given as follows:

$$\mathrm{F(P')} = \delta . \frac{|\mathrm{P'}|}{|\mathrm{B}|} + (1 - \delta) . \mathrm{A} \qquad (1)$$

where $\mathrm{F()}$ is the objective function, $\mathrm{P'}$ denotes the subset of selected features, B denotes the boundary factor per features and A denotes the classification accuracy.

The initial population is chosen randomly which represents different features in the search space and should cover the entire feature space. DE generates new feature vectors by adding the weighted differences between two population vectors to a third vector. This operation is called a mutation process. For each target vector $\mathrm{x_{i,G}};\ \mathrm{i} = 1,2,3, \ldots \mathrm{NP}$, a mutant vector is generated based on

$$\mathrm{v_{i,G+1}} = \mathrm{x_{r1,G}} + \mathrm{CP}(\mathrm{x_{r2,G}} - \mathrm{x_{r3,G}}) \qquad (2)$$

where, $x_{r1,G}$, $x_{r2,G}$ and $x_{r3,G}$ are randomly selected three vectors such that the indices $r1$, $r2$ and $r3$ are distinct and CP denotes the control parameter, it controls the amplification of the differential variation $(x_{r2,G} - x_{r3,G})$.

The mutated vector's features are then mixed with the features of another predetermined vector, the target vector, to yield the so-called trial vector. Feature mixing is often denoted as crossover. To this end, a trial vector is formed which is given as follows:

$$u_{i,G+1} = (u_{1i,G+1}, u_{2i,G+1}, \dots u_{Di,G+1}) \quad (3)$$

where,

$$u_{ji,G+1} = \begin{cases} v_{j,i,G+1}, & \text{if}(\text{rand(j)} \le C) \text{ or } j = \text{random(i)} \\ x_{j,i,G} & \text{if } ((\text{rand(j)} > C)) \text{ and } j \ne \text{random(i)} \end{cases} \quad (4)$$

In equation (4), $\text{rand(j)} \in [0,1]$ is the jth evaluation of an uniform random number generator, $C \in [0,1]$ is the crossover rate, random(i) is the randomly selected index x belongs to $\{1,2, \dots D\}$ which ensures that $u_{j,G+1}$ gets at least one feature from $v_{i,G+1}$. If the trial vector yields a lower objective function than the target vector, the trial vector replaces the target vector in the following generation. This operation is named as selection.

$$x_{i,G+1} = \begin{cases} u_{i,G+1}, & \text{if } f(u_{i,G+1}) > f(x_{i,G}) \\ x_{i,G}, & \text{otherwise} \end{cases} \quad (5)$$

The process of mutation, crossover and selection continues for a fixed number of generations. The efficiency of mutation operator and crossover operator in DE can provide good mixing ability among the population. Hence it provides a better diversity in the population. Simultaneously, DE can also carry out a local search during the process, particularly when approaching the local optimal solutions, and thus this advantage is used to improve both the exploitation and exploration ability of HFDOA. Moreover, updating the current global best in the whole population ensures that solutions can converge to the optimum, while diversification through mixing and regrouping the whole population allows the search algorithm to escape from local optima and may simultaneously increase the diversity of solutions. Instead of generating the new positions from random walks or other operators, HFDOA method only mixes and regroups the individual local information obtained after the main iteration of parallel FA and DE processes. The main advantage of such mixing and regrouping is to guarantee the search focusing on the current locations in the promising areas obtained in the previous phase instead of searching or re-searching less promising regions of the search space. Thus, the proposed HFDOA increases the speed of convergence and the diversity of the population.

3.2 Auto tuned hybridized Firefly and Differential search evolution Optimization Algorithm

The parameters $r1, r2, r3$ and α are control the randomness of HFDOA. These parameters help to decide the search space for the feature selection process. So, these parameters have to be tuned to get better search space for feature selection. These parameters are auto tuned by proposed AFDOA method. A randomness reduction technique is frequently used as iteration is often used as iterations continue, and this is often achieved by using an annealing-like-exponential function

$$\alpha \leftarrow \alpha\eta \quad (6)$$

In equation (6), η is a cooling parameter ranges from 0 to 1. It introduces a cooling schedule to the HFDOA. One simple way to automatically tune α is to set α as proportional to the standard deviation of the current solutions. But, for multi-modal problems, this standard deviation should be calculated for each local mode among local subgroups of population. If two modes S and T with current best solutions x_s^* and x_t^*. There are two standard deviations σ_S and σ_T should be calculated among the solutions relative to x_a^* and x_b^* respectively. Then, the overall α should be a function of σ_S and σ_T. The σ_S and σ_T is combined by calculating weighted average of σ_S and σ_T. It is given as follows,

$$\sigma = \frac{\sigma_S n_1 + \sigma_T n_2}{n_1 + n_2}, n_1 + n_2 = n \quad (7)$$

where, n_1 is the population of S and n_2 is the population of T. As iteration continue, σ decreases in general. So set the parameter α as,

$$\text{autotune\_}\alpha = \zeta\sigma, 0 < \zeta < 1 \quad (8)$$

where, $\text{autotune\_}\alpha$ denotes the auto tuned parameter α, $\zeta = \sqrt{d/(2d+1)}$, d denotes the dimensionality of the feature. Hence, α is automatically associated with the scale of feature selection problem of interest.

In DE, $x_{r1,G}$, $x_{r2,G}$ and $x_{r3,G}$ are randomly selected as current best solution. There are three standard deviations $\sigma_{x_{r1,G}}$, $\sigma_{x_{r2,G}}$ and $\sigma_{x_{r3,G}}$should be calculated among the solutions relative to $x_{r1,G}$, $x_{r2,G}$ and $x_{r3,G}$ respectively. Then, the overall $r1, r2$ and $r3$ should be a function of σ_{r1}, σ_{r2} and σ_{r3}. It is given as follows,

$$\sigma = \frac{\sigma_{r1} n_1 + \sigma_{r2} n_2 + \sigma_{r3} n_3}{n_1 + n_2 + n_3}, n_1 + n_2 = n \quad (9)$$

Set the parameters $r1$, $r2$ and $r3$ as,

$$\text{autotune\_}r1 = \xi\sigma_{r1} \quad (10)$$

$$autotune\_r2 = \xi\sigma_{r2} \quad (11)$$

$$autotune\_r3 = \xi\sigma_{r3} \quad (12)$$

where, $\xi = \sqrt{d/(3d+1)}$, d is the dimensionality of features. The search space becomes narrow by auto-tuning $\alpha, r1, r2$ and $r3$ parameters of HFDOA. The auto-tuned parameters are used in the HFDOA to select the most significant features in lung cancer, leukaemia, and heart diseases. This process is named as Auto tuned HFDOA (AFDOA). Finally, the selected features are used in the Hetro-FRBCS to predict the diseases effectively. The overall process of AFDOA-Hetro-FRBCS is given as follows.

Auto tuned hybridized Firefly and Differential search evolution Optimization Algorithm with Heterogeneous- Fuzzy Rule-Based Classification System

Input: clinical dataset DS, max_itr, $autotune\_\alpha$, β, γ, $autotune\_r1$, $autotune\_r2$, $autotune\_r3$, CP, C, $x_i = \{x_1, x_2, \ldots . x_n\}$
Output: Prediction of disease
Begin
Convert DS into fuzzy sets
Split fuzzified DS into a number of partitions and each partition is processed by mappers.
for each mapper do
Divide the population in each mapper into two groups G_1 and G_2.
Each particles in group G_1 and G_2 randomly selects the features.
Evaluate the objective function of each particle in $G_1$and G_2using equation (1).
Repeat
Do in parallel
for group G_1do
while $(t = 1{:}\, max\_itr)$
for $i = 1{:}\, h$ do
for $j = 1{:}\, h$ do
if $\left(F(G_{1j}) > F(G_{1i})\right)$
Calculate attraction between fireflies
Compute the distance between the fireflies i and j
Move the firefly i towards j using
$x_i = x_i + \beta_0 * exp\left(-\gamma r_{ij}^2\right) * \left(x_j - x_i\right) + autotune\_\alpha * (rand - \frac{1}{2})$
end if
Evaluate new solution and update the objective function
end for j
end for i
Rank the fireflies and find the current best
end while
end for group G_1
for group G_2
while $(t < max\_itr)$
for $i = 1{:}\, m$ in all individuals do
for each x_i, randomly select $x_{autotune\_r1,G}$, $x_{autotune\_r2,G}$ and $x_{autotune\_r3,G}$
Generate a new vector $v_{i,G+1}$ using mutation operation as
$v_{i,G+1} = x_{autotune\_r1,G} + CP\left(x_{autotune\_r2,G} - x_{autotune\_r3,G}\right)$
Create a random index $random(i)$
Create a randomly distributed number $rand(j) \in [0,1]$
for $j = 1{:}\, D$
for each parameter v_{ij}, do crossover operation and update
$u_{ji,G+1} = \begin{cases} v_{j,i,G+1}, & if (rand(j) \le C)\ or\ j = random(i) \\ x_{j,i,G} & if\ \left((rand(j) > C)\right)\ and\ j \neq random(i) \end{cases}$
end for v_{ij}
end for j
Choose operation, select and update the solution x_i
end for i
end while
end Do in parallel
Update the global best in the whole population
Jumble the two groups and regroup then randomly into new groups G_1 and G_2.
Evaluate the fitness of each particle
Until a termination condition is met
end for mappers
Collect the selected features from each mapper and merge them in reducer
Process the selected features in Hetro-FRBCS to predict the diseases
End

In the above algorithm, max_itr denotes the maximum iteration, h denotes the number of fireflies, $x_{i(j)}$ is the position of the firefly $i(j)$, β is the attraction between fireflies, r is the distance, γ is the light absorption coefficient, $autotune_\alpha$ is the auto-tuned α, $autotune\_r1$ is the auto tuned $r1$, $autotune\_r2$ is the auto tuned $r2$, $autotune\_r3$ is the auto tuned $r3$, $rand$ is the random number and D is the dimension of features. The above AFDOA-Hetro-FRBCS algorithm selects the most important features and which are used in Hetro-FRBCS to predict the diseases.

4. RESULTS AND DISCUSSION

The effectiveness of proposed HFDOA-Hetro-FRBCS and AFDOA-Hetro-FRBCS are evaluated in terms of accuracy, precision, recall, and f-measure. For the experimental purpose, lung cancer, leukemia, and heart disease dataset are used. The lung cancer dataset contains 181 tissue samples (31 MPM and 150 ADCA). The training set contains 32 of them, 16 MPM and 16 ADCA. The rest 149 samples are used for testing. Each sample is described by 12533 genes. The leukemia dataset was taken from a collection of leukemia patient samples reported by Golub. The dataset consisted of 72 samples: 25 samples of AML, and 47 samples of ALL. Each sample is measured over 7,129 genes. The heart disease dataset is Cleveland database which consists of 76 attributes.

4.1 Accuracy

Accuracy is defined as the number of all correct disease predictions made divided by the total number of disease prediction made. This is defined as a ratio of appropriately classified data to overall classified data.

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + False\ Positive\ (FP) + TN + False\ Negative\ (FN)}$$

where, if the class label is the presence of disease and the prediction outcome is the presence of disease then it is TP.

If the class label is the absence of disease and the prediction outcome is the absence of disease, then it is called TN.

If the class label is the absence of disease and the prediction outcome is the presence of disease, then it is called FP.

If the class label is the presence of disease and the prediction outcome is the absence of disease, then it is called FN.

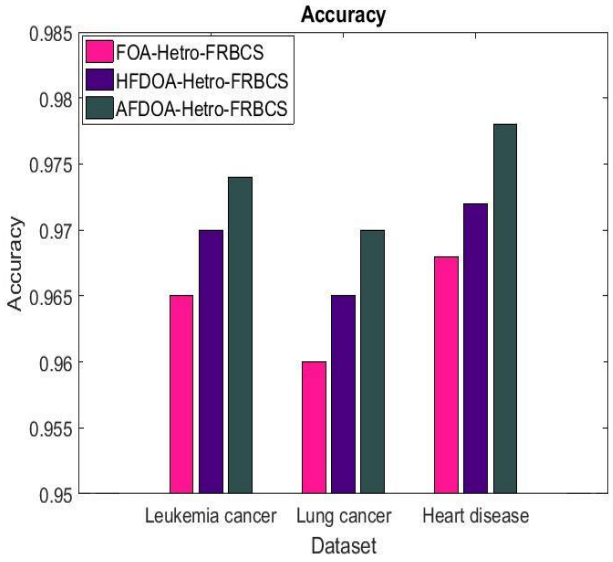

Figure 1. Comparison of Accuracy

Figure 1 shows the comparison between FOA-Hetro-FRBCS, HFDOA-Hetro-FRBCS and AFODA-Hetro-FRBCS in terms of accuracy for leukemia cancer, lung cancer and heart disease datasets. The accuracy of AFODA-Hetro-FRBCS is 0.93% greater than FOA-Hetro-FRBCS and 0.52% greater than HFDOA-Hetro-FRBCS method in leukemia dataset. From figure 1, it came to know that the proposed AFODA-Hetro-FRBCS has high accuracy than HFDOA-Hetro-FRBCS and FOA-Hetro-FRBCS methods for leukemia cancer, lung cancer, and heart disease datasets.

4.2 Precision

Precision is how many selected features are relevant. It is the ability to correctly predict those that do not have the disease. It can be calculated as,

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

Figure 2 shows the comparison between FOA-Hetro-FRBCS, HFDOA-Hetro-FRBCS and AFODA-Hetro-FRBCS in terms of precision for leukemia cancer, lung cancer and heart disease datasets. The precision of AFODA-Hetro-FRBCS is 1.45% greater than FOA-Hetro-FRBCS and 0.82% greater than HFDOA-Hetro-FRBCS method in lung cancer dataset. From figure 2, it came to know that the proposed AFODA-Hetro-FRBCS has high precision than HFDOA-Hetro-FRBCS and FOA-Hetro-FRBCS methods for leukemia cancer, lung cancer, and heart disease datasets.

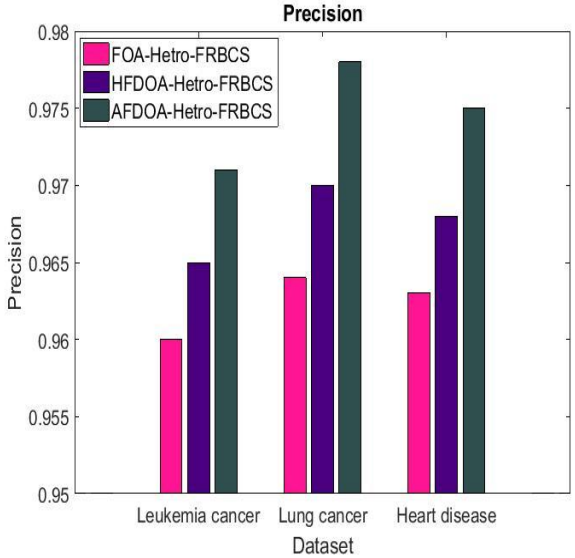

Figure 2. Comparison of Precision

4.3 Recall

Recall is how many relevant items are selected. It is the ability of a test to correctly predict those with the disease. It can be calculated as,

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

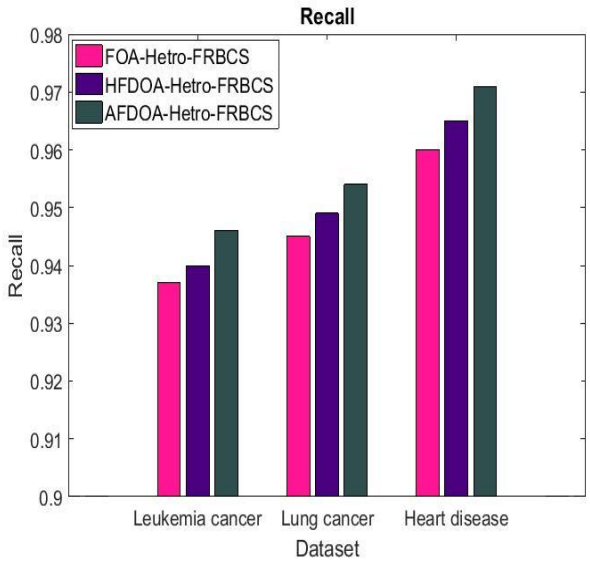

Figure 3. Comparison of Recall

Figure 3 shows the comparison between FOA-Hetro-FRBCS, HFDOA-Hetro-FRBCS and AFODA-Hetro-FRBCS in terms of recall for leukemia cancer, lung cancer and heart disease datasets. The recall of AFODA-Hetro-FRBCS is 1.15% greater than FOA-Hetro-FRBCS and 0.62% greater than HFDOA-Hetro-FRBCS method in heart disease dataset. From figure 3, it came to know that the proposed AFODA-Hetro-FRBCS has high recall than HFDOA-Hetro-FRBCS and FOA-Hetro-FRBCS methods for leukemia cancer, lung cancer and heart disease datasets.

4.4 F-measure

F-measure is an external measure for measuring the accuracy of disease prediction methods. F-measure depend on two factors are precision and recall. It is calculated as,

$$\mathrm{F-Measure} = 2.\frac{\mathrm{Precision}\,.\,\mathrm{Recall}}{\mathrm{Precision}\,+\,\mathrm{Recall}}$$

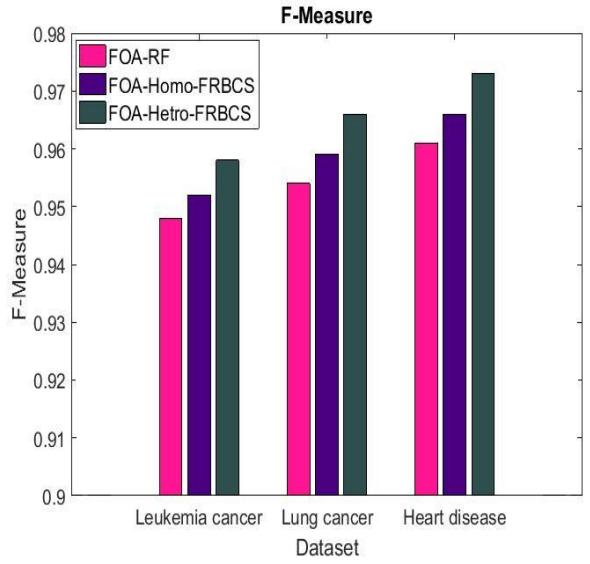

Figure 4. Comparison of F-measure

Figure 4 shows the comparison between FOA-Hetro-FRBCS, HFDOA-Hetro-FRBCS and AFODA-Hetro-FRBCS in terms of f-measure for leukemia cancer, lung cancer and heart disease datasets. The f-measure of AFODA-Hetro-FRBCS is 1.25% greater than FOA-Hetro-FRBCS and 0.74% greater than HFDOA-Hetro-FRBCS method in heart disease dataset. From figure 4, it came to know that the proposed AFODA-Hetro-FRBCS has high recall than HFDOA-Hetro-FRBCS and FOA-Hetro-FRBCS methods for leukemia cancer, lung cancer, and heart disease datasets.

5. CONCLUSION

In this paper, HFDOA and AFDOA are proposed for efficient prediction of lung cancer, leukemia and heart diseases. The HFDOA is the combination of FA and DE algorithm that is used to select the most important features. It increases the speed of convergence and provides better diversity in the population by mixing the global best solution of FA and DE. The AFDOA tunes the randomness parameters of HFDOA and selects the most significant features in the dataset using HFDOA. The selected features fed into Hetro-FRBCS to predict the diseases. The experimental results prove that the HFDOA-Hetro-FRBCS and AFDOA-Hetro-FRBCS have better accuracy, precision, recall, and f-measure than FOA-Hetro-FRBCS method for lung cancer, leukemia and heart disease dataset.

REFERENCES

[1]. Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering, 106, 212-223.

[2]. Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

[3]. Vivekanandan, T., & Iyengar, N. C. S. N. (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. Computers in biology and medicine, 90, 125-136.

[4]. Nagarajan, V. R., & Kumar, V. (2018). An optimized sub group partition based healthcare data mining in big data. International Journal for Innovative Research in Science & Tehnology (IJRST), 4(10), 79-85.

[5]. Nagarajan, V. R., & Kumar, V. (2019). An ensemble big data classification for healthcare data predication analysis. European Journal of Business & Social Sciences, 7(6), 160-172.

[6]. Paul, D., Su, R., Romain, M., Sébastien, V., Pierre, V., & Isabelle, G. (2017). Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Computerized Medical Imaging and Graphics, 60, 42-49.

[7]. Raweh, A. A., Nassef, M., & Badr, A. (2018). A hybridized feature selection and extraction approach for enhancing cancer prediction based on DNA methylation. IEEE Access, 6, 15212-15223.

[8]. Gunasundari, S., Janakiraman, S., & Meenambal, S. (2018). Multiswarm heterogeneous binary PSO using win-win approach for improved feature selection in liver and kidney disease diagnosis. Computerized Medical Imaging and Graphics, 70, 135-154.

[9]. Turabieh, H. (2016). GA-based feature selection with ANFIS approach to breast cancer recurrence. International Journal of Computer Science Issues (IJCSI), 13(1), 36.

[10]. Saravanakumar, S., & Rinesh, S. (2014). Effective Heart Disease Prediction using Frequent Feature Selection Method. International Journal of Innovative Research in Computer and Communication Engineering, 2, 2767-2774.

[11]. Ali, L., Niamat, A., Golilarz, N. A., Ali, A., & Xingzhong, X. (2019). An Expert System Based on Optimized Stacked Support Vector Machines for Effective Diagnosis of Heart Disease. IEEE Access.

[12]. Khushaba, R. N., Al-Ani, A., & Al-Jumaily, A. (2011). Feature subset selection using differential evolution and a statistical repair mechanism. Expert Systems with Applications, 38(9), 11515-11526.