Hybrid Search Optimization Techniques for the Classification of Cancer Data from the Optimal Set of Features

L. Meenachi[1], J. Ramprasath[2], A. S. Muthanantha Murugavel[3]
1,2,3 Department of Information Technology, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India.
E-mail[1] lmeenachi@gmail.com, E-mail[2] jrprasath@gmail.com, E-mail[3] murugavel.asm@gmail.com

Abstract-The paper seeks to find cancer-related data, with the help of the dataset's reduced features. The dataset is searched for globally optimal features using particle swarm optimization. It typically produces higher classification results, but on occasion, these algorithms choose neighborhood features incorrectly, creating a local optimal feature selection trap. In order to select the best features from the microarray gene expression data, this work suggests a method that combines the two algorithms with fuzzy rough set to form Particle Swarm Optimization and Tabu Search with Fuzzy Rough Set for Optimal feature selection (PSTFRO) technique. It simultaneously selects neighborhood features and global features. The usefulness of the suggested selection of features method is evaluated in relation to receiver operating characteristics, exact classification, computation speed, and a positive predicted value using the fuzzy rough nearest neighbor classifier. The proposed approach produces better results than the presently used global optimal, local optimal, and optimal feature selection algorithms.

Keywords-Global Features, Neighborhood Features, Hybridization, Optimal selection of features.

1. Introduction

In the field of biomedical research, data classification with reduced features is crucial [1]. Either globally or locally, the dataset's subsets of features are searched for feature selection. While both search techniques look for the most relevant features, they might not consider all of the pertinent features when looking for subsets [2]. When both search methods are used together, they complement one another and look for optimal features. Selecting the best selection of features for classification, the subsets of features that were found are assessed [3].

There have been numerous studies on how to remove features from a dataset. This paper's goal is to categorize cancer data with suitable features by proposing the algorithms to select the best features. Tabu search, a local optimal search method, has been combined with particle swarm optimization(PSO) to determine the best subsets of characteristics [4,5,6]. These hybridized algorithms evaluated the best subsets to produce the best subset of features using fuzzy rough sets once again. The fuzzy rough set evaluates the data while taking into account the ambiguous and uncertain data [7].

The best feature selection approach, PSTFRO, selects the pertinent features for classification. Using a range of performance indicators, the effectiveness of the suggested approach is compared with that of the already available algorithms. A fuzzy rough nearest neighbor classifier is used to assess the performance of the proposed approach. The following are the few feature selection studies that have already been done and are discussed in this paper:

Nibaran Das et al. [8] proposed the Convolutional neural networks (CNN) and an evolutionary metaheuristic algorithm for deep feature segmentation form the proposed technology. Mengnan Tian et al. [9] suggested developing two distinct local-based mutation operators and probability of an individual-depending selection.

The modified differential evolution(DE) method for feature selection for categorizing cardiovascular illness has been put forth by T. Vivekanandan et al. [10]. The fuzzy AHP and a feed-forward neural network classifier both use selected features. It shows enhanced results in terms of forecast time and accuracy. An efficient solution-based tabu search (SBTS) algorithm was created by Xiangjing Lai et al. [4]. To increase computational efficiency, SBTS has been defined as the combined usage of confined swap neighborhood's tabu status and hashing algorithms. The swarm intelligence algorithm and spiral mechanism was combined and suggested by Ke Chena et al. [11] for selecting the ideal feature subset. The strategy improves position quality of subsequent generation by widening the search and adding new factors to the initial position update.

Four proposed search methods for feature selection were made by L. Meenachi and S.Ramakrishnan[12]. Search techniques for features from random populations that are continuous, global, and local optimum subsets of features include Random Local Optimal (RLO), Random Global and Optimal (RGLO), and Global and Local Optimal (GLO). Three classification phases were proposed by Aytug Onan et al. [13]: while selection of feature uses the consistency-based feature selection technique, selection of instances uses a evaluator technique and a classification uses fuzzy-rough nearest neighbor(FRNN) method,

Joao P. Papa et al.[14] developed the Binary Brain Storm Optimization for feature selection. Various transfer functions are used to transform real-valued solutions into Boolean hypercubes. By employing unsupervised feature selection methods that replicate the dataset's within-cluster level of importance for the feature by using cluster-dependent feature-weighting methods. Deepak Panday et al.'s [15] proposed feature selection method removes the relatively low weight features from the dataset.

Feature evaluation and improve swarm intelligence technique were combined by Indu Jain et al. [16] to create a hybrid model for the classification of cancer. This chooses fewer features to categorise biological samples of tumours. In order to improve feature selection. Bin Yu et al. [17] designed the advantage-disadvantage neighbourhood degree based on a neighbourhood operator. The data are then used to build and optimise a multi-attribute, group-focused executive method based on a rough set on dominance.

In a multitask setting, J. Liang et al. [18] intended to address the issue of multifaceted feature selection alongside numerous auxiliary one-objective feature selection tasks, which examine more promising regions and hasten convergence by utilising the single-objective tasks' expertise in problem-solving. By modifying the Random Subset Feature Selection (RSFS) approach, Lakshmipadmaja D et al. [19] presented the dataset's feature selection. The reduced features are classified using the k-nearest-neighbor (kNN) classifier.

Fuzzy sets with interval values are utilised to compute the memberships in fuzzy-kNN in the evolutionary fuzzy k-nearest neighbours classifier (EF-kNN-IVFS) proposed by Joaquin Derrac et al. [20]. The parameter-independent feature weighted fuzzy kNN technique was created by Nimagna Biswas et al. [21] and optimises the value of k and feature loads using a self-adaptive kind of differential evolution. Dewan Md. Farid et al. [22] to handle biological data multi-class classification with noisy examples, overfitting instances, and class-imbalance data, a novel adaptive rule-based classifier was designed. L. Meenachi and S.Ramakrishnan[23] proposed two-hybrid algorithms for cancer data classification for selecting the best features from microarray gene expression data

The majority of research studies primarily use fewer dimensional data for classification. The global optimal search algorithms have local optimal traps that prevent them from searching neighborhood features, they place less emphasis on search capabilities and space, they fail to hybridize adaptive hybrid neighborhood features, and they only take into account a very small dataset for feature selection and classification. In order to address even the vague features, global and local search algorithms are combined with fuzzy rough sets in this paper in order to overcome the limitations of previous research. Based on it, an optimal selection of features algorithm is proposed and contributions of this paper include the following:

- A fuzzy rough set is hybridized with the swarm intelligence method to choose the feature that is globally optimal.
- To select the local optimal feature, the neighborhood search algorithm is hybridized with a fuzzy rough set.
- To obtain optimum features, a union operation is performed between the global optimal feature and the local optimal feature.

The PSTFRO algorithm that has been proposed to select the top features for classification. The study has the following structure: The dataset is discussed in the paper's Section 2. Next in Section 3,the paper outlines the suggested approach. In part 4, the findings are presented, and in part 5, the conclusion and recommendations for additional research are give

2. Description of Dataset

The efficiency of the suggested algorithm is evaluated in this paper using microarray gene expression data. They are leukemia[7], small-round-blue-cell tumour (SRBCT), breast, and diffuse large B-cell lymphoma (DLBCL)[12].Table 1[24] provides information on the dataset.

Table 1. Dataset Description

S.No.	Datasets
1	Leukaemia is an abnormal cell that develops on the bone marrow and causes an uncontrollable rise in the body's white blood cells. It has 57 instances and 12583 features.
2	One of such malignant tissues is known as a small round blue-cell tumour (SRBCT). These tissues can be seen under a microscope. There are 63 instances and 2309 features.
3	Breast cancer, often known as breast cancer, is the formation of malignant cells on the lobules or ducts of the breast cells. It has 54 instances and 9217 features.
4	The powerful malignant tumour known as diffuse large B cell lymphoma (DLBCL) can spread throughout the entire body. The level of severity of a patient's sickness is used to determine how quickly the abnormal cell mass is expanding. and include

3. Proposed Technique

Using the PSTFRO algorithm for optimal best features for classification were chosen. The effectiveness of the suggested method is assessed using a FRNN classifier. The suggested method exhibits enhanced outcomes in terms of several metrics when its performances are compared to those of existing algorithms.

3.1. PSTFRO Algorithm

This method combines the local optimal search algorithm tabu search and the global optimal search algorithm PSO with fuzzy rough set algorithm. The algorithm is explained, and Figure 1 depicts the workings of the PSTFRO algorithm.

The PSTFRO algorithm operates on the following principle: it is a metaheuristic population-based swarm intelligence system that looks for its global best features. Initialize the swarm, determine the swarm's fuzzy rough dependence degree, and, if the particle fitness test yields an optimal result, update the swarm's personal and global best features accordingly. For the following generation, the particle locations and velocity are updated. The objective of this process is to result in the finest features. It is repeated till it reaches its stopping condition.

If the cost of the candidate feature list is less than the cost of the previous best features, the tabu list is updated with the difference between the candidate list and best features. The features are deleted list if the tabu list size is too large. The greatest ideal features for classification are those produced by combining features obtained locally and globally.

PSTFRO Algorithm

Input: Microarray gene expression data
Output: Optimal Features

Begin
//Global optimal features
Set the particles, V_{e_i} , P_{o_i}
g_{best} =0, p_{best} =0;
// Check the particle fitness
Compute the evaluation function for the swarm with Fuzzy Rough dependence Degree
If the evaluated value is greater than the previous evaluated value
set current value as g_{best} and p_{best}
//Update particle velocity and position
Particle velocity and position for each particle are calculated
$$V_{e_i}(x+1) = V_{e_i}(x) + (L_1 * rand() * p_{best} - P_{o_i}(x)) + (L_2 * rand() * (g_{best} - P_{o_i}(x))$$
Update the position of the particle
$$P_{o_i}(x+1) = P_{o_i}(x) + V_{e_i}(x+1)$$
$$p_{best} = P_{o_i}(x+1)$$
Best position obtained so far is assigned as g_{best}
Endif
//Stopping Condition
If (previous g_{best} = = current g_{best})
GF $= g_{best}$
Else
Calculate the dependency of particles in swarm
Endif
//Local optimal features
Specify T;
Declare Tabulist=0, Tbest-P=0;
//Generation of candidate features
Tbest = Randomly generate a starting population of responses;
Repeat
Tbest-P = Tbest;

CFeature_L=0;

For (CFeature□□Tbest-near)
  If(CFeature not in Tabulist)
    CFeature_L = CandFeatures;
  Endif
Endfor
CFeature = find the most suitable features from CFeature_L;
Determine the degree of dependency (CFeature, Tbest-P)
//Check for the fitness
    If (dependency degree (CFeature) > Tbest-P) then
      Tbest = CFeature;
    Endif
    Tabulist = CFeature;
    If (sizeof(Tabulist > T)
      Delete the first component of Tabulist
    Endif
 Until (Tbest-P = = Tbest)
 LF= Tbest;
 // Optimal Features
 OF= GF □ LF;
 Return OF
 End

Fuzzy rough nearest neighbor classifier performance on various performance measures is used to assess how well the top features selected by the proposed PSTFRO technique perform. By analyzing the performance of several classifiers, the fuzzy rough nearest neighbor classifier, a probability-based classification technique, is selected for performance evaluation.

4. Results and Discussion

The most optimal features are chosen from the four datasets, namely Leukemia, SRBCT, Breast, and DLBCL, by the proposed PSTFRO approach employed in this paper.

4.1. Feature Selection Comparison: In Table 2, results are displayed with relation to the quantity of features chosen using the suggested PSTFRO method algorithm is compared to the quantity of attributes chosen using the existing feature selection techniques, including Tabu Search (TS)[4], DE [10], PSO[11], RLO, and GLO[12].The results of the comparison show that the suggested algorithm for selecting the best features selects the less desirable features.

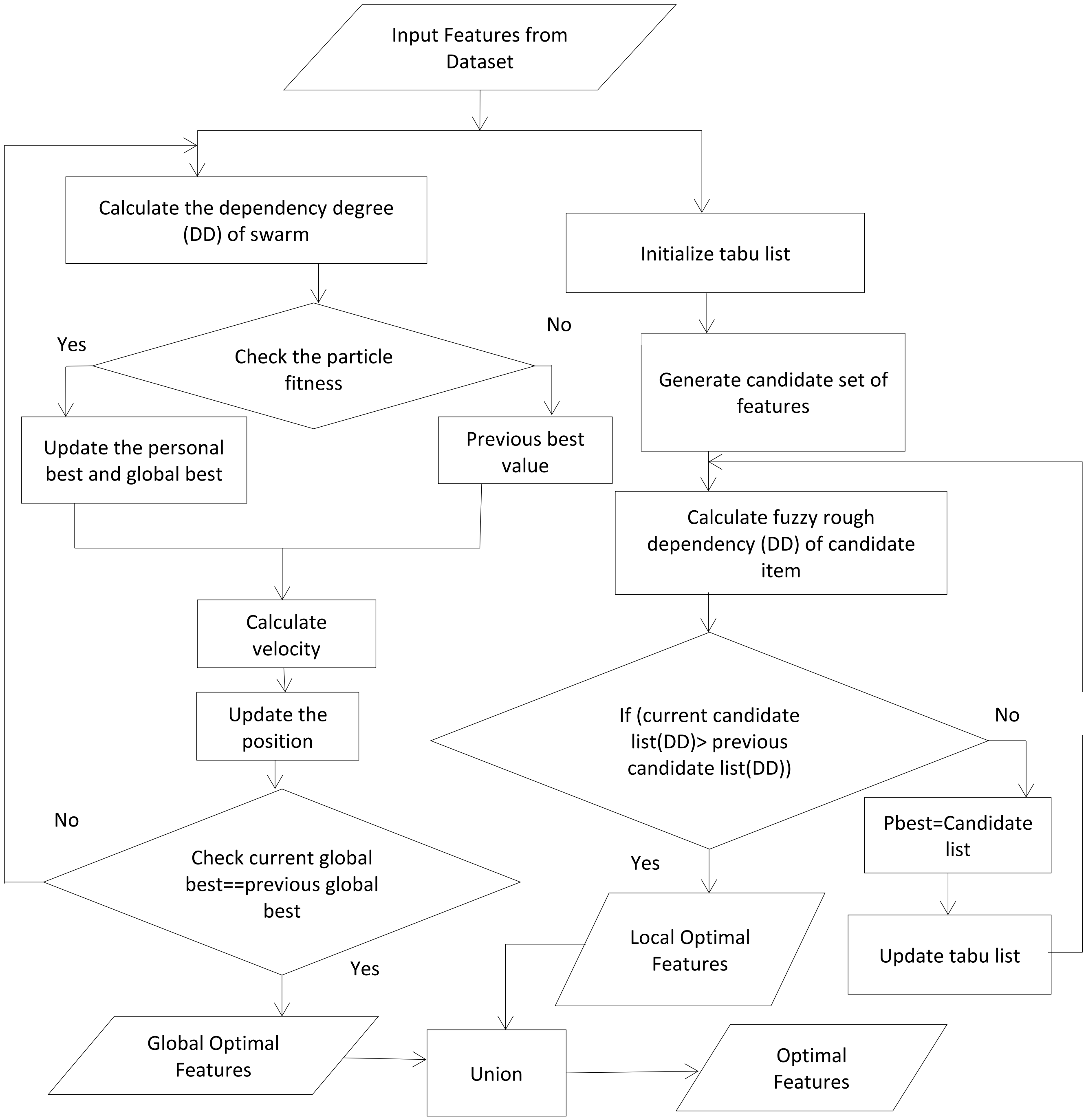

Figure1. Background of PSTFRO Algorithm

Table 2. Number of features selected comparison

Dataset's name	TS[4]	DE[10]	PSO[11]	RLO[12]	GLO[12]	PSTFRO
Leukemia	6976	6923	6834	3984	1582	413
SRBCT	1253	1156	1289	534	167	118
Breast	5023	4489	4345	2128	1436	1421
DLBCL	2797	2890	2673	1634	561	473

4.2. Classifier for Performance Evaluation: To determine whether the suggested selection of features algorithm selects only the necessary features, its performance is assessed and compared to that of competing algorithms. To evaluate how well the suggested and already available selection of features algorithms performs, a classifier is chosen from the following classifiers: Rule-based classifier [22], Nearest Neighbor(NN) classifier [19], Fuzzy Nearest Neighbour (FNN) classifier [20], and FRNN [13]. The classification accuracy of several classifiers for their features that have been condensed by the proposed PSTFRO technique is shown in Figure 3.

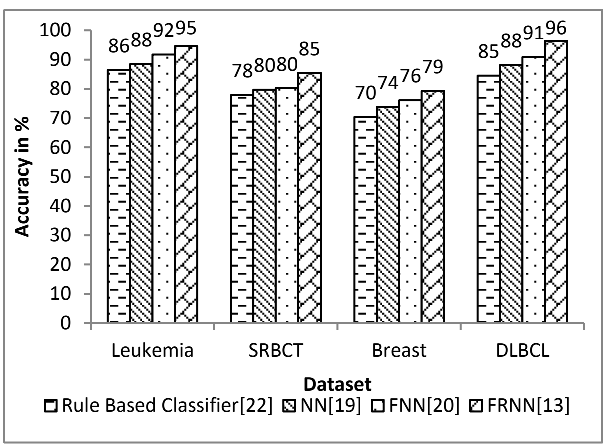

Figure 3 PSTFRO Algorithm's Accuracy from Classifiers

FRNN had shown the enhanced classifier accuracy among the different classifiers. Therefore, FRNN classifier is utilized to compare feature selection methods' performance. The fuzzy rough uncertainty in nearest neighbor is exploited by FRNN, an extension of nearest neighbor classifier [13]. Experimental evidence demonstrates that the suggested algorithms pick out fewer features than the other algorithms existing in use.

4.3. Performance Analysis of Proposed Approach: The effectiveness of selection of feature method is evaluated by means of the FRNN classifier. When the instances are divided into 10-fold groups, it is performed to validate the prediction. Each segment will serve as training and test data throughout the course of 10 iterations. The ratios of correctly predicted cases to the entire population are measured as the performance metrics in this paper, which focus on classifier accuracy. Necessary features are chosen by the proposed algorithm for selecting features that contribute to the prediction of cancer data, therefore the final value should be high when compared to that of the existing methods shows that appropriate features have been selected. Figure 4 illustrates the enhanced classifier accuracy of the proposed feature selection approach.

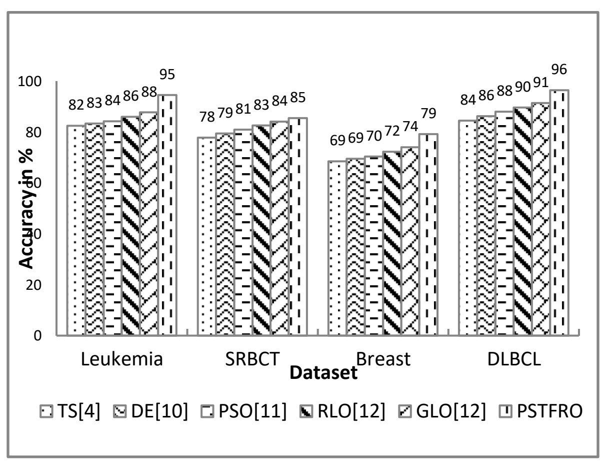

Figure 4. Classifier Accuracy Comparison

The computation time for the feature selection procedure is calculated similarly to classification accuracy. Performance is greater if computation involves less time. The computation times for the proposed and existing techniques are displayed in Figure 5. The proposed approach has shorter computing times, which suggests a shorter computation time for prediction.

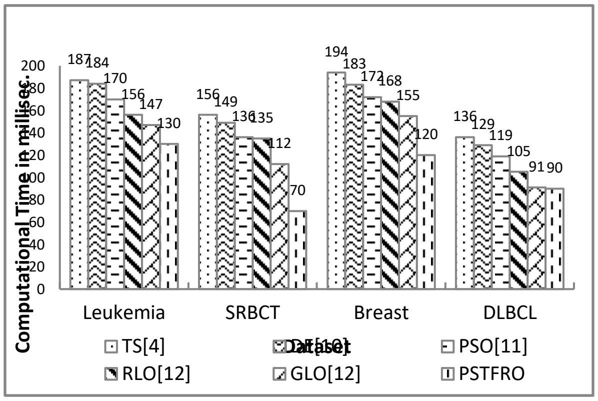

Figure 5. Feature selection algorithms' computation times, both proposed and existing

The classifier prediction rate is measured with graphical analysis using the receiver operating characteristic (ROC) curve. It is shown on the y-axis against Sensitivity and the x-axis against 1-Specificity [25]. If the curve is above the diagonal and at the top corner of the ROC space, the forecast is accurate. The output of suggested PSTFRO is shown in Figure 6 and appears as mentioned, proving the predictor's accuracy.

As a result, the suggested PSTFRO algorithm outperform the prevailing methods in relation to accuracy of the classifier, processing time, ROC, and PPV.

5. Conclusion

The proposed feature selection technique PSTFRO combines particle swarm optimization, fuzzy rough set and tabu search to produce the optimal subset of features. The proposed algorithms address the flaws of the earlier algorithms. The FRNN classifier is chosen from a pool of classifiers based on performance to prove the effectiveness of the suggested method. By way of the number of features selected, accuracy, receiver operating

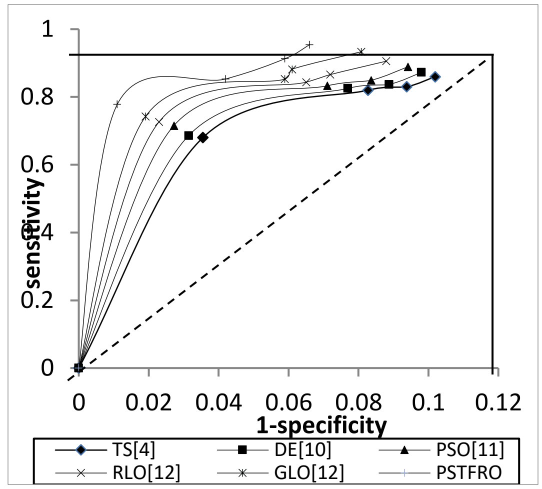

Figure 6. Receiver Operating Characteristics Comparison

The ratio of genuine positives to conditions expected is known as the positive predictive value (PPV) [26]. Otherwise, it suggests that additional positive discoveries are probably false positives, hence the result's value should be high. The PPV findings are displayed in Table 3. The suggested algorithm selecting features performs more effectively than the current method, showing that the anticipated true results are the actual positive results. Projected false positive instances are infrequent.

Table 3. Positive Predictive Values Compare

Dataset's name	TS [4]	DE [10]	PSO [11]	RLO [12]	GLO [12]	PSTFRO
Leukemia	0.84	0.84	0.85	0.86	0.87	0.94
SRBCT	0.77	0.79	0.80	0.82	0.84	0.85
Breast	0.70	0.71	0.73	0.74	0.75	0.80
DLBCL	0.86	0.88	0.89	0.91	0.94	0.98

characteristics, and positive anticipated value, the suggested selection of features algorithm performs better than the prevailing methods. In the future, fuzzy rough set approach can be combined with the comparable global search and neighborhood search algorithms

References

[1]. Alam MZ, Rahman MS, Rahman MS (2019) A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked, doi: https:// doi.org/10.1016/j.imu.2019.100180.

[2]. Yunus Emre Yildiz, Ali Osman Topal(2019) Large scale continuous global optimization based on micro differential evolution with local directional search. Information Sciences, 477 : 533–544.

[3]. Chengjun Zhan, Shuanghong Li, Yupu Yang (2019) Improved process monitoring based on global–local manifold analysis and statistical local approach for industrial process. Journal of Process Control 75:107–119.

[4]. Xiangjing Lai , Dong Yue , Jin-Kao Hao, Fred Glover (2018) Solution-based tabu search for the maximum min-sum dispersion problem. Information Sciences 441:79–94

[5]. Radka Polakova, Josef Tvrdik, Petr Bujok (2019) Differential evolution with adaptive mechanism of population size according to current population diversity. Swarm and Evolutionary Computation, https://doi.org/10.1016/j.swevo.2019.03.014.

[6]. Laith Mohammad Abualigaha,Ahamad Tajudin Khader, Essam Said Hanandeh(2018) A new feature selection method to improve the document clustering using particle swarm optimization algorithm. Journal of Computational Science, 25: 456–466.

[7]. Loganathan Meenachi, Srinivasan Ramakrishnan(2018) Evolutionary sequential genetic search technique-based cancer classification using fuzzy rough nearest neighbour classifier.Healthcare Technology Letters, 5: 130–135.

[8]. Nibaran Das, Bodhisatwa Mandal, KC Santosh, Linlin Shen, Sukanta Chakraborty(2023), Cervical cancerous cell classification: opposition-based harmony search for deep feature selection. Int. J. Mach. Learn. Cyber. https://doi.org/10.1007/s13042-023-01872-z

[9]. Mengnan Tian, Xingbao Gao(2019) Differential evolution with neighborhood-based adaptive evolution mechanism for numerical optimization. Information Sciences, 478: 422–448.

[10]. T. Vivekanandan , N. Ch Sriman Narayana Iyengar (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.

[11]. Ke Chena, Feng-Yu Zhoua, Xian-Feng Yuanb(2019) Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. Expert Systems With Applications 128:140–156.

[12]. L.Meenachi and S.Ramakrishnan (2018) Random Global and Local Optimal Search Algorithm Based Subset Generation for Diagnosis of Cancer. Current Medical Imaging Reviews 14:1. (doi:https://doi.org/10.2174/ 1573405614666180720 152838).

[13]. Aytug Onan (2015) A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Systems with Applications 42:6844–6852.

[14]. Joao P. Papa, Gustavo H. Rosa , Andre N. de Souza , Luis C.S. Afonso(2018) Feature selection through binary brain storm optimization. Computers and Electrical Engineering, 72: 468–48.

[15]. Deepak Panday, Renato Cordeiro de Amorim, Peter Lane (2018) Feature weighting as a tool for unsupervised feature selection. Information Processing Letters, 129: 44–52.

[16]. Indu Jain, Vinod Kumar Jain, Renu Jain (2018) Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. Applied Soft Computing, 62:203–215.

[17]. Bin Yu, Ruihui Xu, Zeshui Xu, Jianhua Dai (2023),Optimizing group-oriented multiple attribute decision-making method based on dominance-based rough set model with

advantaged (disadvantaged) neighborhoods, Engineering Applications of Artificial Intelligence, 124,106602

[18]. J. Liang, Y. Zhang, B. Qu, K. Chen, K. Yu and C. Yue(2023), A Multiform Optimization Framework for Multi-objective Feature Selection in Classification, IEEE Transactions on Evolutionary Computation, doi: 10.1109/TEVC.2023.3284867.

[19]. Lakshmipadmaja D, B. Vishnuvardhan (2018) Classification Performance Improvement Using Random Subset Feature Selection Algorithm for Data Mining. Big Data Research, 12: 1–12.

[20]. Joaquin Derrac, Francisco Chiclana, Salvador Garciac, Francisco Herrera (2016) Evolutionary fuzzy k-nearest neighbors algorithm using interval-valued fuzzy sets. Information Sciences, 329 :144–163.

[21]. Nimagna Biswas, Saurajit Chakraborty, Sankha Subhra Mullick, Swagatam Das (2018) A parameter independent fuzzy weighted k-Nearest neighbor classifier. Pattern Recognition Letters, 101: 80–87.

[22]. Dewan Md. Farid,Mohammad Abdullah Al-Mamun, Bernard Manderick , Ann Nowe (2016) An adaptive rule-based classifier for mining big biological data. Expert Systems With Applications, 64 :305–316

[23]. L. Meenachi, S. Ramakrishnan(2021), Metaheuristic Search Based Feature Selection Methods for Classification of Cancer,Pattern Recognition, 119,108079.

[24]. Jin-Mao Wei, Shu-Qin Wang, and Xiao-Jie Yuan(2010) Ensemble Rough Hypercuboid Approach for Classifying Cancers. IEEE Transactions on Knowledge and Data Engineering,22: 381-391.

[25]. Jonathan Aaron Cook (2017)ROC curves and nonrandom data. Pattern Recognition Letters,85:35-41.

Chia-Yen Lee, Bo-Syun Chen(2018) Mutually-exclusive-and-collectively-exhaustive feature election scheme. Applied Soft Computing 68: 961–971.