Vol.64: e21210240, 2021
https://doi.org/10.1590/1678-4324-2021210240
ISSN 1678-4324 Online Edition

Article - Engineering, Technology and Techniques

Decision Tree Based Salp Swarm Optimization for Multi Medical Data Classification with Feature Reduction Technique

Sakunthala Prabha Kadaksham Sarala[1]*
https://orcid.org/0000-0001-8447-9888

Mahesh Chitraivel[1]
https://orcid.org/0000-0003-0091-8025

Raja Soosaimarian Peter Raj[2]
http://orcid.org/0000-0002-7216-2207

[1]VelTech RangarajanDr.Sagunthala R & D Institute of Science and Technology, Department of Information Technology, Chennai, Tamilnadu, India; [2]Vellore Institute of Technology, School of Computer Science and Engineering, Vellore, Tamilnadu, India.

Editor-in-Chief: Alexandre Rasi Aoki
Associate Editor: Fabio Alessandro Guerra

Received: 2021.04.15; Accepted: 2021.06.28.

*Correspondence: sakunthalaprabhaks@veltech.edu.in (S.P.K.S.).

HIGHLIGHTS

- This paper proposes a new hybrid feature selection method.
- This paper proposes a novel Decision Tree based salp optimization algorithm.
- The proposed model is trained using four datasets namely Leukemia, Diffuse Larger B-cell Lymphomas (DLBCL), Lung cancer and Colon.
- This paper produces an accuracy of 98.88.

Abstract: The ambitious task in the domain of medical informatics is medical data classification. From medical datasets, intention to ameliorate human burden with the medical data classification entails to taking in classification designs. The medical data classification is the major focus of this paper, where a Decision Tree based Salp Swarm Optimization (DT-SWO) algorithm is proposed. After pre-processingthe hybrid feature selection method selects the medical data features. The high dimensional features are reduced by Discriminant Independent Component Analysis (DICA) and DT-SWO is to classify the most relevant class of medical data. The details of four datasets namely Leukemia, Diffuse Larger B-cell Lymphomas (DLBCL), Lung cancer and Colon relating to four diseases for heart, liver, cancer and lungs are collected from the UCI machine learning repository. Ultimately, the experimental outcomes demonstrated that the proposed DT-SWO algorithm is suitable for medical data classification than other algorithms.

Keywords: Medical data; Feature selection; Decision Tree based Salp Swarm Optimization; Discriminant Independent Component Analysis and dataset.

INTRODUCTION

In disease diagnosis, the medical data classification can effectively assist physicians and that treat many diseases with appropriate prediction. The medical data classification performance is improved by adopting various efficient efforts. Practically, the class imbalance issues are quite widespread [1,2]. The development of computing technologies with the medical field plays a significant role which leads to acquire the digital storage activities and various medical activities such as prognosis, screening and diagnosis to gain vital knowledge [3,4]. Different kinds of data mining algorithms for medical data are developed owing to knowledge discovery and henceforth, the improvement of medical data accuracy is accomplished by this potentially useful and novel information [5,6]. The two different classifications such as server assigned task and worker selected task are adopted. The aim of this is health care quality improvement and to learn classification in medical data. The medical data classification serves as a purpose for diagnosis and prognosis. Moreover, the noise included in medical data exhibit unique features by missing values and systematic errors [7,8].

Physician's dependency is significant to this system for the precision and accuracy exhibited in the diagnosis. The medical decision support systems computerize the deployment. The larger amounts of digital information stored is relatively easy to acquire. The prediction is the goal of the classification task [9,10], which is quite challenging in medical informatics. The medical data classification is applied with statistical techniques. The arduous task is to learn the properties of the dataset. Each parameter defines the types of medical databases containing an enormous collection of medical data [11,12]. Due to the substantial development of automation, the medical data classification engenders excellent importance in the field of medicine. This classification enhances to achieve the chances of recovery to be significantly high, if diagnosis is successful in the early stages.

Two conflict criterions such as intensification of the optimal solution and the diversification of the search space are defined during the design of meta heuristics. The search space algorithm performance is improved with the help of a genuine balance between these criteria. The local and global search algorithm fusion is the memetic algorithm [13]. To ensure diversification, the global search approach [14] is employed. The error prediction of the classifier in the wrapper methods is the most Discriminant set of features. Around the classification algorithms, a feature selection method is wrapped. The optimal feature set is generated from different kinds of searching algorithms. While compared to filter methods, the features generated by wrapper methods [15,16] are significant. The proposed DT-SWO algorithm selects only the relevant features and discards the highly correlated or noisy features, this is the advantage of the proposed DT-SWO algorithm over Random Forest. To balance the diagnosis of medical classification samples, there exist minor and major classes. In this paper, Decision Tree based Salp Swarm Optimization (DT-SWO) algorithm is proposed for medical data classification. The steps of major contribution are:

- Discriminant Independent Component Analysis (DICA) for feature selection and DT-SWO algorithm for medical data classification is proposed.The proposed Discriminant ICA (*D*ICA) method jointly maximizes the inter-class variance and Negentropy of a given feature. Hence it can able remove all the noisy or highly correlated features.
- We use Leukemia, Diffuse Larger B-cell Lymphomas (DLBCL), and Prostate Tumor and Colon datasets with four diseases.

The concise organization of the paper is: Section 2 explains the related works and section 3 elucidates the problem definition. Section 4 formulates the proposed work followed by the results discussed in section 5. Finally, the paper is concluded in section 6.

Related Work

The medical data classification based techniques were proposed over the last few years. Few of them are discussed in this section.

The feature ranking based method was proposed by Md. ZahangirAlam and coauthors [17] to classify the medical data. Few suitable ranking algorithms were used for dataset feature ranking and the high ranked features are predicted using Random Forest classifier. Ten-fold cross-validations with many other feature ranking algorithms are applied to evaluate the performance. Khanmohammadi and coauthors [18] proposed the model of Gaussian Mixture based Discretization (GMBD) algorithm for medical data classification, which preserves the most recurrent figures of the original dataset considering the multimodal distribution of the numerical variables. The six various publicly available datasets used verifies the efficiency of the GMBD algorithm. Gorzalczany and coauthors [5] proposed Multi-Objective Genetic Fuzzy Optimization (MOGFO) for fuzzy rule-based classification system (FRBCS) design from medical data. The collection of medical FRBCS solutions obtained is balanced by characterizing various stages of accuracy and interpretability. For medical data processing, a rule base representation of special coding-free and original genetic operators is introduced.

Bania and coauthors [19] proposed a selection method of parameter-free greedy ensemble attribute (R-Ensembler). The attribute-attribute relevance measure, attribute significant and attribute class adopts the rough set theory concept. The various rough set produces multiple subset combination, which is combined by Ensembler method. During the attribute selection process time, the new *n* number of intersection method was proposed for the reduction of biasness and the dataset is preprocessed with kNN imputation. From the different subsets of the attribute pool, the Ensembler method presumed to be highly effective, selects the high relevant features from the UCI medical dataset. The emperor penguin (EPO) and social engineering optimization (SEO) (memetic algorithm, a fusion of EPO with SEO) were proposed by Baliarsingh and coauthors [20] for the classification of medical data. SVM is one of the faster classification methods, but the selection of regularization and kernel parameters are the wearisome issues of SVM. Hence, the author used a memetic algorithm for tuning of both regularization and kernel parameters (Memetic based SVM). The Memetic based SVM provided optimal medical data classification whose results are better than other fifteen existing algorithms but its computational complexity is high. Hybridized harmony search and Pareto optimization method were proposed by Dash [13] for the selection of high dimensional features from the medical data. For both feature subset prediction and sample classification, the hybridized method provided high potentiality results for high dimensional databases. For medical data classification, Fan and coauthors [21] proposed a Hybridized model of fuzzy decision tree and integrating case-based clustering method. The authors collected both breast cancer Wisconsin and liver disorder datasets from the UCI machine learning repository. The dataset pre-processed by case based clustering method and fuzzy decision tree is applied for disease identification. These methods provided the average forecasting accuracy of 94.4% for breast cancer and 81.6% for liver disorders when compared to existing methods.

Problem Definition

- The most trivial feature in the dataset is to evaluate the enhancement for assessing classification accuracy. Joining the forecast of various classifiers with Troupe representation has proposed classification accuracy enhancement [8].
- Rendered to prompting wrong diagnosis and treatment of the disease with the previous methods which produced incorrect disease analysis [20].
- For a precise medical diagnosis, fairly less effective features are removed from the existing medical diagnosis strategies.
- The accuracy execution is improved in a decent case by binary encoded else optimization relating to the issue of feature selection [6].
- Based on the specified data with the proper kernel identification is demanding. Depending on the previous information, most of the algorithms choose the right kernel.
- Considerable data sets are well observed to reduce the quantity of operations in the learning mode.

Easier kernel functions with the specific end goal are developed from existing ones.

MATERIALS AND METHODS

The proposed algorithm in this is designed for medical data classification. For this work, we have chosen four disease datasets relating to heart, liver, cancer and lung diseases from UCI machine learning repository. To start with, the medical dataset is pre-processed and the high dimensionality features are reduced in this method. Then, the normal and abnormal diseases are classified with the help of Decision Tree based Salp Swarm Optimization algorithm (DT-SWO). The proposed framework of medical data classification is depicted in Figure 1 and the creeping process of proposed work is:

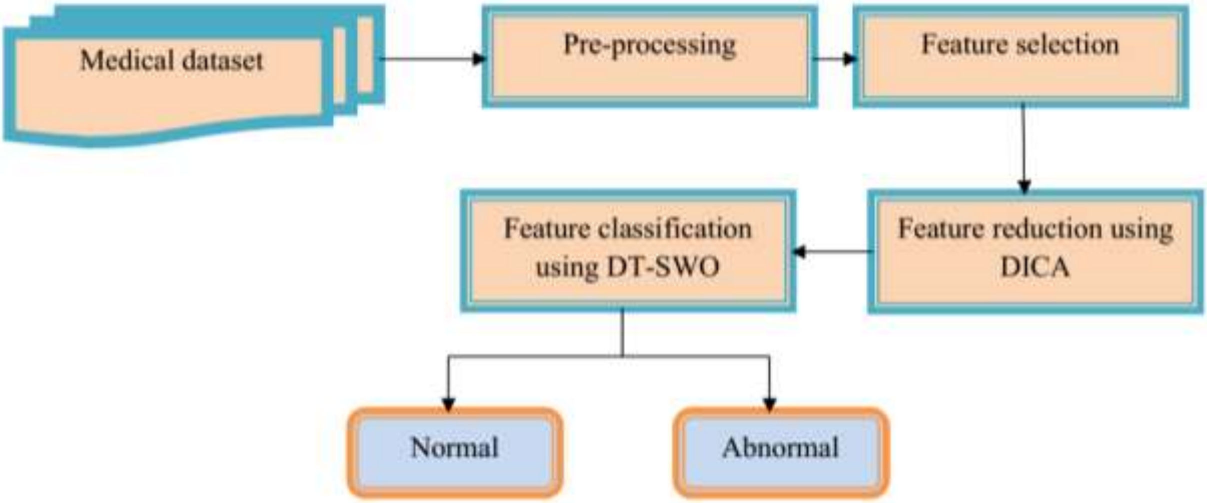
Medical dataset
Pre-processing
Feature selection
Feature reduction using DICA
Feature classification using DT-SWO
Normal
Abnormal

Figure 1. Proposed framework of medical data classification

Pre-processing

First, the medical data is pre-processed to proceed with the further process. The preprocessing capacity consolidates the discretization of numeric qualities, the assurance of trademark subset(s), and treatment of missing values. Picking the best blend of preprocessing systems for a specific dataset is ridiculous without preliminary and assessments. To decrease the proportion of the arrangement space, the amount of characteristics is limited in this stage. Then the non-numerical information is cleared and obtained by the numerical dataset for further procedure.

Feature selection

Thereafter preprocessing, the most relevant features are selected in the feature selection process that may provide good classification performance. Here, the Fisher, T-test and Bayesian Logistic Regression (*BLogReg*) method ranked the features based on the characteristics [22]. This method provided a final ranking taken from the sum of the feature rankings. Out of this, the best and top-ranked features are selected to proceed henceforth.

Feature reduction using Discriminant Independent Component Analysis (DICA)

The Negentropy maximization obtains the independent features and lower dimensions with multivariate data in DICA (Discriminant Independent Component Analysis) method. Simultaneously it maximizes the sum of the marginal Negentrophy of independent extracted features and Fisher criterion in DICA. In order to develop a better classification, the DICA combines the properties of Independent Component Analysis (ICA) [23].

Negentropy maximization for independent feature extraction

Subsequent to feature selection, high dimensional features of medical data are reduced by means of DICA method. For non-Gaussian random variables, Negentrophy is the better statistical scale and the marginal Negentropy approximation is given in Equation (1):

$$R(x_i) \approx k_1(H(M^1(x_i)))^2 + k_2(H(M^1(x_j)))^2 - H(M^2(\sigma)))^2 \tag{1}$$

where, x_i is the standard deviation and the same mean with the univariate Gaussian distribution is σ. For random vector x_i, the below equation is used to prove the random vectors M^1 and M^2. Where, $k_1 = 36/(8\sqrt{3}-9)$ and $k_2 = 16\sqrt{3}-27)$. The values of $M^1(x_i)$ and $M^2(x_i)$ is given in Equations (2) to (4).

$$M^1(x_i) = x_i^3 \tag{2}$$

$$M^2(x_i) = \frac{1}{d_1}\log\cosh d_1 x_i, \qquad 0 < d_1 \le 1 \tag{3}$$

$$M^2(x_i) = -\exp\left({x_i^2}/{2} \right) \tag{4}$$

The Lagrange formula in equation (5) explains the unit covariance with maximization sum of the marginal Negentropy.

$$\hat{L}(N) = \sum_{i=1}^{R}\left[H(M(n_i^T z)) - H(M(\sigma))\right]^2 + \sum_{i=1}^{R}\alpha_i (n_j^T n_j - 1) \tag{5}$$

From Equation (5), the target function maximization is attained by the features. The Lagrange formula in Equation (6) defines the independent extracted features in Negentropy and the functional criterion of classification performance is maximized by an optimization problem.

$$\hat{L} = \hat{L}(N) + k\varphi(N,Z,D) \tag{6}$$

Therefore, the constant is k and the given D and $\hat{L}(N)$ with the feature classification of X is measured by the function φ . The below Equations (7) to (10) explains the rules of learning.

$$\Delta n_i = \lambda(\gamma_i (H(Z_n(n_i^T Z)) + k\frac{\partial\varphi(N,Z,D)}{\partial m_i} + 2\alpha_i n_i) \tag{7}$$

$$\alpha_i = -\frac{1}{2}\gamma_i H(x_i m(x_i)) \tag{8}$$

$$\gamma_j = 2\left(-\frac{\sum_{n=1}^{n}\exp\left(-\frac{x_{in}^2}{2}\right)}{N} + \frac{1}{2} \right) \tag{9}$$

which applies for the symmetric orthogonalization of N.

$$N \leftarrow (NN^T)^{(-0.5)} N \tag{10}$$

Performance of classification for functional measures

The classification performance of the function is given in Equation (11):

$$\varphi(N,Z,D) = \sum_{i=1}^{R}\log\frac{n_i^T C_S n_i}{n_i^T C_w n_i} = \sum_{i=1}^{R}\log\frac{\sum_{d=1}^{D} M_s(\sigma_{id}-\sigma_i)^2}{\sum_{d=1}^{D} M_s \rho_{id}^2} \tag{11}$$

Hence,

$$\sigma_{id} = \frac{1}{M_S}\sum_{m\in class} x_{in} \tag{12}$$

$$\rho_{id} = \frac{1}{M_S}\sum_{m\in class}(x_{in} - \rho_{id})^2 \tag{13}$$

From the above equations, the scatter matrix and the between-scatter matrix class are C_w and C_S are given in Equation (14). The entire amount of class is C and the mean σ_{id} and variance are ρ_{id} of i^{th} the feature class is D.

$$\frac{\partial\varphi(N,Z,D)}{\partial n_{li}} = 2\sum_{d=1}^{D}\sum_{m\in Class}(U_{id} - V_{id})\,Z_{in} \tag{14}$$

Here,

$$U_{id} = \frac{(\sigma_{id} - \sigma_d)}{\sum_{d^{`}=1}^{D} M_{s^{`}}(\sigma^{`}_{id} - \sigma_i)^2} \tag{15}$$

$$V_{id} = \frac{(x_{in} - \sigma_i)}{\sum_{d^{`}=1}^{D} M^{`}_{s}\rho^{2}_{id^{`}}} \tag{16}$$

From Z, obtain the l^{th} dimension of the m^{th} sample regarding Z_{in} [24]. For feature reduction, the separating matrix N is examined using DICA. The cost function of the gradient is used to obtain the independent variables and the approach of gradient tends to local optima [22,25]. Due to the complexity involved in ICA, gradient-based method never achieves high accuracy level. Thus, DICA is more efficient to solve the complexities. At the same time, the Negentropy and the fisher ration are maximized with the help of DICA. Algorithm 1 explains the steps of DICA for feature reduction. Finally, the DICA method is used to reduce the high dimensionality features from the medical data.

Decision Tree based Salp Swarm Optimization algorithm (DT-SWO) for medical data classification

Next to feature reduction, the Decision Tree based Salp Swarm Optimization (DT-SWO) algorithm is used to classify the medical data.

Decision Tree

Based on each attribute A, the decision tree learning algorithm calculates the information gain G, which is expressed in Equation (17):

$$G(R,A) = entropy(R) - \sum_{u\in values}\frac{|R_u|}{|R|}entropy(R_u) \tag{17}$$

From the above equation (17), the total input space is R and the subset for R is R_u. The value u is an attribute A. The $\sum_{j=1}^{D} -P_j \log_2(P_j)$ gives the entropy R and the probability classes of “j” are P_j. Where, C is the highest information gain with the attribute and the root node of the tree is chosen [21]. By using the training subspace $S - \{S_C\}$, the novel decision tree is built over each value of C. If the entire instances present in the training subspace then the decision tree is formed. The 0 denotes the normal and 1 denote the anomaly class assignment to test for detecting an anomaly.

Evolving DT by SWO (DT-SWO) data classification

In medical data classification, the Salp Swarm Optimization (SWO) enhances the accuracy performance of Fuzzy Decision Tree (FDT). Based on each data, the SWO algorithm will determine the best number of fuzzy terms. After every new fuzzy terms are numbered, the fitness function is re-calculated. The classification accuracy of medical data is represented by the fitness function of SWO [26]. Next, continue the process of SWO till the end of the stopping criterion.

(i) Representation of the solution

Each solution is limited to binary 0 and 1 in medical data classification. The binary should be developed for the salp swarm optimization algorithm that is utilized by the feature classification work. The one-dimensional vector defines the solution, according to the number of features in the real dataset and represents the vector length. All vector cells contain 0 and 1 value. The relevant medical data class is chosen as 1 else 0. The continuous value into binary value mapping is represented in Equation (18).

$$X_{nm} = \begin{cases} 1 & if\, Y_{nm} > 0.5 \\ 0 & otherwise \end{cases} \tag{18}$$

Hence, the solution vector Y represents a discrete form X_{nm}. At dimension m, the continuous position of the search agent n is Y_{nm}. For a dataset of 7 attributes, the sample features for subset solution is explained in Figure 2. From this, 4 selected features (1, 4, 6 and 7) are for classification task and the remaining one value is discarded.

1	0	0	1	0	1	1

Figure 2. Salp solution representation for classification

(ii) Fitness function Evaluation

In this section, medical data classification is represented as the multi-objective optimization issue and it is accomplished by two conflicting objectives such as a minimal number of selected features and maximal amount of classification accuracy. Based on all the search agents, the classification performance is enhanced by the fitness function. The selected features are balanced among classification accuracy and selected features in each solution. The fitness function is assesses the fitness function in SSA, as shown below in Equation (19):

$$F = \delta\, Error(d) + \phi \frac{|f|}{|t|} \tag{19}$$

From the above equation (19), the classifier error rate of the identified subset is $Error(d)$. The feature reduction and the classification are controlled by the parameters ϕ and δ. The identified feature subset is $|f|$ and the total numbers of features are $|t|$. In our work, the value of δ is to 0.1, then $\phi = (1-\delta)$. (i.e.) the value of $\phi = 0.9$.

(iii) The output of classification results

The DT-SWO is involved for medical data classification. Hence the step by step process of DT-SWO algorithm for medical data classification is shown in Figure 3. Here, the parameters of the decision tree moth flame and tunicate salp swarm algorithm are initialized, thereby using the maximal number of iterations. We define the classification accuracy and its efficiency in the multi-objective function. The decision tree position is updated by iterating from equation (1)-(8) thereby the equations (9) - (15) updates the position of Salp swarm. Then, evaluate the fitness function until the stopping criterion is met. The classification is checked for the required optimal and relevancy of the particular disease, otherwise, the process is repeated. Finally, the disease is diagnosed with best classification and is more accurate.

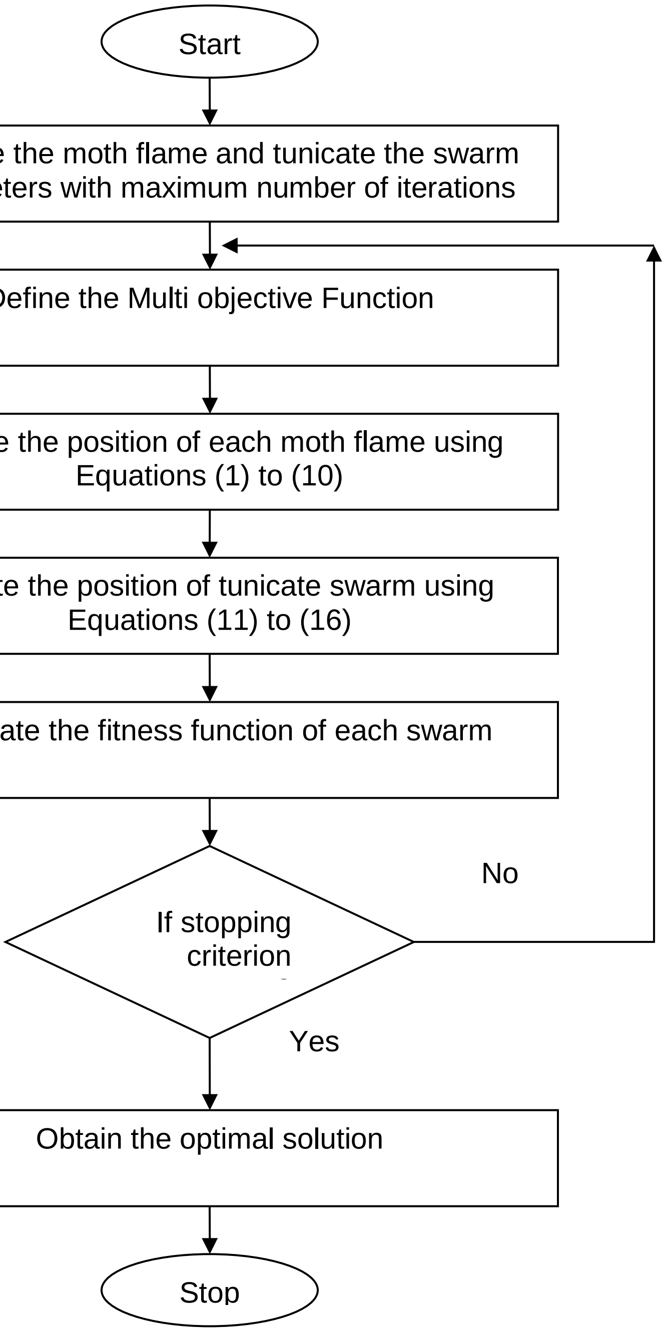
Start
Initialize the moth flame and tunicate the swarm parameters with maximum number of iterations
Define the Multi objective Function
Update the position of each moth flame using Equations (1) to (10)
Update the position of tunicate swarm using Equations (11) to (16)
Evaluate the fitness function of each swarm
No
If stopping criterion
Yes
Obtain the optimal solution
Stop

Figure 3. Medical data classification using DT-SWO algorithm

RESULT AND DISCUSSION

The proposed work performance for medical data classification is evaluated in this section. Moreover, MATLAB 2016a with an i5 processor and 4GB RAM was used to actualize the performance of proposed work. Four disease datasets for heart, liver, cancer and lung diseases from UCI machine learning repository is used. Furthermore, the comparison experiment with different feature selection and classification method evaluates the performance of the proposed work. The data description model is explained as follows.

Dataset description

Four disease datasets relating to heart, liver, cancer and lung diseases from UCI machine learning repository are chosen. In this work, Leukemia, Diffuse Larger B-cell Lymphomas (DLBCL), Prostate Tumor and Colon datasets are used. The description of dataset is depicted in Table 1 and dataset details are explained as follows:

- *Leukemia dataset:* The Leukemia dataset is widely used in the research field of medical classification. This dataset consists of 73 samples. Here, the Acute Myeloid Leukemia and Acute Lymphoblastic Leukemia are the class attributes. Among the entire samples, 53 samples are utilized for training and the remaining 20 samples are for the testing operation.
- *Diffuse Larger B-cell Lymphomas (DLBCL):* This dataset consists of 79 samples that are collected from a patient. For this experiment, there are 56 samples for employed for training and 23 samples are employed for testing operation.
- *Lung cancer dataset:* This dataset has the lung cancer gene of Adenocarcinoma and Malignant Pleural Mesothelioma. The dataset consists of 185 tissue samples. From this, 154 samples are utilized for training and the balance 31 samples are for the testing operation.
- *Colon dataset:* The colon dataset contains two classes, namely Adenocarcinomas and gastrointestinal, which contains 143 samples with 2100 genes. From this, 105 samples are availed for training and the remaining 38 samples for the testing procedure.

Table 1. Dataset description for medical data classification

Parameters	**Name of the dataset**			
	Leukemia dataset	***Diffuse Larger B-cell Lymphomas***	***Lung cancer dataset***	***Colon dataset***
Instances	294	672	899	678
Attributes	72	72	72	72
Missing Values	Yes	Yes	Yes	Yes
Class	2	2	2	2

Performance metrics

The medical data classification is gaugedby various performance metrics such as accuracy, specificity, sensitivity, precision and recall. The below equations express each performance metrics.

$$\mathrm{Pr}\,ecison = \frac{PT}{PT + PF} \tag{20}$$

$$F_{measure} = \frac{2TP}{(2TP + FP + FN)} \tag{21}$$

$$A = \frac{TP + TN}{Total\,no.of\ samples} \times 100 \tag{22}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{23}$$

$$Specificity = \frac{TN}{TN + FN} \times 100 \tag{24}$$

Where,

TP - True Positive is the correctly identified abnormal people

TN – True Negative is correctly identified normal people

FP - False Positive is the incorrectly identified abnormal people

FN – False Negative is incorrectly identified normal people

Performance evaluation

The proposed performance analysis based on accuracy is shown in Table 1. This work uses four datasets namely Leukemia, Diffuse Larger B-cell Lymphomas (DLBCL), Prostate Tumor and Colon datasets with four diseases namely heart, liver, cancer and lung diseases. Here, three sets of optimizations are sorted that determine the percentage of accuracy. The Leukemia dataset classifies 86.21% without optimization and MFO showed 90.5% and the proposed DT-SWO produces 95.6% accuracy results. Based on Diffuse Larger B-cell Lymphomas dataset, the accuracy results are 89.11%, 95.64% and 98.89% without optimization, MFO and DT-SWO respectively. For Lung cancer dataset, the accuracy results are 80.89%, 96.02% and 97.85% without optimization, MFO and DT-SWO respectively. Based on Colon dataset 90.42%, 93.5% and 98.88% are the accuracy results produced without optimization, MFO and DT-SWO respectively.

Table 2. Performance analysis of accuracy (%).

Datasets	Without Optimization	MFOA	DT-SWO
Leukemia dataset	86.21	90.5	95.6
Diffuse Larger B-cell Lymphomas	89.11	95.64	98.89
Lung cancer dataset	80.89	96.02	97.85
Colon dataset	90.42	93.5	98.88

Figure 4 explains the results based on proposed performance metrics. Here, four datasets like Leukemia, Diffuse Larger B-cell Lymphomas (DLBCL), Lung cancer and Colon datasets with four diseases such as heart, liver, cancer and lung disease are chosen. Moreover, the accuracy, specificity, sensitivity, precision and F-measure validates the performance of proposed medical data classification. Based on the system base, the execution of esteem can be changed. From this experiment; the leukemia dataset provides 97.9% accuracy, 96.09% specificity, 96.09% sensitivity, 95.81% precision and 97.9% F-measure outputs. Likewise, the Diffuse Larger B-cell Lymphomas achieved 97.81% accuracy, 93.9% specificity, 91.89% sensitivity, 92.09% precision and 93.9% F-measure outcomes. Furthermore, 92.78% accuracy, 90.89% specificity, 95.81% sensitivity, 93.9% precision and 95.6% F-measure results are produced by lung cancer dataset. Finally, the colon showed 90.6% accuracy, 95.6% specificity, 97.85% sensitivity, 90.89% precision and 90.9% F-measure outcomes.

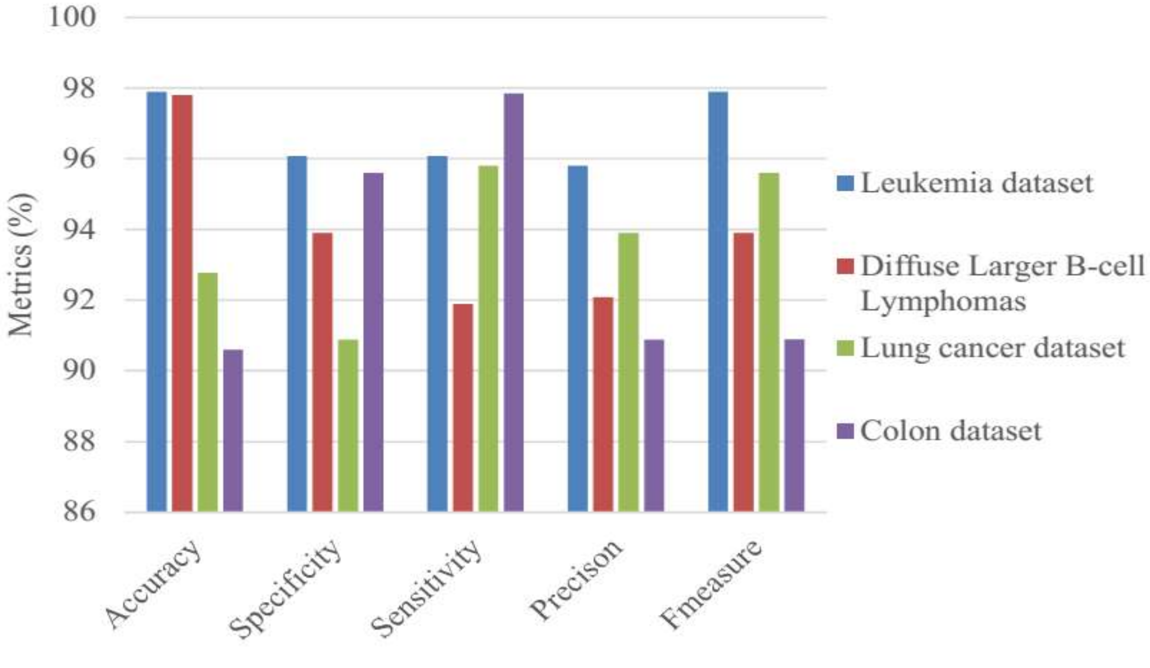

Figure 4. Proposed performance metrics results.

Figure 5. shown the number of selected features with state-of-art results. In this experiment, we have chosen ICA, PCA and proposed DICA are the feature selection approaches. Thereafter feature selection, the feature reduction operation is performed via DICA method. We select 5, 10, 15 and 20 features for this experiment. While compared to all the methods, the proposed DICA provides higher performance results and it is well appropriate of feature reduction in medical data classification.

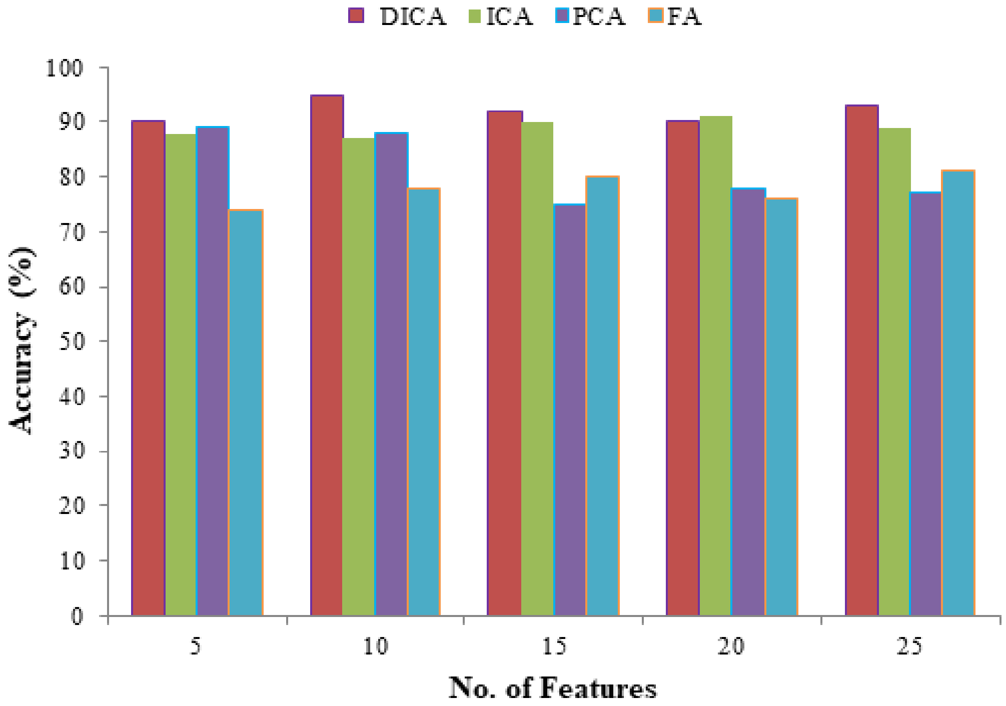

Figure 5. Number of selected features with state-of-art results.

Figure 6 explains the convergence performance of the proposed optimization algorithm. Hence, the convergence performance is one of the important methods to validate the performance of optimization algorithms. The convergence performances increase the accuracy level and performance. So, we use state-of-art convergence performance analysis. For our experiment, the Firefly Algorithm (FA), Salp Swarm Optimization (SWO), Particle Swarm Optimization (PSO), Moth Flame Optimization (MFO) and proposed Decision Tree based Salp Swarm Optimization (DT-SWO) algorithms were considered. Finally, the proposed DT-SWO provided better convergence results than other existing algorithms.

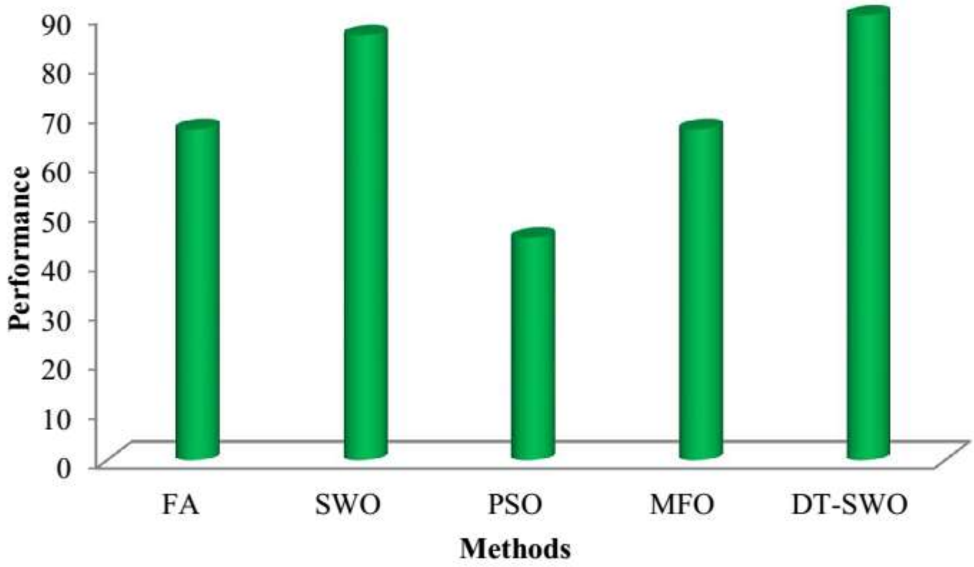

Figure 6. State-of-art results with respect to convergence performance

Table 3 explains the proposed feature selection method with feature reduction results. In this, four datasets are chosen namely Leukemia dataset, Diffuse Larger B-cell Lymphomas (DLBCL), Lung cancer dataset and Colon dataset with four diseases such as heart disease, liver disease, cancer disease and lung disease. All four datasets are compared with FA, ICA and DICA methods. When compared to all methods, the proposed DICA method provided better feature selection performance. Based on Leukemia dataset, the DICA provided 97.90%, 95.81% and 96.09% feature reduction performance. According to Diffuse Larger B-cell Lymphomas, the DICA delivers 90.60%, 92.78% and 90.89% feature reduction performance. Based on Lung cancer dataset, the DICA delivers 93.06%, 97.81% and 90.89% feature reduction performance. According to colon dataset, the DICA provides 93.90%, 91.89% and 92.09% feature reduction performance.

Table 3. Proposed selection approach (Ttest+Blogreg+Fisher) with feature reduction performance.

Dataset	***Methods***				
	Number of selected features	***Without feature extraction***	***FA***	***ICA***	***DICA***
Leukemia dataset [27]	1234	86	87.98	79.56	97.90
			75.91	94.90	95.81
			78.67	90.78	96.09
Diffuse Larger B-cell Lymphomas [28]	138	89	67.98	78.67	90.60
			78.73	89.68	92.78
			78.82	60.78	90.89
Lung cancer [29]	1000	88	88.89	80.79	93.06
			80.89	87.17	97.81
			89.90	90.78	90.89
Colon [30]	348	95	89.78	88.89	93.90
			78.12	89.90	91.89
			78.90	89.78	92.09

Table 4 explains the state-of-art results based on classification. Four datasets namely Leukemia, Diffuse Larger B-cell Lymphomas (DLBCL), Lung cancer and Colon datasets with four diseases such as heart, liver, cancer and lung disease is chosen here. Support Vector Machine (SVM), Decision tree, Naive Bayes, Kernel SVM (K-SVM), Salp Swarm Optimization (SWO) and Decision Tree based Salp Swarm Optimization (DT-SWO) algorithms are used. The proposed DT-SWO delivers better medical data classification performance than other methods because of the ability in determining the best number of fuzzy terms. Finally, the Diffuse Larger B-cell Lymphomas (DLBCL), Leukemia, Lung cancer and Colon datasets achieved 95%, 97%, 94% and 98% classification performance.

Table 4. State-of-art results based on the classification.

Methods	**Dataset**			
	Diffuse Larger B-cell Lymphomas	***Leukemia dataset***	***Lung cancer***	***Colon dataset***
SVM	90%	88%	91%	94%
Decision tree	89%	89%	95%	87%
Naive Bayes	73%	78%	86%	95%
Kernel SVM	67%	84%	88%	80%
SWO	78%	89%	67%	67%
DT-SWO	95%	97%	94%	98%

CONCLUSION

This paper proposed a Decision Tree based Salp Swarm Optimization (DT-SWO) algorithm for medical data classification. The medical data details are collected from the UCI machine learning repository by choosing four datasets namely Leukemia, Diffuse Larger B-cell Lymphomas (DLBCL), Lung cancer and Colon datasets relating to four diseases for heart, liver, cancer and lungs. The performance metrics such as accuracy, specificity, sensitivity, precision and F-measurevalidates the performance of proposed work. The convergence performance of proposed DT-SWO provides better convergence performances than other algorithms such as PSO, SWO, MFO and DT. The feature reduction performances of DICA is higher than other methods such as FA and ICA. Finally, the proposed DT-SWO accomplishes better medical data classification performance than other existing algorithms such as SVM, K-SVM, SWO and DT.

REFERENCES

1. Gan D, Shen J, An B, Xu M, Liu N. Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Comput Ind Eng [Internet]. 2020 Feb;140:106266. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0360835219307351
2. Seera M, Lim CP. A hybrid intelligent system for medical data classification. Expert Syst Appl [Internet].2014 Apr; 41(5):2239–49. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417413007562
3. Shen L, Chen H, Yu Z, Kang W, Zhang B, Li H, et al. Evolving support vector machines using fruit fly optimization for medical data classification. Knowledge-Based Syst [Internet]. 2016 Mar;96:61–75. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0950705116000125
4. Tu MC, Shin D, Shin D. A comparative study of medical data classification methods based on decision tree and bagging algorithms. 8th IEEE Int Symp Dependable, Auton Secur Comput DASC 2009. 2009;4(1):183–7.
5. Gorzałczany MB, Rudziński F. Interpretable and accurate medical data classification – a multi-objective genetic-fuzzy optimization approach. Expert Syst Appl [Internet]. 2017 Apr;71:26–39. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417416306467
6. Tang T, Wang P. A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis. Ind Eng Manag Syst. 2005;4(1):102–8.
7. AlMuhaideb S, Menai MEB. An Individualized Preprocessing for Medical Data Classification. Procedia Comput Sci [Internet]. 2016;82:35–42. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877050916300205
8. Dennis B, Muthukrishnan S. AGFS: Adaptive Genetic Fuzzy System for medical data classification. Appl Soft Comput [Internet]. 2014 Dec;25:242–52. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1568494614004852
9. Cohen S, Jannot A-S, Iserin L, Bonnet D, Burgun A, Escudié J-B. Accuracy of claim data in the identification and classification of adults with congenital heart diseases in electronic medical records. Arch Cardiovasc Dis [Internet]. 2019 Jan;112(1):31–43. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1875213618301839
10. Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath. Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients. Int J Eng Adv Technol [Internet]. 2012;1(3):147–51. Available from: http://www.galaxy.gmu.edu/interface/I01/I2001Proceedings/Jbreault
11. Babu PH, Gopi ES. Medical Data Classifications Using Genetic Algorithm Based Generalized Kernel Linear Discriminant Analysis. Procedia Comput Sci [Internet]. 2015;57:868–75. Available from: https://linkinghub.elsevier.com/retrieve/pii/S187705091502027X
12. Huda S, Yearwood J, Jelinek HF, Hassan MM, Fortino G, Buckland M. A Hybrid Feature Selection With Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis. IEEE Access [Internet]. 2016;4:9145–54. Available from: http://ieeexplore.ieee.org/document/7809136/
13. Dash R. An Adaptive Harmony Search Approach for Gene Selection and Classification of High Dimensional Medical Data. J King Saud Univ - Comput Inf Sci [Internet]. 2021 Feb;33(2):195–207. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1319157817304883
14. Sivasankar S, Nair S, Judy M. Feature Reduction in Clinical Data Classification using Augmented Genetic Algorithm. Int J Electr Comput Eng [Internet]. 2015 Dec 1;5(6):1516. Available from: http://ijece.iaescore.com/index.php/IJECE/article/view/5779
15. Karim AM, Güzel MS, Tolun MR, Kaya H, Çelebi F V. A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing. Biocybern Biomed Eng [Internet]. 2019 Jan;39(1):148–59. Available from: https://linkinghub.elsevier.com/retrieve/pii/S020852161830322X
16. Lee CS. Feature reduction using a GA-Rough hybrid approach on bio-medical data. Int Conf Control Autom Syst [Internet]. 2011;1339–43. Available from: https://ieeexplore.ieee.org/abstract/document/6106133/
17. Alam MZ, Rahman MS, Rahman MS. A Random Forest based predictor for medical data classification using feature ranking. Informatics Med Unlocked [Internet]. 2019;15:100180. Available from: https://linkinghub.elsevier.com/retrieve/pii/S235291481930019X
18. Khanmohammadi S, Chou C-A. A Gaussian mixture model based discretization algorithm for associative classification of medical data. Expert Syst Appl [Internet]. 2016 Oct;58:119–29. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417416301440
19. Bania RK, Halder A. R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data. Comput Methods Programs Biomed [Internet]. 2020 Feb;184:105122. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0169260719306972

20. Baliarsingh SK, Ding W, Vipsita S, Bakshi S. A memetic algorithm using emperor penguin and social engineering optimization for medical data classification. Appl Soft Comput [Internet]. 2019 Dec;85:105773. Available from: https://linkinghub.elsevier.com/retrieve/pii/S156849461930554X
21. Fan C-Y, Chang P-C, Lin J-J, Hsieh JC. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput [Internet]. 2011 Jan;11(1):632–44. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1568494609002774
22. Cawley GC, Talbot NLC. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics [Internet]. 2006 Oct 1;22(19):2348–55. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl386
23. Dhir CS, Soo-Young Lee. Discriminant Independent Component Analysis. IEEE Trans Neural Networks [Internet]. 2011 Jun;22(6):845–57. Available from: http://ieeexplore.ieee.org/document/5756242/
24. Liu K-H, Li B, Zhang J, Du J-X. Ensemble component selection for improving ICA based microarray data prediction models. Pattern Recognit [Internet]. 2009 Jul;42(7):1274–83. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0031320309000508
25. Mollaee M, Moattar MH. A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybern Biomed Eng [Internet]. 2016;36(3):521–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0208521616300973
26. Tubishat M, Idris N, Shuib L, Abushariah MAM, Mirjalili S. Improved Salp Swarm Algorithm based on opposition based learning and novel local search algorithm for feature selection. Expert Syst Appl [Internet]. 2020 May;145:113122. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417419308395
27. Raetz EA, Perkins SL, Bhojwani D, Smock K, Philip M, Carroll WL, et al. Gene expression profiling reveals intrinsic differences between T-cell acute lymphoblastic leukemia and T-cell lymphoblastic lymphoma. Pediatr Blood Cancer [Internet]. 2006 Aug;47(2):130–40. Available from: https://onlinelibrary.wiley.com/doi/10.1002/pbc.20550
28. Spidlen J, Breuer K, Rosenberg C, Kotecha N, Brinkman RR. FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. Cytom Part A [Internet]. 2012 Sep;81A(9):727–31. Available from: https://onlinelibrary.wiley.com/doi/10.1002/cyto.a.22106
29. Hong Z-Q, Yang J-Y. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern Recognit [Internet]. 1991 Jan;24(4):317–24. Available from: https://linkinghub.elsevier.com/retrieve/pii/003132039190074F
30. Schneider M, Huber J, Hadaschik B, Siegers GM, Fiebig H-H, Schüler J. Characterization of colon cancer cells: a functional approach characterizing CD133 as a potential stem cell marker. BMC Cancer [Internet]. 2012 Dec 20;12(1):96. Available from: http://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-12-96