Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2021.016817

ARTICLE

An Effective Feature Generation and Selection Approach for Lymph Disease Recognition

Sunil Kr. Jha[1,*] **and Zulfiqar Ahmad**[2]

[1]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China

[2]Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072, China

[*]Corresponding Author: Sunil Kr. Jha. Email: 002891@nuist.edu.cn

Received: 29 March 2021 Accepted: 15 July 2021

ABSTRACT

Health care data mining is noteworthy in disease diagnosis and recognition procedures. There exist several potentials to further improve the performance of machine learning based-classification methods in healthcare data analysis. The selection of a substantial subset of features is one of the feasible approaches to achieve improved recognition results of classification methods in disease diagnosis prediction. In the present study, a novel combined approach of feature generation using latent semantic analysis (LSA) and selection using ranker search (RAS) has been proposed to improve the performance of classification methods in lymph disease diagnosis prediction. The performance of the proposed combined approach (LSA-RAS) for feature generation and selection is validated using three function-based and two tree-based classification methods. The performance of the LSA-RAS selected features is compared with the original attributes and other subsets of attributes and features chosen by nine different attributes and features selection approaches in the analysis of a most widely used benchmark and open access lymph disease dataset. The LSA-RAS selected features improve the recognition accuracy of the classification methods significantly in the diagnosis prediction of the lymph disease. The tree-based classification methods have better recognition accuracy than the function-based classification methods. The best performance (recognition accuracy of 93.91%) is achieved for the logistic model tree (LMT) classification method using the feature subset generated by the proposed combined approach (LSA-RAS).

KEYWORDS

Disease data mining; feature selection; classification; lymph; diagnosis

Nomenclature

K	Cohen's kappa coefficient
C	Accuracy in %
LSA	Latent semantic analysis
RSA	Ranker search

1 Introduction

The machine learning approaches are playing a vital role in the development of computer-aided diagnosis systems [1–3]. The highly efficient machine learning-based soft disease diagnosis system provides an economical, non-invasive, and quick diagnostic facility for the patient. Such a system also eases the effort of physicians in decision-making and interpretation of disease diagnosis results. The lymphatic system improves the immune system, maintains the balance of body fluids, removes the waste product, bacteria, and virus, and supports the absorption of nutrients, etc. [4,5]. Any blockage and infection of the tissue in lymph vessels result in lymphoma, lymphadenitis, and lymphedema, etc. [6]. Imaging techniques are used in the examination of lymph nodes [7,8]. Moreover, the classification approaches of machine learning can be implemented to improve the prediction accuracy of the initial status of the lymph node by modeling the measurements of imaging techniques and physical observations.

Generally, the classification techniques of machine learning are the backbone of the soft disease diagnosis system for the class recognition of the specific disease in diagnosis purposes by analyzing preliminary observations and instrumental measurements [9–11]. The performance of the classification methods has been affected by the size of the data, the number of attributes, nature of attributes, noise and outliers in data, and uneven distribution of instances of different attributes, etc. [12,13]. Consequently, addressing the earlier issues is crucial for the real-time diagnosis and recognition of diseases by a machine learning-based system. Among the previous concerns, reducing the dimensionality (attributes) of a dataset is one of the significant steps for the disease recognition performance improvement of the classification method [13–16]. The dimensionality of any disease data set can be reduced in two ways (i) selecting a significant subset of attributes from the original attribute set, and (ii) generating novel features by a transformation of the original attributes of the dataset into new feature space and subsequently, selecting a significant subset of features. In the present study, both of the earlier approaches of dimensionality reduction have been implemented for efficient recognition of lymph disease. Moreover, a novel approach of feature generation and selection has been implemented for the dimensionality reduction of the lymph dataset and its effect on the recognition performance of five different classification methods has been examined. Besides, some other feature generation methods like principal component analysis (PCA), and attribute selection methods based on the genetic algorithm (GA), greedy forward and backward search, random search, and rank search, etc. have been implemented for performance comparison of the proposed approach.

1.1 Literature Survey

Classification approaches to machine learning have been implemented in the recognition of lymph disease in past studies [17–28]. Mainly, single classification methods [19,22], in combination with the feature selection approach [17,20,21,24], and combination with other classification methods [18,23,25,27,28] have been used in the analysis of the lymph disease dataset. Tab. 1 presents a short review of the classification approaches used in the analysis of the lymph disease dataset. Based on category wise analysis of the classification methods, it is obvious that the tree-based classification methods have been used mostly in the lymph disease recognition [17,22,28]. The maximum accuracy of 92.2% has been achieved using the selected features and random forest (RF) classifier [17]. The artificial neural network (ANN) classification methods implemented in some studies [24,25], like multi-layer perceptron (MLP) [24], and hybrids of radial basis function neural network and evolutionary algorithm [25]. The hybrid ANN method achieved improved recognition accuracy of 85.47% than MLP. The Bayesian classifiers [23,24] result in

average recognition accuracy. Besides, in some recent studies, deep learning approaches have been implemented in disease diagnosis, like convolutional long short-term memory neural network in heartbeat classification [29], atrial fibrillation detection using adaptive residual network [30], and arrhythmia classification using fully connected neural networks [31], etc.

Table 1: A review of previous approaches in lymph disease classification

Classification method	Classification accuracy in %	Ref.
Feature selection (Genetic algorithm (GA), Principal component analysis, and Relief-F, etc. + Random forest	75.5%–92.2%	[17]
Data gravitation classification (DGC+), DGC, and K nearest neighbour, etc.	75.99%–81.90%	[18]
Differential evolution	$73.93 \pm 2.68\%$–$80.79 \pm 1.66\%$	[19]
Evolutionary instance selection-rough set feature selection, and fuzzy rough set theory, etc.	73.87%–82.65%	[20]
Information gain, Relief (RLF), and Consistency based subset evaluation (CNS), etc. + naïve Bayes (NB) and C4.5	79.67%–83.24%–73.09%	[21]
Functional tree, and Sequential minimal optimization for training a support vector classifier	86.49%	[22]
NB, Evolutional naïve Bayes-classification accuracy (ENB-ACC), and SBC-ACC), etc.	78.40%–85.39%	[23]
NB, multi-layer perceptron, and Feature selection approach + J48	77.02%–84.46%	[24]
Hybrid of radial basis function neural network and co-operative co-evolutionary algorithm, GA-RBFNN, and Decaying radius selection clustering	85.47%, 83.04%, and 91.01%	[25]
Forest method using nested dichotomies	82.16%–83.51%	[26]
Artificial immune System based self-adaptive attribute weighting naïve Bayes, naïve Bayes (NB), and correlation based feature selection + naïve Bayes (CFSWNB), etc.	78.78%–85.97%	[27]
Super parent-one-dependence estimators, averaged one-dependence estimators, and correlation feature selection based weighted, etc.	85.39%–85.92%	[28]
Latent semantic analysis (LSA)-ranker search (RAS)-logistic model tree (LMT)	**93.91%**	**Present study**

1.2 Motivation and Contribution of Present Study

It is obvious from the literature survey that the selected feature improves the recognition performance of the classification methods. The selection of an optimal set of features that can result in the maximum lymph disease recognition accuracy is still an existing challenge. With this motivation, an effective approach of feature generation (latent semantic analysis (LSA)) and selection (ranker search (RAS)) has been proposed which results in the maximum lymph disease

recognition accuracy of classification methods. The main findings of the present study include the followings:

- An efficient approach of dimensionality reduction using the combination of feature generation and selection approaches.
- An effective recognition approach of lymph disease using the combination of an optimal subset of selected features.
- Comprehensive performance comparison of the proposed feature generation and selection method with other methods of attribute selection.
- Performance validation of the proposed approach using functions and tree-based classification methods.
- The maximum recognition accuracy of the classification methods using the feature subset selected with the proposed approach with methods in the reviewed literature.

Details of the lymph disease dataset are available in Section 2, Section 3 presents the proposed approach, analysis results are presented in Section 4, discussed in Section 5, and concluded in Section 6.

2 Experimental Lymph Dataset

The lymphography dataset was accessed from the University of California Irvin's (UCI) machine learning repository [32]. A description of the lymphography dataset is available in Tab. 2. It contains two instances of normal cases and eighty-one, sixty-one, and four instances of metastases, malign lymph, and fibrosis cases of lymph disease, respectively. Fifteen nominal attributes (lymphatics, block of afferent, and block of lymph c, etc.) and three numerical attributes (lymph nodes diminish, lymph nodes enlarge, and number of nodes) of each of the instances have been observed without missing values.

Table 2: Details of the lymph disease dataset

Attribute name	Attribute type	Label/value range of attribute	Count of label of attribute
Lymphatics	Nominal	Normal	2
		Arched	67
		Deformed	46
		Displaced	33
Block of afferent	Nominal	No	66
		Yes	82
Block of lymph c	Nominal	No	112
		Yes	26
Block of lymph s	Nominal	No	141
		Yes	7
By pass	Nominal	No	112
		Yes	36
Extravasates	Nominal	No	73
		Yes	75

(Continued)

Table 2 (continued)

Attribute name	Attribute type	Label/value range of attribute	Count of label of attribute
Regeneration of	Nominal	No	138
		Yes	10
Early uptake	Nominal	No	44
		Yes	104
Changes in lymph	Nominal	Bean	6
		Oval	77
		Round	65
Defect in node	Nominal	No	3
		Lacunar	49
		Lacunar marginal	46
		Lacunar central	50
Changes in node	Nominal	No	6
		Lacunar	42
		Lacunar marginal	75
		Lacunar central	25
Changes in structure	Nominal	No	2
		Grainy	14
		Drop-like	19
		Coarse	31
		Diluted	28
		Reticular	2
		Stripped	7
		Faint	45
Special forms	Nominal	No	28
		Chalices	43
		Vesicles	77
Dislocation of	Nominal	No	50
		Yes	98
Exclusion of node	Nominal	No	31
		Yes	117
Lymph nodes diminish	Numeric	1.0–1.2	142
		1.8–2.0	3
		2.8–3.0	3
Lymph nodes enlarge	Numeric	1.0–1.6	13
		1.6–2.2	72
		2.8–3.4	43
		3.4–4.0	20
Number of nodes	Numeric	1.0–2.2	94
		2.2–3.3	18
		3.3–4.5	10
		4.5–5.7	8
		5.7–6.8	8
		6.8–8.0	10

3 Feature Generation, Selection, and Classification

The LSA method generates an effective set of features by combining the original attributes. The RAS selects an optimal subset of features from the LSA generated set. Subsequently, the selected optimal subset of features results in the improved recognition accuracy of MLP, simple linear logistic regression (SL), and sequential minimal optimization (SMO), functional tree (FT), and logistic model tree (LMT) classification methods. Fig. 1 presents a schematic diagram of the analysis and validation procedures. A PC (64-bit Windows 10, Intel(R) Core(TM) i5-4590 CPU@3.30 GHz, 8 GB RAM) was used in the implementation of attribute selection, feature generation and selection, classification methods, and their combination in WEKA [33]. A short description of attribute selection, feature generation, and selection, classification methods are as follows.

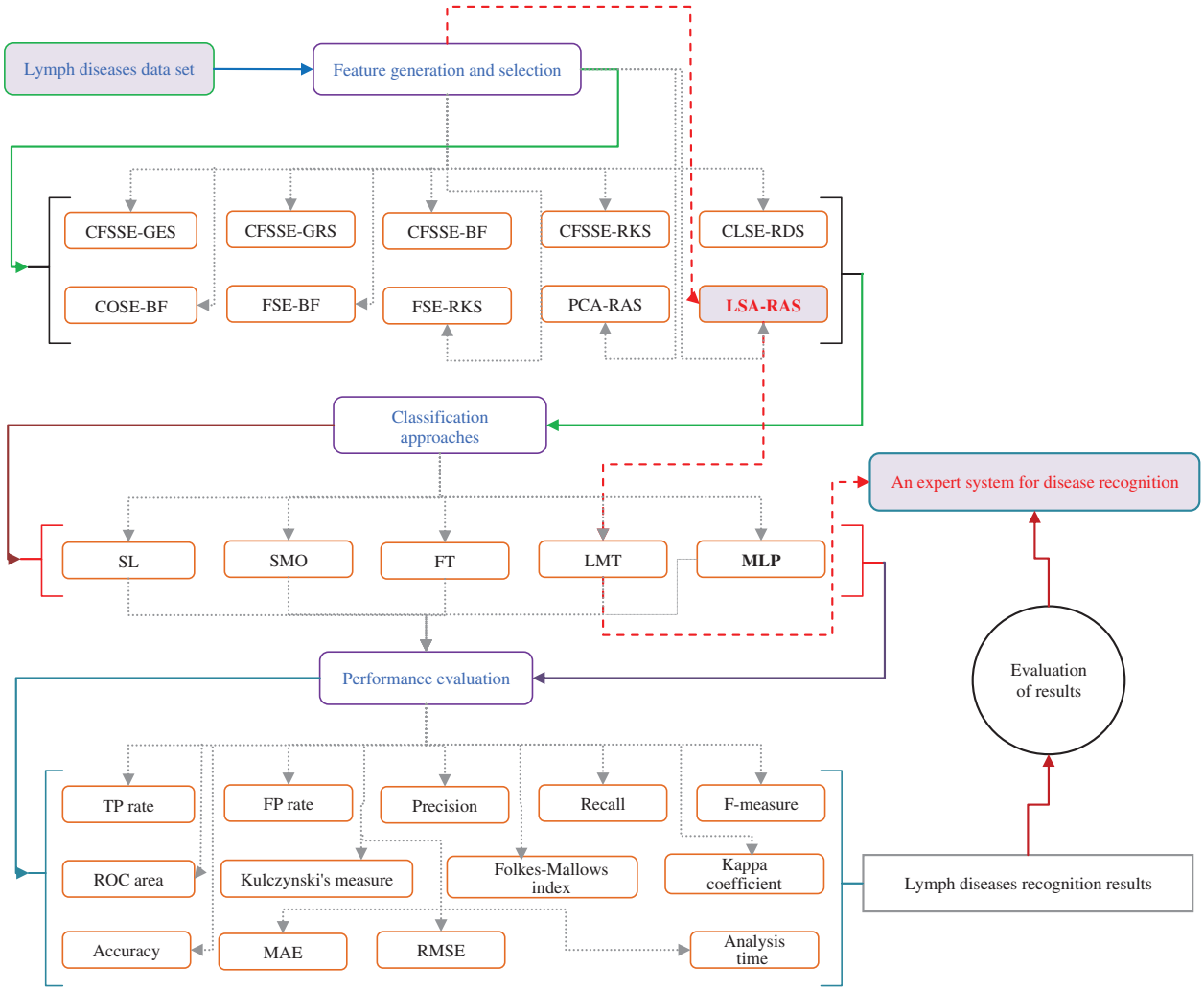

Figure 1: An overview of attribute selection, feature generation and selection, classification, and performance evaluation methods of the lymph disease

3.1 The Proposed LSA-RAS Approach of Feature Generation and Selection

The LSA measures the textual coherence of the nominal attributes. It is suitable for selecting an optimal subset of features, discarding the inappropriate features, and representation of instances in a novel semantic space for better discrimination, etc. [34]. The details of the LSA method are available in [34]. Firstly, the frequency ($\mathbf{T}_{m,n} = a$) of each term (m) in the documents (n) of the original dataset is used to calculate a term-document matrix $\mathbf{T}_{i\times j}$. Subsequently, $\mathbf{T}_{i\times j}$ is normalized as $\mathbf{T}_{i\times j} \rightarrow (\log(\mathbf{T}_{i\times j})/\text{entropy of attribute})$ and analyzed with the singular value decomposition as $\mathbf{T} = \mathbf{U}_{i\times r} \times \mathbf{S}_{r\times r} \times \mathbf{V}^{\mathrm{T}}_{j\times r}$. Finally, some largest singular values are selected in the approximation of the $\mathbf{T}_{i\times j}$ as $\overline{\mathbf{T}} = \mathbf{U}_{i\times k} \times \mathbf{S}_{k\times k} \times \mathbf{V}^{\mathrm{T}}_{j\times k}$. The RAS approach implements the combination of entropy, gain ratio, and reliefF methods [33] in the selection of an optimal subset of the latent variables. Entropy is defined as $Info(S) = -\sum_{i=1}^{m} p_i(\log_2(p_i))$ (the essential information for identification of an instance) [35]. The gain ratio is the ratio of $Gain_A = -\sum_{i=1}^{m} p_i(\log_2(p_i)) - \sum_{j=1}^{v} (|S_j|/|S|) \times Info(S_j)$ and splitting information $Split\inf_A(S) = -\sum_{j=1}^{v} (|S_j|/|S|) \times \log_2|S_j|/|S|$ [35]. The Manhattan norm of the nearest hit, and the nearest miss was used in the reliefF method to update the initial weights of features in their selection [35].

3.2 Functions and Tree-Based Classification Approaches

Three functions-based classifiers (MLP, SL, and SMO) and two tree-based classifiers (FT, and LMT) have been implemented to test the efficiency of the LSA-RAS selected feature subset and other feature subsets in a 10-fold cross-validation.

3.2.1 MLP Classifier

It is a systematic arrangement of artificial neurons in different layers (input, hidden, and output). The input of a neuron is defined as $Y = \sum W_n X_n + b$ by using the weights W_n and bias b of attributes X_n [36]. In the present study, the sigmoid activation function $O = 1/1 + \exp(-Y) = 1/1 + \exp\left(-\left(\sum W_n X_n + b\right)\right)$ was used to compute the output of neurons in the hidden layer. The linear activation function was used to calculate the output of neurons in the output layer. The MLP uses a feed-forward back-propagation strategy to update the weights and bias of each of the neurons till the error is minimized. The weight is updated using the error gradient (δ_j) and learning rate (η) in delta rule as $w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p)$, where $\Delta w_{ij}(p) = \eta\delta_j x_{ji}$, and $\delta_j = (t_j - o_j) \times o_j \times (1 - o_j)$. Using a momentum term (α), the weight update (Δw_{ij}) is defined as $\Delta w_{ij}(p) = \eta\delta_j x_{ji} + \alpha\Delta w_{ij}(p-1)$. The error $(t_j - o_j)$ decreases with the number of training epochs. The optimal MLP classifier was built using $\eta = 0.3$, $\alpha = 0.2$, training epoch $= 500$, and hidden layers $= (attributes + classes/2)$, etc. [33]. Moreover, the decaying value of η (dividing the η with the current epoch number) was used to limit the divergence from the target. Normalized values of the attributes and the classes (-1 to $+1$) are used in the training and validation. A nominal to binary filter was used for nominal attributes [33].

3.2.2 SL Classifier

It implements regression function and boosting algorithm (LogitBoost) [37] in class recognition of an instance. The LogitBoost algorithm starts with the weight initialization as $w_{ij} = 1/n$, where $i = 1, 2, 3 \ldots n$ and j represents the number of classes. Thereafter, the working response

(z_{ij}) and weights (w_{ij}) are updated on each iteration as $z_{ij} = (y^*_{ij} - p_j(x_i))/(p_j(x_i)(1-p_j(x_i)))$, and $w_{ij} = p_j(x_i)(1-p_j(x_i))$, respectively. Next, weighted least square regression z_{ij} and w_{ij} are used in $f_{mj}(x)$. The value of $F_j(x) = \sum_{m=1}^{M} f_{mj}(x)$ and $f_{mj}(x)$, are set as $F_j(x) \leftarrow F_j(x) + f_{mj}(x)$, and $f_{mj}(x) \leftarrow ((j-1)/j)\left(f_{mj}(x) - (1/j)\sum_{k=1}^{J} f_{mk}(x)\right)$. Finally, the class probability $p_j(x_i)$ is updated as $p_j(x_i) = e^{F_j(x)}/\sum_{k=1}^{J} e^{F_k(x)}$ to compute the classifier argmax $F_j(x)$. The SL classifier was implemented using the following parameters: heuristic stop = 50, maximum boosting iterations = 500, and $\beta = 0$ (value of weight trimming), etc. [33]. The heuristic stop criterion was used to terminate the iteration of LogitBoost after achieving the error minima to reduce the analysis time of the lymph dataset. Due to the variations among the attributes in the lymph dataset, the weight trimming was not implemented.

3.2.3 SMO

SMO [36] is used in the training of the support vector machine (SVM) classifier. The decision function of binary SVM is defined as $f(x) = w^T x + b$, where the class $y = +1$ for $f(x) \geq 0$, and $y = -1$ for $f(x) < 0$. Considering the inner product of the input vectors, the earlier decision function is defined as $f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^i, x \rangle + b$. For computational simplicity, the kernel function can be used for the inner product of the vectors as $f(x) = \sum_{i=1}^{m} \alpha_i y_i k(x_i, x_j) + b$. SMO is used to obtain the solution of the dual problem of SVM $\max\ W(\alpha) = \sum_{i=1}^{m} \alpha_i - \sum_{i-1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle / 2$, under constraint of $0 \leq \alpha_i \leq C$ for $i = 1, 2, 3 \ldots m$ and $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$. The optimal value of α_i, α_j and threshold b are obtained as $\alpha_j = H$ if $\alpha_j >$ H, α_j if L $\leq \alpha_j \leq H$, and L if $\alpha_j <$ L. The SMO classifier was built using the polynomial kernel function ($k(x_i, x_i) = (x_i.x_j + c)^d$), tolerance parameter equal to 0.01, complexity parameter equal to 1, and $\epsilon = 1 \times 10^{-12}$, etc. [33]. The nominal attributes were converted to binary form and normalized before training and validation of SVM.

3.2.4 FT Classifier

FT [38] uses logistic regression functions at the inner nodes and/or leaves and a constructor function (generalized linear model (GLM)) to build the decision tree. GLM combines the original attributes to generate the novel attributes. Firstly, the constructor function is used to build the initial model. In the second step, the model is mapped to new attributes of dimension equal to the number of classes in the dataset. The new attributes represent the class belonging probability of an instance computed by using the constructor function. A merit function is used to evaluate the attribute with the original attributes. The FT was built using the following parameters: boosting iterations equal to 150, number of instances equal to 15 for the splitting of nodes, and $\beta = 0$, etc. [33].

3.2.5 LMT Classifier

LMT is a combination of linear logistic regression (low variance high bias) and tree induction (high variance low bias) classification methods [39]. Logistic regression functions are generated at every node of the tree using the LogitBoost algorithm. An information gain criterion was used for the splitting of the tree, and after the complete formation of the tree, the CART algorithm was used for its pruning. The heuristic cross-validation was used to control the number of iterations of LogitBoost to avoid data overfitting. The additive logistic regression of the LogitBoost algorithm for each class M_i is defined as $L_M(x) = \sum_{i=1}^{n} \beta_i x_i + \beta_0$. The posterior probability of leaf node is defined as $P(M/x) = \exp(L_M(x))/\sum_{M=1'}^{D} \exp(L_{M'}(x))$. The optimal performance of LMT was achieved in a number of instances equal to 15 of the splitting of nodes, boosting iterations equal to 150, and $\beta = 0$ (weight trimming value), etc. [33].

3.3 Additional Attributes and Feature Selection Methods

Nine feature selection methods (Tab. 3) have been used in the performance comparison analysis. The correlation-based feature selection genetic search (CFS-GES) method combines two approaches, correlation-based feature selection (CFS) and genetic search (GES). The CFS implements a correlation measure to select the feature which is highly correlated with the class and less correlated with the other features.

Table 3: Attribute and feature selection methods, and selected attributes and features

Attribute selection and feature extraction and selection method	Abbreviations used	Number of selected attributes/features out of 18	Selected attributes/features
Correlation based feature selection (CFS) genetic search	CFSSE-GES	10	lymphatics, block_of_afferent, regeneration_of, early_uptake_in, lym_nodes_diminish, changes_in_lymph, changes_in_node, special_forms, dislocation_of, no_of_nodes_in
CFS-greedy stepwise	CFSSE-GRS	10	lymphatics, block_of_afferent, regeneration_of, early_uptake_in, lym_nodes_diminish, lym_nodes_enlarge, changes_in_lymph, changes_in_node, special_forms, no_of_nodes_in
CFS-best first	CFSSE-BF	10	lymphatics, block_of_afferent, regeneration_of early_uptake_in, lym_nodes_diminish, lym_nodes_enlarge, changes_in_lymph, changes_in_node, special_forms, no_of_nodes_in
CFS-rank search	CFSSE-RKS	11	lymphatics, block_of_afferent, block_of_lymph_s, regeneration_of, early_uptake_in, lym_nodes_diminish, lym_nodes_enlarge, changes_in_lymph, changes_in_node special_forms, no_of_nodes_in

(Continued)

Table 3 (continued)

Attribute selection and feature extraction and selection method	Abbreviations used	Number of selected attributes/features out of 18	Selected attributes/features
Classifier subset evaluation random search	CLSE-RDS	9	block_of_afferent, block_of_lymph_c, block_of_lymph_s, extravasates, changes_in_lym defect_in_node, changes_in_node, special_forms, no_of_nodes_in
Consistency subset evaluation best first	COSE-BF	9	lymphatics, block_of_afferent, block_of_lymph_c, changes_in_lymph, defect_in_node, changes_in_node, changes_in_stru, special_forms exclusion_of_no
Filtered subset evaluation best first	FSE-BF	10	lymphatics, block_of_afferent, regeneration_of early_uptake_in, lym_nodes_diminish, lym_nodes_enlarge, changes_in_lymph, changes_in_node, special_forms, no_of_nodes_in
Filtered subset evaluation rank search	FSE-RKS	11	lymphatics, block_of_afferent, block_of_lymph_s regeneration_of, early_uptake_in, lym_nodes_diminish, lym_nodes_enlarge, changes_in_lymph, changes_in_node, special_forms, no_of_nodes_in
Principal component analysis ranker search	PCA-RAS	25	Principal components (PC1-PC25)
Latent semantic analysis ranker search	**LSA-RAS**	**13**	**Latent variables (LV1-LV13)**

On the basis of the earlier correlation values, a merit measure M_s of feature subset is defined as $M_s = kr_{cf}/\sqrt{k+(k+1)r_{ff}}$ (r_{cf} and r_{ff} denotes an average feature-class and feature-feature correlation) [33]. The GES method basically implements a simple genetic algorithm in searching for an optimal set of attributes. Selection, crossover, and mutation operators have been used in GA to adopt the process of evolution of nature [36]. The GES method was built using the initial population of 20 features, cross-over probability equal to 0.6, the number of generations equal to 20, and mutation probability equal to 0.033 [33]. The GES method was used for the ranking of the attributes in combination with CFS (Tab. 3). The attributes are selected either in the forward or backward direction using the greedy stepwise (GRS) method. The best subset of the attribute is selected by including attributes step by step till the merit of the feature subset increases. A forward selection approach is implemented in the selection of an optimal subset of attributes. The GRS method was used in combination with the CFS method in attribute ranking and selection of an optimal subset of attributes (Tab. 3).

Redundant attributes are discarded using a threshold value of −1.80 [33]. A hill-climbing approach with a backtracking search approach was used in the selection of optimal attributes in the best first (BF) method. A forward search approach and search termination threshold value equal to 5 was used in the BF method in the present analysis [33]. BF method was used

in combination with CFS, consistency subset evaluation, and filtered subset evaluation methods (Tab. 3). The rank search (RKS) approach uses a forward selection search method to generate an optimal subset of the attribute of maximum merit. The attributes are included one by one with the best attribute in each step to generate an optimal subset of attributes. Attribute evaluator (gain ratio with starting point equal to 0 and the step size equal to 1) was used to evaluate the attribute subset in each step after including an attribute until the merit of the attribute subset increases [21]. The RKS method was combined with the CFS and filtered subset evaluation (FSE) methods for the attribute ranking and selection (Tab. 3). Classifier subset evaluation (CLSE) implements a classification method in the selection of an optimal subset of attributes. The CLSE uses a ZeroR classification approach to compute the merit of the feature subset. For the numeric class, ZeroR predicts the mean of the numeric class and mode for the nominal class. This concept is used to compute the merit of an attribute subset in CLSE [33]. The random search (RDS) uses a random search approach to select an optimal subset of attributes. The RDS selects a random subset of attributes in finding the optimal subset. Another parameter used in the RDS method was a seed equal to 1 to generate a random number, and 25% of the search space [33]. The RDS was used in combination with the CLSE (Tab. 3). Consistency subset evaluation (COSE) selects an optimal subset of attributes on the basis of its level of consistency in class. The consistency (C_s) of a subset of the attribute is defined as $C_s = 1 - \left\{\sum_{i=1}^{J} |D_i| - |M_i|\right\}/N$ (N represents the number of instances in the dataset, s denotes the attribute subset, J stand for different combinations of attributes, and $|D_i|$ and $|M_i|$ denote the frequency and the cardinality of the majority class of i^{th} attribute value combination, respectively) [21]. The COSE was used in combination with the BF method (Tab. 3). Filtered subset evaluation (FSE) is a combined approach of CFS and a random subsample filter used in the selection of an optimal subset of attributes. Basically, an initial subset of features is selected casually by the random subsample filter and used as the input of the CFS method. A spread value is always defined in the FSE to control the effect of least and most recurrent classes [40,41]. The FSE was used in combination with the BF and RKS method to select an optimal subset of features (Tab. 3). The principal component analysis (PCA) method doesn't select the attributes directly; nonetheless, it first transforms the original attributes in a novel principal component (PC) space and then selects a significant subset of the attributes using some ranking method. Basically, the original attributes are projected along the PC directions to obtain the novel subset of features. The PC component matrix $\mathbf{PC}_{mxk}$ of an original data matrix $\mathbf{O}_{mxk}$ (m instances and n attributes) is achieved as $\mathbf{O}_{mxk} = \mathbf{PC}_{mxk}\mathbf{L}_{mxk}^{T} + \mathbf{R}_{mxn}$ [36]. The loading matrix $\mathbf{L}_{mxk}^{T}$ denotes the significance of the attributes in the formation of the PC components. The RAS method is combined with the PCA to select an ideal subset of PC components [33].

3.4 Performance Assessment Measures of Classification Approaches

Performance of classification approaches is evaluated on the basis of the average value of true positive (TP) rate, false positive (FP) rate, precision, recall, Kulczynski's measure (arithmetic mean of precision and recall), Folkes-Mallows index (geometrical mean of precision and recall), F-measure (harmonic mean of precision and recall), Kappa coefficient, receiver operating characteristic (ROC) area, and analysis time. The Kappa coefficient is computed using the number of instances in a row ($x_{i.}$), column (x_j), and diagonal (x_{ii}) of the confusion matrix of the classification method and the total number of instances in the dataset (N) as $k = \left(N\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i.}x_{.i}\right)/\left(N^2 - \sum_{i=1}^{k} x_{i.}x_{.i}\right)$ [40]. The TP, FP, false negative (FN), and true negative (TN) are used to compute the ROC area as $ROC_{area} = 1 + (TP/TP + FN) - (FP/FP + TN)/2$ [18].

4 Validation Results of the Proposed Approach and Comparative Analysis Results

Tab. 3 summarizes the attributes and features selected by the different approaches. It is obvious that the CFSSE-RKS method selects a maximum number of attributes (11 out of 18) and PCA-RAS generates a maximum number of features. CFSSE-GES, CFSSE-GRS, and CFSSE-BF methods select a similar number of attributes (Tab. 3). The different attributes subsets are selected by the CFSSE-GES, and CFSSE-GRS methods, while the attribute subset selected by the CFSSE-GRS and CFSSE-BF is the same. It is also noticeable that the PCA-RAS and LSA-RAS methods select the optimal subset of features considering the contributions of all attributes, while the rest of the methods in Tab. 3 select an optimal subset of attributes. The parametric details of the attributes/feature selection methods and the merits/ranking of the selected subset of the attributes/features are summarized in Tab. 4. The performance of classification methods using selected subsets of attributes/features is summarized in Tab. 5. Tab. 6 presents the performance evaluation metrics of classification methods.

Table 4: Merits and ranking values of selected attributes and features

Approach	Parameters	Merits/ranking values, and rank		
		Merit value	Selected attributes subset	
CFSSE-GES	Population size: 20 Number of generations: 20 Probability of crossover: 0.6 Probability of mutation: 0.033 Report frequency: 20 Random number seed: 1	0.4132 0.4132 0.3909 0.3948 0.4047 0.4037 0.4132 0.4083 0.4132 0.4044 0.4232	1, 2, 7, 8, 9, 11, 13, 15, 18 1, 2, 7, 8, 9, 11, 13, 15, 18 1, 2, 7, 8, 9, 11, 13, 15 1, 2, 5, 8, 9, 11, 13, 15, 18 2, 7, 8, 9, 13, 15, 18 1, 2, 7, 9, 11, 13, 15, 18 1, 2, 7, 8, 9, 11, 13, 15, 18 1, 2, 7, 8, 9, 11, 13, 18 1, 2, 7, 8, 9, 11, 13, 15, 18 1, 2, 5, 7, 8, 9, 11, 13, 15, 18 1, 2, 7, 8, 9, 11, 13, 15, 16, 18	
		Merit of best subset (first ten attributes): 0.414		
		Ranking value	Attribute number	Attribute
CFSSE-GRS	Greedy Stepwise (forwards). Start set: no attributes Threshold: −1.79	0.281 0.331 0.364 0.385 0.390 0.399 0.406 0.412 0.413 0.414	13 18 9 2 8 7 1 15 11 10	changes_in_node no_of_nodes_in lym_nodes_diminish block_of_afferent early_uptake_in regeneration_of lymphatics special_forms changes_in_lymph lym_nodes_enlarge

(Continued)

Table 4 (continued)

Approach	Parameters	Merits/ranking values, and rank			
CFSSE-BF	Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 172	Merit of best subset (1, 2, 7, 8, 9, 10, 11, 13, 15, 18): 0.414			
CFSSE-RKS	Start set: no attributes Step size: 1	Merit of best subset (1, 2, 4, 7, 8, 9, 10, 11, 13, 15, 18): 0.4			
CLSE-RDS	Start set: no attributes Number of iterations: 65536 (25.0% of the search space)	Merit of best subset (2, 3, 4, 6, 11, 12, 13, 15, 18): 0.453			
COSE-BF	Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 166	Merit of best subset (1, 2, 3, 11, 12, 13, 14, 15, 17): 1			
FSE-BF	Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 172	Merit of best subset (1, 2, 7, 8, 9, 10, 11, 13, 15, 18): 0.414			
FSE-RKS	Attribute evaluator: gain ratio Start point: no attribute Step size: 1	Merit of best subset (1, 2, 4, 7, 8, 9, 10, 11, 13, 15, 18): 0.4			
		Ranking value	Principal component	Eigen value	Cumulative variance
PCA-RAS	Threshold: −1.79	0.87	PC1	4.79	0.13
		0.77	PC2	3.79	0.23
		0.69	PC3	3.27	0.31
		0.62	PC4	2.54	0.38
		0.57–0.1	PC5–25	2.03–0.41	0.43–0.95

(Continued)

Table 4 (continued)

Approach	Parameters	Merits/ranking values, and rank		
		Ranking value	**Latent variable**	**Singular value**
LSA-RAS	**Start set: no attributes**	**0.76114**	**LV1**	**58.4074**
	Threshold: −1.79	**0.05955**	**LV2**	**16.33678**
		0.02665	**LV3**	**10.9288**
		0.01914	**LV4**	**9.26279**
		0.02–0.01	**LV5-13**	**7.96–5.46**

Notes: attribute no. 1: lymphatics, 2: block_of_afferent, 3: block_of_lymph_c, 4: block_of_lymph_s, 5: by_pass, 6: extrava sates, 7:regeneration_of, 8: early_uptake_in, 9: lym_nodes_diminish, 10: lym_nodes_enlar, 11: changes_in_lym, 12: d efect_in_node, 13:changes_in_node, 14: changes_in_stru, 15: special_forms, 16: dislocation_of, and 17: exclusion_of_no, 18: no_of_nodes_in.

Table 5: Performance of the classification methods using different features subsets in the lymph disease recognition

Method	Results	Original attributes	Selected attributes, and generated and selected features subsets									
			CFSSE-GES	CFSSE-GRS	CFSSE-BF	CFSSE-RKS	CLSE-RDS	COSE-BF	FSE-BF	FSE-RKS	PCA-RAS	LSA-RAS
MLP	CC	84.46	81.08	81.76	81.76	81.76	82.43	81.76	84.46	81.76	85.81	93.24
	MAE	0.08	0.10	0.09	0.09	0.10	0.09	0.10	0.08	0.10	0.07	0.04
	RMSE	0.26	0.28	0.27	0.27	0.27	0.26	0.28	0.26	0.27	0.24	0.15
	k	0.70	0.64	0.65	0.65	0.64	0.66	0.65	0.70	0.64	0.72	0.87
SL	CC	83.11	80.41	83.78	83.78	82.43	81.76	83.78	83.11	82.43	89.86	92.57
	MAE	0.10	0.12	0.10	0.10	0.10	0.11	0.11	0.10	0.10	0.08	0.05
	RMSE	0.27	0.27	0.25	0.25	0.25	0.27	0.25	0.27	0.25	0.20	0.18
	k	0.68	0.63	0.69	0.69	0.66	0.64	0.69	0.68	0.66	0.81	0.86
SMO	CC	86.49	83.11	85.81	85.81	85.81	81.76	85.14	86.49	85.81	86.49	92.57
	MAE	0.26	0.26	0.26	0.26	0.26	0.27	0.26	0.26	0.26	0.26	0.26
	RMSE	0.33	0.33	0.33	0.33	0.33	0.34	0.33	0.33	0.33	0.33	0.32
	k	0.74	0.68	0.73	0.73	0.73	0.64	0.71	0.74	0.73	0.74	0.86
FT	CC	86.49	86.49	85.14	85.14	84.46	77.70	85.14	86.49	84.46	84.46	93.24
	MAE	0.08	0.08	0.08	0.08	0.09	0.12	0.08	0.08	0.09	0.08	0.03
	RMSE	0.26	0.26	0.26	0.26	0.26	0.30	0.26	0.26	0.26	0.28	0.18
	k	0.74	0.74	0.72	0.72	0.70	0.57	0.72	0.74	0.70	0.70	0.87
LMT	**CC**	**83.11**	**80.41**	**83.78**	**83.78**	**82.43**	**81.76**	**83.78**	**83.11**	**82.43**	**89.86**	**93.92**
	MAE	**0.10**	**0.12**	**0.10**	**0.10**	**0.10**	**0.11**	**0.11**	**0.10**	**0.10**	**0.08**	**0.03**
	RMSE	**0.27**	**0.27**	**0.25**	**0.25**	**0.25**	**0.27**	**0.25**	**0.27**	**0.25**	**0.20**	**0.17**
	k	**0.68**	**0.63**	**0.69**	**0.69**	**0.66**	**0.64**	**0.69**	**0.68**	**0.66**	**0.81**	**0.89**

Note: CC-correct classification rate, MAE-mean absolute error, RMSE-root means square error, and k-Kappa coefficient.

The LSA-RAS generated feature subset results in the maximum accuracy of classification methods (Tab. 5). Among the three functions-based classification methods, the maximum classification accuracy (93.24%) and the minimum value of mean absolute error (MAE) (0.04) have been achieved for the MLP. The LMT classifier achieved higher accuracy (93.92%) and the maximum value of the kappa coefficient (0.89) than the FT method.

Table 6: Evaluation metrics of classification methods using selected attribute and features subsets in the lymph disease recognition

Classification method	Attributes/features subsets	TP rate	FP rate	Precision	Recall	F-measure	ROC area	Kulczynski's measure	Folkes-Mallows index
MLP	Original attributes	0.845	0.157	0.837	0.845	0.834	0.920	0.841	0.841
	CFSSE-GES	0.811	0.174	0.805	0.811	0.805	0.912	0.808	0.808
	CFSSE-GRS	0.818	0.173	0.811	0.818	0.808	0.915	0.815	0.814
	CFSSE-BF	0.818	0.173	0.811	0.818	0.808	0.915	0.815	0.814
	CFSSE-RKS	0.818	0.179	0.810	0.818	0.807	0.905	0.814	0.814
	CLSE-RDS	0.824	0.170	0.796	0.824	0.809	0.916	0.810	0.810
	COSE-BF	0.818	0.175	0.816	0.818	0.810	0.892	0.817	0.817
	FSE-BF	0.845	0.157	0.837	0.845	0.834	0.920	0.841	0.841
	FSE-RKS	0.818	0.179	0.810	0.818	0.807	0.905	0.814	0.814
	PCA-RAS	0.858	0.136	0.842	0.858	0.848	0.948	0.850	0.850
	LSA-RAS	**0.932**	**0.065**	**0.895**	**0.932**	**0.913**	**0.993**	**0.914**	**0.913**
SL	Original attributes	0.831	0.163	0.832	0.831	0.831	0.893	0.832	0.831
	CFSSE-GES	0.804	0.189	0.805	0.804	0.804	0.885	0.805	0.804
	CFSSE-GRS	0.838	0.162	0.839	0.838	0.837	0.914	0.839	0.838
	CFSSE-BF	0.838	0.162	0.839	0.838	0.837	0.914	0.839	0.838
	CFSSE-RKS	0.824	0.175	0.825	0.824	0.823	0.912	0.825	0.824
	CLSE-RDS	0.818	0.178	0.816	0.818	0.809	0.892	0.817	0.817
	COSE-BF	0.838	0.154	0.836	0.838	0.837	0.905	0.837	0.837
	FSE-BF	0.831	0.163	0.832	0.831	0.831	0.893	0.832	0.831
	FSE-RKS	0.824	0.175	0.825	0.824	0.823	0.912	0.825	0.824
	PCA-RAS	0.899	0.108	0.904	0.899	0.898	0.950	0.902	0.901
	LSA-RAS	**0.926**	**0.057**	**0.927**	**0.926**	**0.920**	**0.970**	**0.927**	**0.926**
SMO	Original attributes	0.865	0.135	0.869	0.865	0.864	0.869	0.867	0.867
	CFSSE-GES	0.831	0.166	0.833	0.831	0.830	0.844	0.832	0.832
	CFSSE-GRS	0.858	0.148	0.861	0.858	0.856	0.869	0.860	0.859
	CFSSE-BF	0.858	0.148	0.861	0.858	0.856	0.869	0.860	0.859
	CFSSE-RKS	0.858	0.148	0.861	0.858	0.856	0.869	0.860	0.859
	CLSE-RDS	0.818	0.178	0.790	0.818	0.803	0.844	0.804	0.804
	COSE-BF	0.851	0.151	0.858	0.851	0.846	0.861	0.855	0.854
	FSE-BF	0.865	0.135	0.869	0.865	0.864	0.869	0.867	0.867
	FSE-RKS	0.858	0.148	0.861	0.858	0.856	0.869	0.860	0.859
	PCA-RAS	0.865	0.143	0.856	0.865	0.858	0.866	0.861	0.860
	LSA-RAS	**0.926**	**0.072**	**0.912**	**0.926**	**0.918**	**0.936**	**0.919**	**0.919**
FT	Original attributes	0.865	0.136	0.866	0.865	0.864	0.890	0.866	0.865
	CFSSE-GES	0.865	0.128	0.865	0.865	0.864	0.871	0.865	0.865
	CFSSE-GRS	0.851	0.137	0.851	0.851	0.851	0.886	0.851	0.851
	CFSSE-BF	0.851	0.137	0.851	0.851	0.851	0.886	0.851	0.851
	CFSSE-RKS	0.845	0.157	0.846	0.845	0.844	0.849	0.846	0.845
	CLSE-RDS	0.777	0.197	0.768	0.777	0.772	0.841	0.773	0.772
	COSE-BF	0.851	0.142	0.853	0.851	0.850	0.907	0.852	0.852
	FSE-BF	0.865	0.136	0.866	0.865	0.864	0.890	0.866	0.865
	FSE-RKS	0.845	0.157	0.846	0.845	0.844	0.849	0.846	0.845
	PCA-RAS	0.845	0.154	0.846	0.845	0.844	0.857	0.846	0.845
	LSA-RAS	**0.932**	**0.047**	**0.940**	**0.932**	**0.934**	**0.964**	**0.936**	**0.936**
LMT	Original attributes	0.831	0.163	0.832	0.831	0.831	0.893	0.832	0.831
	CFSSE-GES	0.804	0.189	0.805	0.804	0.804	0.885	0.805	0.804
	CFSSE-GRS	0.838	0.162	0.839	0.838	0.837	0.914	0.839	0.838
	CFSSE-BF	0.838	0.162	0.839	0.838	0.837	0.914	0.839	0.838
	CFSSE-RKS	0.824	0.175	0.825	0.824	0.823	0.912	0.825	0.824
	CLSE-RDS	0.818	0.178	0.816	0.818	0.809	0.892	0.817	0.817
	COSE-BF	0.838	0.154	0.836	0.838	0.837	0.905	0.837	0.837
	FSE-BF	0.831	0.163	0.832	0.831	0.831	0.893	0.832	0.831
	FSE-RKS	0.824	0.175	0.825	0.824	0.823	0.912	0.825	0.824
	PCA-RAS	0.899	0.108	0.904	0.899	0.898	0.950	0.902	0.901
	LSA-RAS	**0.939**	**0.039**	**0.946**	**0.939**	**0.940**	**0.970**	**0.943**	**0.943**

The LSA-RAS selected feature subset results in the improvement of 10.81% in the classification accuracy of the LMT classifier than the original attributes. Moreover, the accuracy of classification methods using most of the selected attribute subset except PCA-RAS and LSA-RAS selected feature subset is lower or comparable than using the original attributes. The LSA-RAS selected feature subset results in improved evaluation measures (maximum value of average the TP rate, Precision, Recall, F-measure, ROC area, Kulczynski's measure, and Folkes-Mallows index, and the minimum average value of FP rate) of each of the classification methods than other selected attribute subsets, selected feature subset, and original attributes. Furthermore, the LMT classification method using the LSA-RAS selected feature subset has the best values of earlier evaluation metrics than the rest of the classification method. A detailed class confusion matrix of each of the classification methods using the best performing LSA-RAS selected feature subset is summarized in Tab. 7. The MLP classification recognizes 138 out of 148 instances of lymph disease correctly. The maximum number of instances (80 out of 81) of metastases class is identified correctly (accuracy of 98.77%). Fig. 2 presents the error curve of the MLP method using three attribute subsets and one feature subset.

Table 7: The class confusion matrix in recognition of the lymph diseases using the LSA-RAS features subset

Total	Normal	Metastases	Malign lymph	Fibrosis	Total
MLP					
Normal	**0**	2	0	0	2
Metastases	0	**80**	1	0	81
Malign lymph	0	3	**58**	0	61
Fibrosis	0	0	4	**0**	4
Total	0	85	63	0	**93.24%**
SMO					
Normal	**0**	2	0	0	2
Metastases	0	**80**	1	0	81
Malign lymph	0	5	**55**	1	61
Fibrosis	0	0	2	**2**	4
Total	0	87	58	3	**92.57%**
SL					
Normal	**0**	2	0	0	2
Metastases	2	**77**	2	0	81
Malign lymph	0	2	**59**	0	61
Fibrosis	0	0	3	**1**	4
Total	2	81	64	1	**92.57%**
FT					
Normal	**0**	2	0	0	2
Metastases	3	**77**	1	0	81
Malign lymph	0	2	**59**	0	61
Fibrosis	0	0	2	**2**	4
Total	3	81	62	2	**93.24%**

(Continued)

Table 7 (continued)

Total	Normal	Metastases	Malign lymph	Fibrosis	Total
LMT					
Normal	**0**	2	0	0	2
Metastases	3	**77**	1	0	81
Malign lymph	0	1	**60**	0	61
Fibrosis	0	0	2	**2**	4
Total	3	79	63	2	**93.92%**

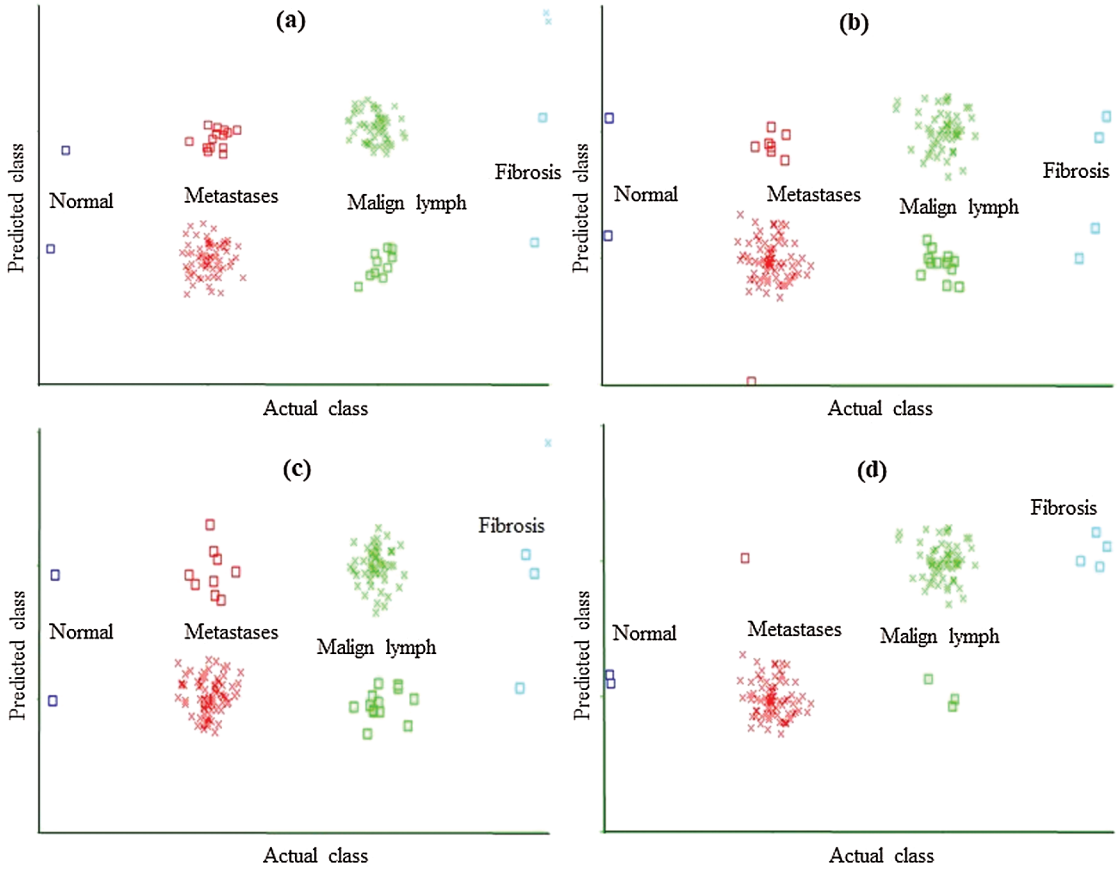

Figure 2: Classification error curve of MLP using (a) CFSSE-GES, (b) CLSE-RDS, (c) FSE-RKS, and (d) LSA-RAS selected attribute and feature subsets

The classification error in Fig. 2 is denoted by the square symbol. It is obvious that the LSA-RAS feature subset results in the minimum classification error (10) of MLP than rest three attribute subsets (Fig. 2d). It is analogous to the confusion matrix of MLP in Tab. 7. The CFSSE-GES selected attribute subset results in the maximum error of MLP (Fig. 2a). The error curve of the SL is presented in Fig. 3. The SL classification recognizes 137 out of 148 instances of

lymph disease correctly. The maximum number of instances (59 out of 61) of malign lymph class is identified correctly (accuracy of 96.72%, Tab. 7). The LSA-RAS selected feature subset results in the minimum classification error (11) of the SL than rest three attribute subsets (Fig. 3d) which is similar to the confusion matrix of SL in Tab. 7. The maximum error of SL has been obtained for the CFSSE-GES selected attribute subset (Fig. 3a). Fig. 4 presents the error curve of the SMO classification method. Like SL, the SMO classification method also recognizes 137 out of 148 instances of lymph disease correctly though there is some difference in the confusion matrix (Tab. 7). The maximum number of instances (80 out of 81) of metastases class is identified correctly (accuracy of 98.77%, Tab. 7). The LSA-RAS selected feature subset results in the minimum classification error (11) of SMO than rest three attribute subsets (Fig. 4d) (analogous to the confusion matrix of SL in Tab. 7).

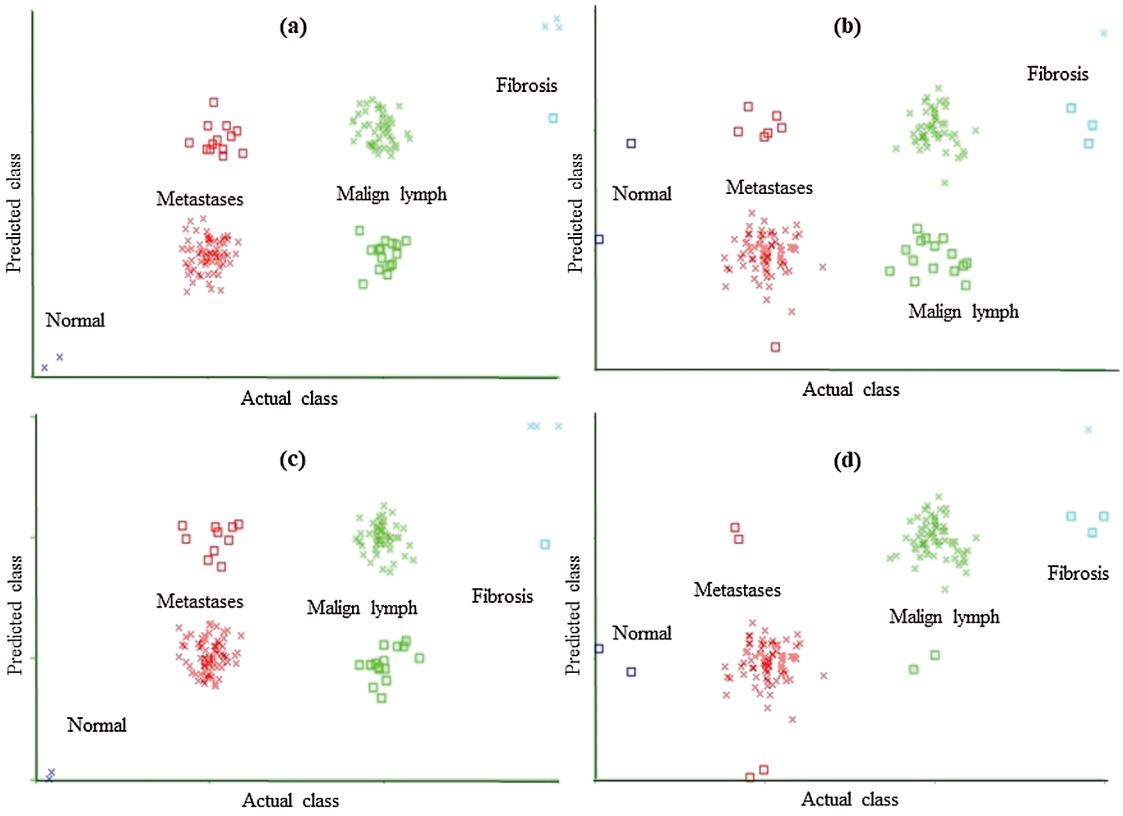

Figure 3: Classification error curve of SL using (a) CFSSE-GES, (b) CLSE-RDS, (c) FSE-RKS, and (d) LSA-RAS selected attribute and feature subsets

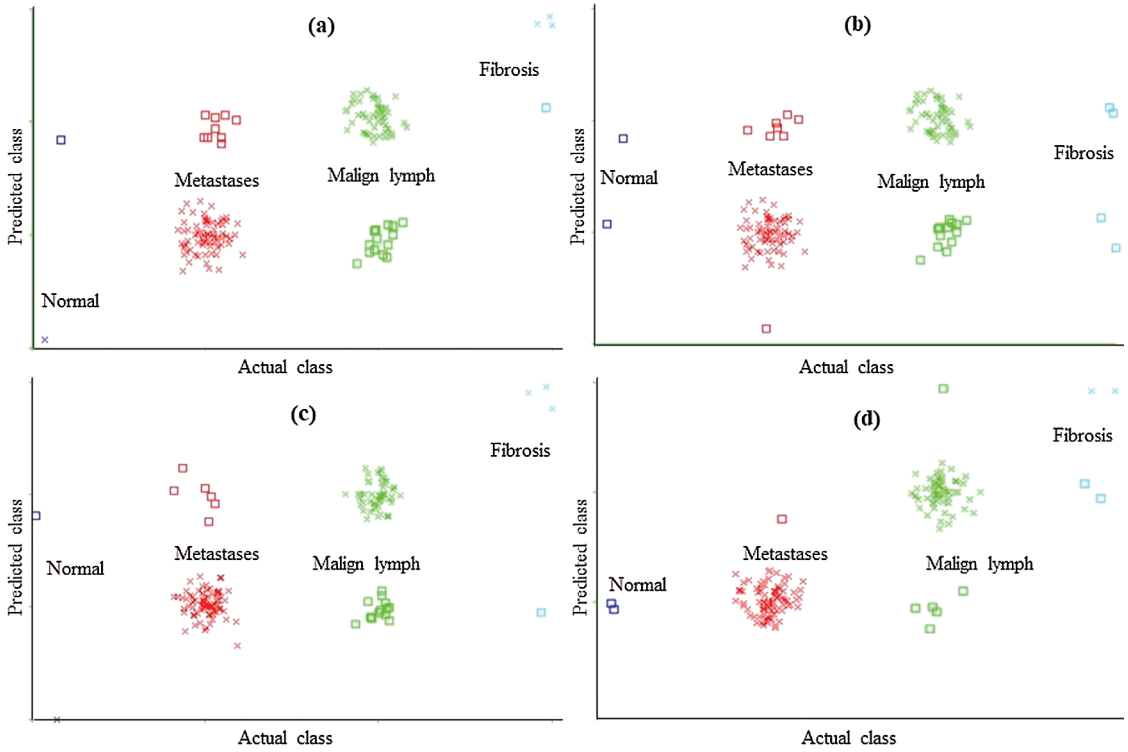

Figure 4: Classification error curve of SMO using (a) CFSSE-GES, (b) CLSE-RDS, (c) FSE-RKS, and (d) LSA-RAS selected attribute and feature subsets

The CLSE-RDS selected attribute subset results in the maximum error of the SMO (Fig. 4a). The error curve of the FT classification method using CFSSE-GES, CLSE-RDS, and FSE-RKS selected attribute subsets, and LSA-RAS selected feature subset is presented in Fig. 5. The FT classification method identifies 138 out of 148 instances of lymph disease correctly (confusion matrix in Tab. 7). The maximum number of instances (59 out of 61) of malign lymph class is identified correctly (accuracy of 96.72%). The LSA-RAS selected feature subset results in the minimum classification error (10) of the FT classifier than the rest three attributes subsets (Fig. 5d) (similar to the confusion matrix of FT in Tab. 7). CFSSE-GES and FSE-BF have the maximum and similar errors (Figs. 5a and 5c). Fig. 6 presents the error curve of the LMT classification method. The error curve in Fig. 6d represents that 139 out of 148 instances have been correctly identified by the LMT method using the LSA-RAS selected feature subset. It is analogous to the confusion matrix of the LMT method in Tab. 7. The maximum number of instances (60 out of 61) of malign lymph class is identified correctly (accuracy of 98.36%). The CFSSE-GES selected attribute subset results in the maximum error of the LMT (Fig. 6a). The LSA-RAS selected feature subset results in the improved value of the area under ROC of the classification methods than any other selected attribute subset, feature subset, and original attributes (Tab. 6). Furthermore, the maximum average area under the ROC was achieved for the LMT using the LSA-RAS selected feature subset.

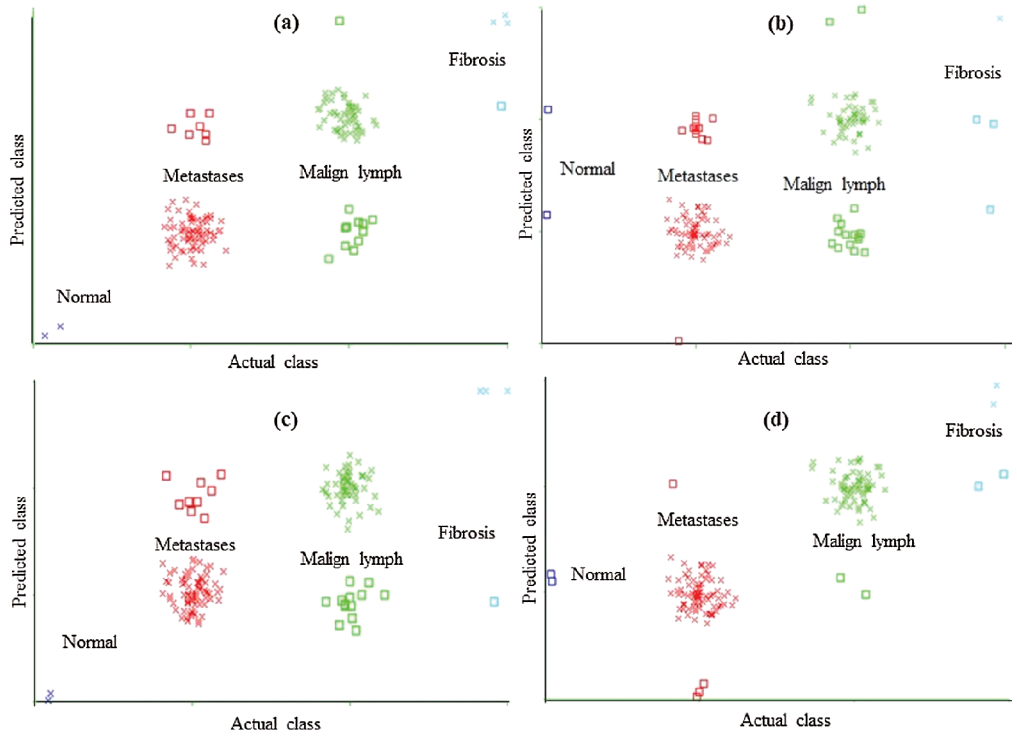

Figure 5: Classification error curve of FT using (a) CFSSE-GES, (b) CLSE-RDS, (c) FSE-RKS, and (d) LSA-RAS selected attribute and feature subsets

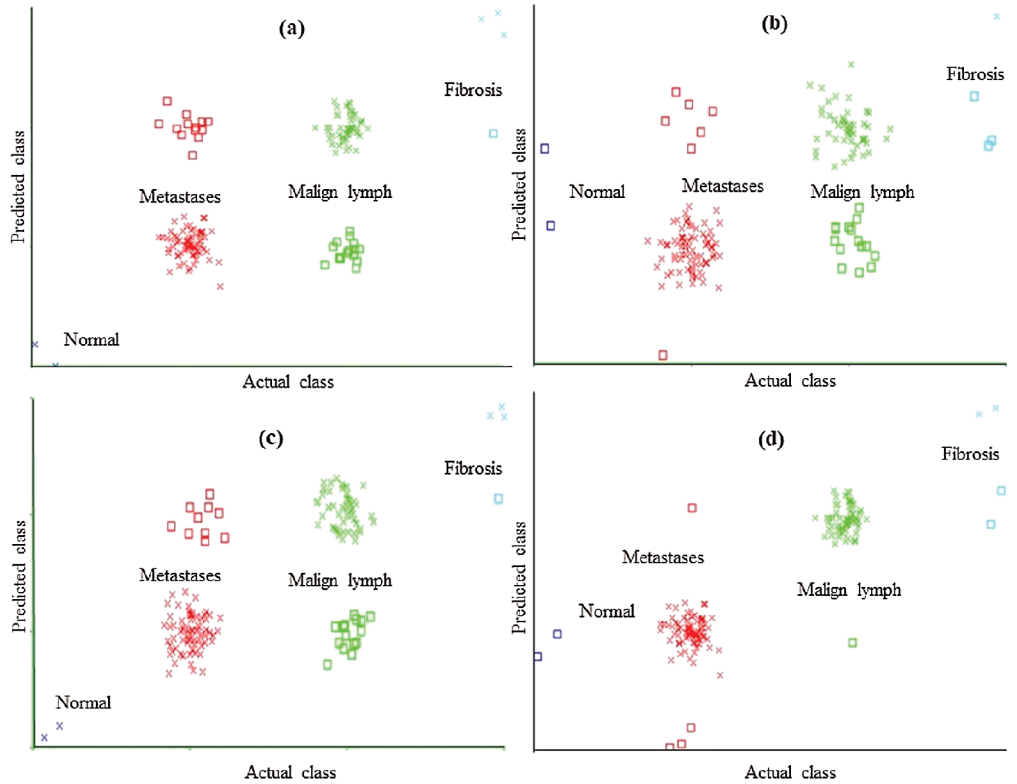

Figure 6: Classification error curve of LMT using (a) CFSSE-GES, (b) CLSE-RDS, (c) FSE-RKS, and (d) LSA-RAS selected attribute and feature subsets

The class-wise area under the ROC curve of LMT using the LSA-RAS selected feature subset is demonstrated in Fig. 7. It is obvious that the three positive classes (metastases, malign lymph class, fibrosis) of the lymph disease have an area under ROC $\geq$ 0.96. The minimum ROC area (0.43) was obtained for the normal class of the lymph disease while the fibrosis class of the lymph has the maximum ROC area (0.998). Fig. 8 represents the cost/benefit curve of LMT using the LSA-RAS selected feature subset of normal and three classes of lymph. The cost/benefit curve represents the error rate on the Y-axis and the probability of belonging to the positive class on the X-axis. The normal and fibrosis classes of lymph have higher error (100%, and 50%, respectively), consequently, the cost curve has a positive slope. The rest of the two classes metastases and malign have a lower error rate. The analysis time of each of the classification methods using the original attributes, selected attributes, and selected features are presented in Fig. 9. The MLP classification method has a maximum analysis time of 2.29 s using the FSE-BF selected attribute subset and the SMO method has a minimum analysis time of 0.01 s using the FSE RKS selected attribute subset. Moreover, the MLP has the maximum average analysis time of 0.793 s and the SMO has a minimum average analysis time of 0.037 s using original attributes and selected attributes, and selected features.

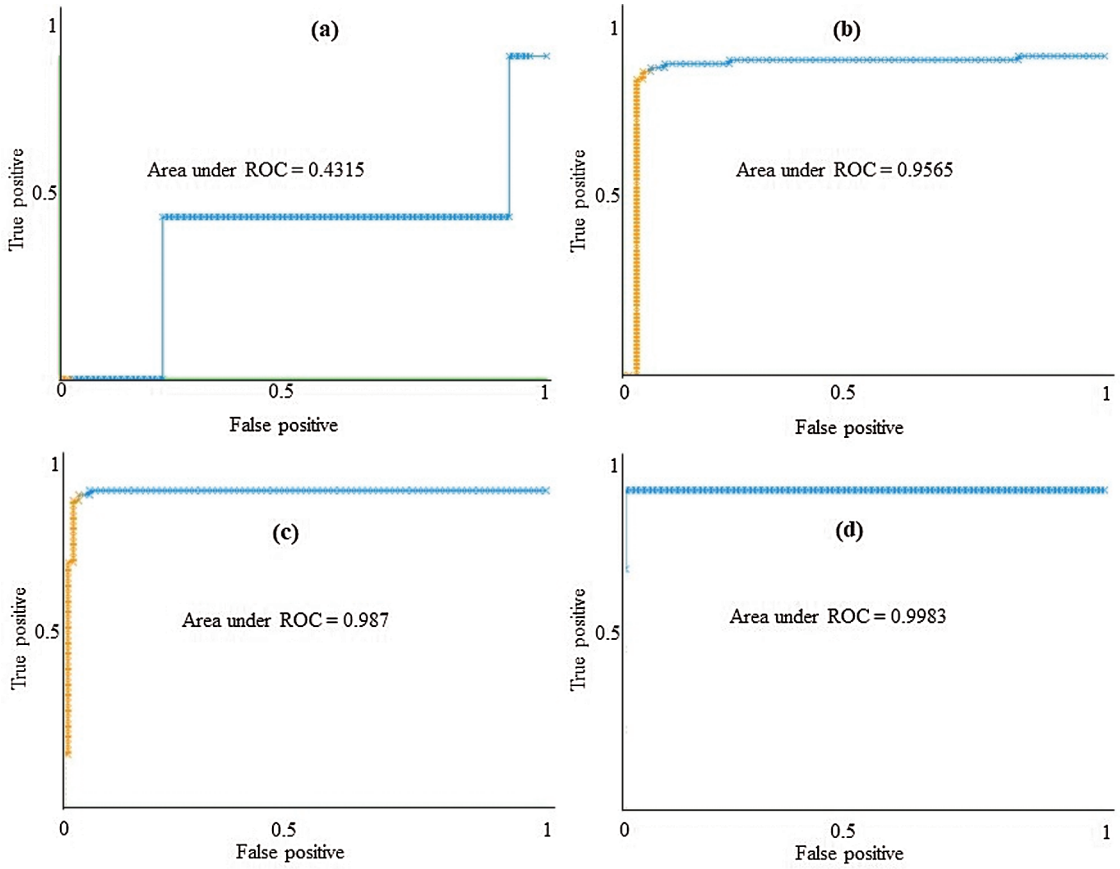

Figure 7: The area under the ROC curve of LMT using LSA-RAS selected feature subset for (a) normal, (b) metastases, (c) malign lymph, and (d) fibrosis classes of lymph diseases

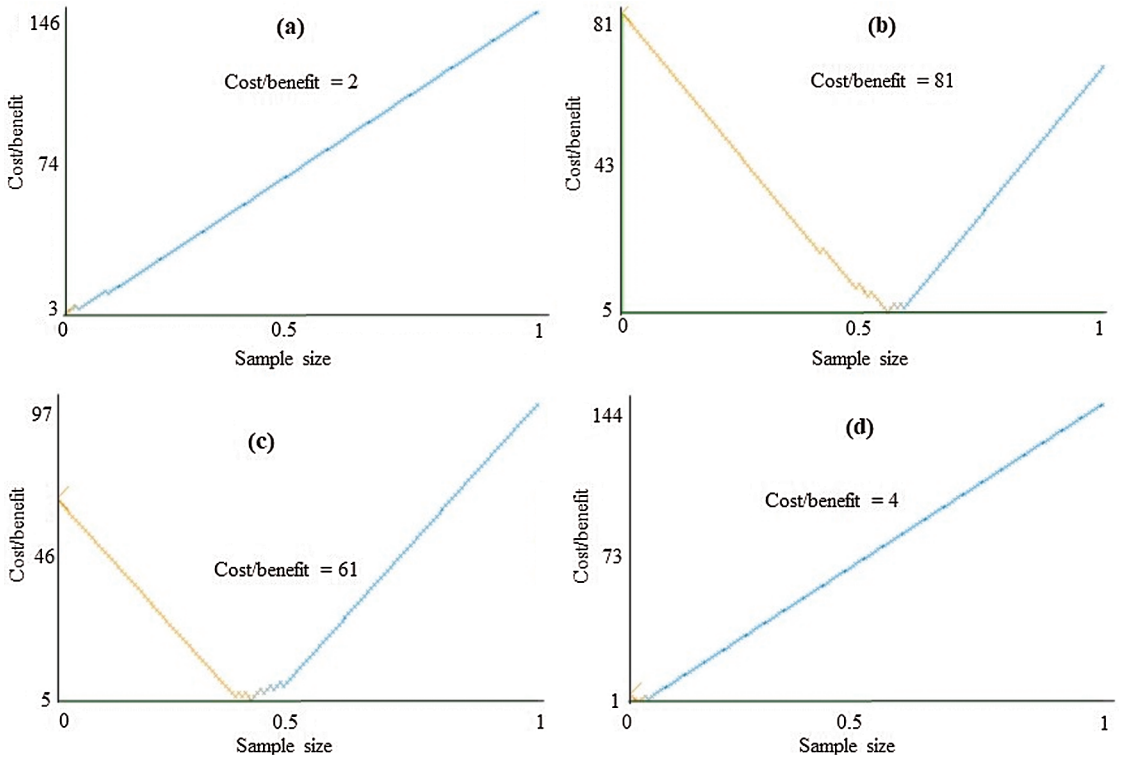

Figure 8: Cost/Benefit curve of LMT using LSA-RAS selected feature subset for (a) normal, (b) metastases, (c) malign lymph, and (d) fibrosis classes of lymph diseases

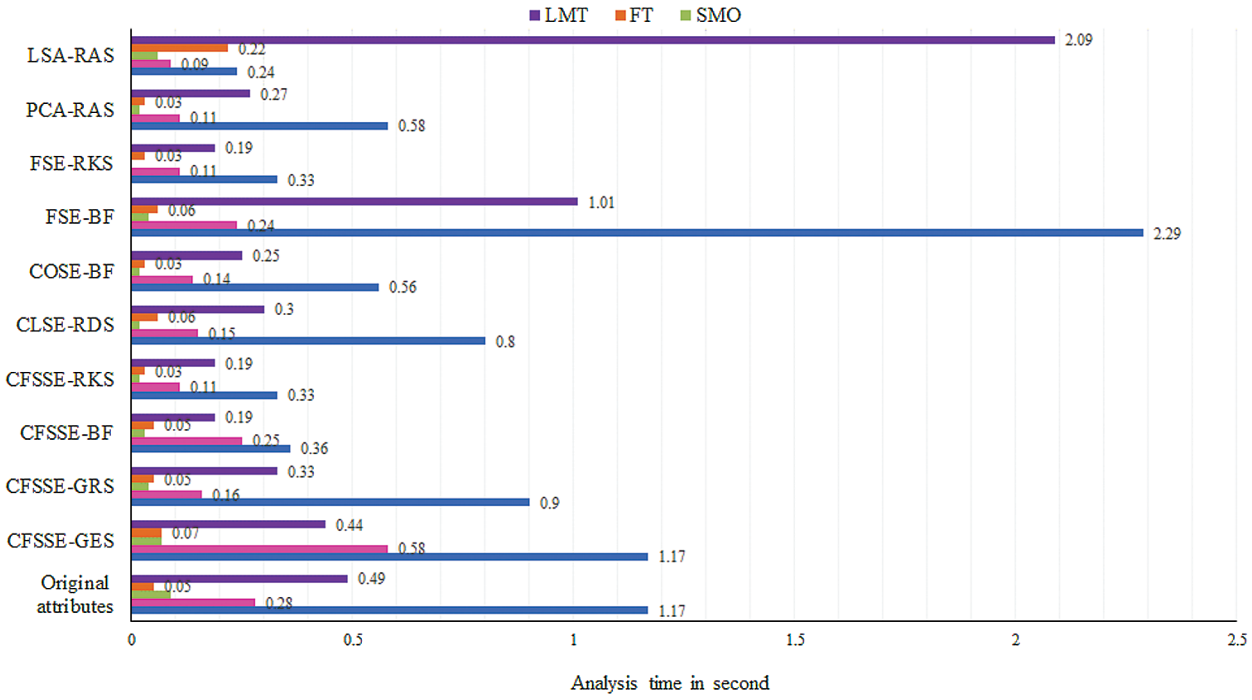

Figure 9: Analysis time of classification methods using original attributes and selected attributes, and selected and generated feature subsets

5 Discussions of Validation and Comparative Analysis Results

The PCA-RAS and LSA-RAS methods consider the contribution of all attributes in the generation and selection of an optimal subset of novel features. This is the reason for the better performance of classification methods in lymph disease recognition using the PCA-RAS and LSA-RAS selected feature subset than the original attributes and other selected subset of attributes. Though the better performance of the LSA-RAS selected feature than the PCA-RAS selected feature, in-class recognition is due to the majority of the nominal attribute (15 out of 18) in the lymph disease dataset.

The PCA minimizes the correlation and maximizes the variance of the three original numerical attributes, while the LSA measures the textual coherence of most of the nominal attributes effectively and generates novel latent variables that result in a significant improvement in the accuracy of classification methods. The deprived performances of the eight attribute selection methods (Tab. 3) in class recognition of the lymph diseases are due to the selection of a few significant attributes of the original data. It may cause the loss of the class identity information on the discarded attributes and hence the substandard recognition performance of the classification methods using the selected attributes than the original attribute, and the PCA-RAS and LSA-RAS selected feature subsets. Analysis results in Figs. 2–8 and Tabs. 2–7 confirm the better performance of the tree-based than the function-based classification methods using the LSA-RAS selected feature subset. Specifically, the LMT achieved the best recognition accuracy than the rest of the classification methods and the performance of the FT is comparable to MLP and better than the SL and SMO methods. Using the efficient LSA-RAS selected feature subset is the reason for the improved recognition accuracy of each of the classification methods. The improved performance of classification methods in recognition of lymph disease in combination with the feature selection methods is also discussed in some past studies summarized in Tab. 1, like the performance improvement of RF using GA, PCA, and ReliefF, etc. (maximum accuracy of 92.2% using GA selected feature subset) [17]; the maximum accuracy of 82.65% of classification method using the rough set selected feature subset [20]; improved accuracy of NB and C4.5 classification methods using information gain (IG), relief, and consistency-based subset evaluation (CNS), etc. (maximum accuracy of 83.24%) [21]; improved accuracy of NB, MLP, and J48 classification method using the IG, gain ratio, and symmetrical, etc. (maximum accuracy of 84.46%) [24]; and improved accuracy ($84.94 \pm 8.42\%$) of the NB method using the artificial immune system self-adaptive attribute weighting method [27], etc.

The LSA or the combination of the LSA with RAS in lymph disease recognition is not implemented before in the previously published research. Though, the LSA method has been implemented in different applications [41–45], including topic modeling [41], remote sensing image fusion [42], patient-physician communication [43], essay evaluation [44], and psychosis prediction [45], etc. The semantic information is obtained by combining the likelihood of the co-occurrence in the LSA. Also, the latent variables attempt to link the nominal attributes of the instance to their respective class maximally, which causes the improved performance of the classification methods. The improved performance of the RAS method in feature selection is due to its characteristics to combine the entropy, gain-ratio, and relief criteria. The combination of the earlier three criteria reduces the redundancy in the selected feature subset. Some of them have been used independently in the feature selection of the lymph dataset [17,21,24], like reliefF (accuracy of 84.2%) [17], information gain, and reliefF (accuracy of 82.63%, and 81.47%) [21], and information gain, gain ratio, and reliefF (accuracy of 77.02%–80.40%) [24], etc. Among the three functions-based classification methods, the MLP results in the maximum accuracy, using

the LSA-RAS selected feature subset. The tree-based LMT achieved the maximum recognition accuracy, using a similar feature subset. The best accuracy of the tree-based classification method and improved accuracy of the function-based classification method is also confirmed in the earlier studies [17,22,24–26], like the best recognition accuracy of 92.2% of random forest method [17], maximum accuracy of 86.49% of SMO and FT methods using the original attributes of lymph dataset [22], the accuracy of 84.46% of MLP, using chi-square selected and original attributes [24], the training accuracy of 85.47% of hybrid radial basis function neural network [25], and the maximum accuracy of 83.51% of ensembles of decision trees [26], etc.

The better performance of the tree-based classification methods; LMT and FT are due to the less number of adjustable parameters after using a significant subset of features selected by the LSA-RAS method, a reduced amount of noise of original attributes in latent variables, and negligible influence of noise, etc. The improved performance of the MLP method is due to the reduced uncertainty of the input and output by using the LSA-RAS selected feature subset. Among the implemented feature selection methods in the recognition of the lymph disease in the previous study [17,20,21,24,27], the best accuracy has been achieved for the combination of the GA and random forest classification methods [17]. The proposed approach LSA-RAS-LMT in the present study achieved the maximum recognition accuracy of the lymph disease than previously published reports. A significant improvement in the accuracy of the LMT (10.81%), SL (9.46), and ML (8.78%) has been achieved (Tab. 5). The analysis time of the LSA-RAS-LMT approach in the present analysis of the lymph dataset was 2.09 s (Fig. 9). It is in between the analysis time 0.02 s–11.77 s of [20] and, 0.0004 s–0.0051 s (Linux cluster node (Inter(R) Xeon(R) @3.33 GHz, and 3 GB memory) [28]. The area under the ROC of the LSA-RAS-LMT approach in the present analysis is equal to 0.97 (Tab. 6). It is higher than the area under ROC of other approaches [17,19,23,27], like 0.843-0.954 [17], 80.48 [19], 91.3757 ± 3.25–91.8005 ± 3.61 [23], and 92.99 ± 4.15–95.01 ± 4.87 [27]. The LSA-RAS-LMT method has the maximum value of the kappa coefficient (0.89) (Tab. 5) in the present analysis. It is also higher than the earlier achieved value of the kappa coefficient [17,18], like 0.512–0.879 [17] and 0.500–0.629 [18].

Moreover, the LSA-RAS approach has been validated in the recognition of other benchmark diseases (primary tumor, breast cancer, audiology, fertility, and post-operative patient) [32]. The performance of classification approaches is summarized in Tab. 8. It is obvious that the LSA-selected features subset results in improved accuracy of each of the classification methods than the original attributes. Specifically, a major improvement in accuracy of MLP in the primary tumor (55.45%), SL and SMO in the post-operative patient (26.61%), and FT (53.39%) and LMT (48.95%) in primary tumor has been noticed. The LMT classifier has an improved recognition performance in the analysis of most of the disease datasets. Deep neural networks such as convolutional and recurrent neural networks are used mainly in the preprocessing and classification of the image, text, and continuous data successfully in the past studies [11,12,29–31]. Though the lymph and other disease datasets selected in the present study contains the discrete values of numeric and nominal attributes, therefore, the direct implementation of the deep neural networks and its comparison with the proposed approach is not feasible. However, there is a need to explore the possibility in the future research.

Table 8: Performance of the classification methods in recognition of other diseases

Dataset	Classification method	Original attributes				LSA-RAS selected features			
		Performance measures				Performance measures			
		CC	MAE	RMSE	k	CC	MAE	RMSE	k
Primary tumor	MLP	38.36	0.06	0.20	0.31	93.81	0.02	0.08	0.93
	SL	48.67	0.06	0.18	0.41	93.81	0.01	0.07	0.93
	SMO	46.90	0.08	0.20	0.39	60.18	0.08	0.20	0.53
	FT	43.66	0.06	0.21	0.37	97.05	0.00	0.05	0.97
	LMT	48.68	0.06	0.18	0.41	97.63	0.00	0.05	0.97
Breast cancer Wisconsin	MLP	95.27	0.05	0.20	0.90	96.28	0.05	0.17	0.92
	SL	95.99	0.05	0.17	0.91	96.42	0.05	0.17	0.42
	SMO	95.89	0.05	0.17	0.91	96.42	0.05	0.17	0.42
	FT	95.79	0.05	0.17	0.91	96.71	0.03	0.17	0.93
	LMT	95.99	0.05	0.17	0.91	96.41	0.05	0.17	0.92
Audiology	MLP	83.19	0.02	0.10	0.80	86.28	0.02	0.09	0.84
	SL	84.07	0.01	0.10	0.81	89.38	0.01	0.08	0.88
	SMO	81.86	0.08	0.20	0.79	86.19	0.02	0.10	0.84
	FT	84.51	0.02	0.10	0.82	90.26	0.01	0.08	0.89
	LMT	84.07	0.01	0.10	0.81	90.71	0.01	0.08	0.89
Fertility	MLP	88	0.14	0.33	0.34	93	0.07	0.25	0.63
	SL	88	0.5	0.3	0.0	91	0.14	0.26	0.48
	SMO	88	0.12	0.35	0.0	89	0.12	0.35	0.0
	FT	86	0.18	0.37	0.03	92	0.13	0.26	0.59
	LMT	88	0.49	0.49	0.11	91	0.14	0.26	0.48
Post-operative patient	MLP	55.56	0.30	0.49	0.09	97.78	0.03	0.11	0.95
	SL	71.11	0.44	0.47	0.0	97.79	0.04	0.11	0.95
	SMO	67.78	0.30	0.39	0.06	97.78	0.23	0.29	0.95
	FT	64.44	0.29	0.44	0.12	97.78	0.01	0.10	0.95
	LMT	71.11	0.44	0.47	0.0	97.79	0.03	0.10	0.95

Note: CC-correct classification rate in %, RMSE-root means square error, MAE-mean absolute error, and k-Kappa coefficient.

6 Conclusions and Research Scope

In the present study, a competent feature generation and selection method (LSA-RAS) of lymph disease recognition has been implemented and validated. The LSA-RAS method results in the improved accuracy of different classification methods. The tree-based methods achieved better performance than the function-based classification methods using the LSA-RAS selected feature subset. Furthermore, hybrids approach (LSA-RAS-LMT) using the combination of feature generation and selection, and classification methods achieved the maximum recognition accuracy and improved the value of other evaluation metrics than other approaches available in the published literature. The LSA-RAS-LMT approach is efficient in the recognition of the lymph disease and analogous disease datasets. Future research will focus on further improvement in the accuracy of the classification methods for lymph disease recognition.

Acknowledgement: This work is supported by the Startup Foundation for Introducing Talent of NUIST. The authors acknowledge the anonymous reviewers for their valued comments and suggestions.

Funding Statement: This work is supported by the Startup Foundation for Introducing Talent of NUIST, Project No. 2243141701103.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Yassin, N. I., Omran, S., El Houby, E. M., Allam, H. (2018). Machine learning techniques for breast cancer computser aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine, 156,* 25–45. DOI 10.1016/j.cmpb.2017.12.012.
2. Chan, H. P., Hadjiiski, L. M., Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Medical Physics, 47(5),* e218–e227. DOI 10.1002/mp.13764.
3. Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A. et al. (2016). Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Transactions on Medical Imaging, 35(5),* 1170–1181. DOI 10.1109/TMI.2015.2482920.
4. Moore Jr, J. E., Bertram, C. D. (2018). Lymphatic system flows. *Annual Review of Fluid Mechanics, 50,* 459–482. DOI 10.1146/annurev-fluid-122316-045259.
5. Suami, H. (2017). Lymphosome concept: Anatomical study of the lymphatic system. *Journal of Surgical Oncology, 115(1),* 13–17. DOI 10.1002/jso.24332.
6. Leone, A., Diorio, G. J., Pettaway, C., Master, V., Spiess, P. E. (2017). Contemporary management of patients with penile cancer and lymph node metastasis. *Nature Reviews Urology, 14(6),* 335–347. DOI 10.1038/nrurol.2017.47.
7. Hu, D., Li, L., Li, S., Wu, M., Ge, N. et al. (2019). Lymphatic system identification, pathophysiology and therapy in the cardiovascular diseases. *Journal of Molecular and Cellular Cardiology, 133,* 99–111. DOI 10.1016/j.yjmcc.2019.06.002.
8. Arrivé, L., Monnier-Cholley, L., Cazzagon, N., Wendum, D., Chambenois, E. et al. (2019). Non-contrast MR lymphography of the lymphatic system of the liver. *European Radiology, 29(11),* 5879–5888. DOI 10.1007/s00330-019-06151-6.
9. Masoudi, M., Pourreza, H. R., Saadatmand-Tarzjan, M., Eftekhari, N., Zargar, F. S. et al. (2018). A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. *Scientific Data, 5(1),* 1–9. DOI 10.1038/sdata.2018.180.
10. Wang, H., Zhou, Z., Li, Y., Chen, Z., Lu, P. et al. (2017). Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 F-fDG PET/CT images. *EJNMMI Research, 7(1),* 1–11. DOI 10.1186/s13550-017-0260-9.
11. Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z. et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging, 35(5),* 1285–1298. DOI 10.1109/TMI.42.
12. Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine, 161,* 1–13. DOI 10.1016/j.cmpb.2018.04.005.
13. Jain, D., Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal, 19(3),* 179–189. DOI 10.1016/j.eij.2018.03.002.
14. Rahman, M., Usman, O. L., Muniyandi, R. C., Sahran, S., Mohamed, S. et al. (2020). A review of machine learning methods of feature selection and classification for autism spectrum disorder. *Brain Sciences, 10(12),* 949–972. DOI 10.3390/brainsci10120949.

15. Cilia, N. D., De Stefano, C., Fontanella, F., Raimondo, S., Scotto di Freca, A. (2019). An experimental comparison of feature-selection and classification methods for microarray datasets. *Information, 10(3),* 109–122. DOI 10.3390/info10030109.
16. Agor, J., Özaltın, O. Y. (2019). Feature selection for classification models via bilevel optimization. *Computers & Operations Research, 106,* 156–168. DOI 10.1016/j.cor.2018.05.005.
17. Azar, A. T., Elshazly, H. I., Hassanien, A. E., Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine, 113(2),* 465–473. DOI 10.1016/j.cmpb.2013.11.004.
18. Cano, A., Zafra, A., Ventura, S. (2013). Weighted data gravitation classification for standard and imbalanced data. *IEEE Transactions on Cybernetics, 43(6),* 1672–1687. DOI 10.1109/TSMCB.2012.2227470.
19. de Falco, I. (2013). Differential evolution for automatic rule extraction from medical databases. *Applied Soft Computing, 13(2),* 1265–1283. DOI 10.1016/j.asoc.2012.10.022.
20. Derrac, J., Cornelis, C., García, S., Herrera, F. (2012). Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection. *Information Sciences, 186(1),* 73–92. DOI 10.1016/j.ins.2011.09.027.
21. Hall, M. A., Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering, 15(6),* 1437–1447. DOI 10.1109/TKDE.2003.1245283.
22. Jha, S. K., Pan, Z., Elahi, E., Patel, N. (2019). A comprehensive search for expert classification methods in disease diagnosis and prediction. *Expert Systems, 36(1),* e12343. DOI 10.1111/exsy.12343.
23. Jiang, L., Cai, Z., Zhang, H., Wang, D. (2012). Not so greedy: Randomly selected naive Bayes. *Expert Systems with Applications, 39(12),* 11022–11028. DOI 10.1016/j.eswa.2012.03.022.
24. Karabulut, E. M., Özel, S. A., Ibrikci, T. (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology, 1,* 323–327. DOI 10.1016/j.protcy.2012.02.068.
25. Li, M., Tian, J., Chen, F. (2008). Improving multiclass pattern recognition with a co-evolutionary RBFNN. *Pattern Recognition Letters, 29(4),* 392–406. DOI 10.1016/j.patrec.2007.10.019.
26. Rodríguez, J. J., García-Osorio, C., Maudes, J. (2010). Forests of nested dichotomies. *Pattern Recognition Letters, 31(2),* 125–132. DOI 10.1016/j.patrec.2009.09.015.
27. Wu, J., Pan, S., Zhu, X., Cai, Z., Zhang, P. et al. (2015). Self-adaptive attribute weighting for naive Bayes classification. *Expert Systems with Applications, 42(3),* 1487–1502. DOI 10.1016/j.eswa.2014.09.019.
28. Wu, J., Pan, S., Zhu, X., Zhang, P., Zhang, C. (2016). Sode: Self-adaptive one-dependence estimators for classification. *Pattern Recognition, 51,* 358–377. DOI 10.1016/j.patcog.2015.08.023.
29. Wang, H., Shi, H., Lin, K., Qin, C., Zhao, L. et al. (2020). A high-precision arrhythmia classification method based on dual fully connected neural network. *Biomedical Signal Processing and Control, 58,* 101874. DOI 10.1016/j.bspc.2020.101874.
30. Jin, Y., Qin, C., Liu, J., Lin, K., Shi, H. et al. (2020). A novel domain adaptive residual network for automatic atrial fibrillation detection. *Knowledge-Based Systems, 203,* 106122. DOI 10.1016/j.knosys.2020.106122.
31. Jin, Y., Qin, C., Liu, J., Li, Z., Shi, H. et al. (2021). A novel incremental and interactive method for actual heartbeat classification with limited additional labeled samples. *IEEE Transactions on Instrumentation and Measurement, 70,* 1–12. DOI 10.1109/TIM.2021.3069021.
32. Dua, D., Graff, C. (2019). *UCI Machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.
33. Frank, E., Hall, M. A., Witten, I. H. (2016). The WEKA Workbench. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques",* pp. 1–128. USA: Morgan Kaufmann.
34. Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38(1),* 188–230. DOI 10.1002/aris.1440380105.
35. Liu, H., Motoda, H. (2012). *Feature selection for knowledge discovery and data mining.* USA: Springer.
36. Christopher, M. B. (2016). *Pattern recognition and machine learning.* USA: Springer.
37. Kotsiantis, S. B. (2005). Logitboost of simple Bayesian classifier. *Informatica, 29(1),* 53–59. http://www.informatica.si/index.php/informatica/article/viewFile/17/11.

38. Gama, J. (2004). Functional trees. *Machine Learning, 55(3),* 219–250. DOI 10.1023/B:MACH.0000027782.67192.13.
39. Landwehr, N., Hall, M., Frank, E. (2005). Logistic model trees. *Machine Learning, 59(1–2),* 161–205. DOI 10.1007/s10994-005-0466-3.
40. Ben-David, A. (2008). Comparison of classification accuracy using cohen's weighted kappa. *Expert Systems with Applications, 34(2),* 825–832. DOI 10.1016/j.eswa.2006.10.022.
41. Kim, S., Park, H., Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-lSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications, 152,* 113401. DOI 10.1016/j.eswa.2020.113401.
42. Fernandez-Beltran, R., Haut, J. M., Paoletti, M. E., Plaza, J., Plaza, A. et al. (2018). Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(12),* 4982–4993. DOI 10.1109/JSTARS.4609443.
43. Vrana, S. R., Vrana, D. T., Penner, L. A., Eggly, S., Slatcher, R. B. et al. (2018). Latent semantic analysis: A new measure of patient-physician communication. *Social Science & Medicine, 198,* 22–26. DOI 10.1016/j.socscimed.2017.12.021.
44. Zupanc, K., Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems, 120,* 118–132. DOI 10.1016/j.knosys.2017.01.006.
45. Rezaii, N., Walker, E., Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *NPJ Schizophrenia, 5(1),* 1–12. DOI 10.1038/s41537-019-0077-9.