Alexandria Engineering Journal (2018) 57, 1557–1567

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

ORIGINAL ARTICLE

Computer aided diagnosis of pulmonary hamartoma from CT scan images using ant colony optimization based feature selection

J. Dhalia Sweetlin [a], H. Khanna Nehemiah [a,*], A. Kannan [b]

[a] Ramanujan Computing Centre, Anna University, Chennai 600025, Tamil Nadu, India
[b] Department of Information Science and Technology, Anna University, Chennai 600025, Tamil Nadu, India

Received 27 September 2016; revised 25 March 2017; accepted 23 April 2017
Available online 17 May 2017

KEYWORDS

Computer aided diagnosis;
Pulmonary hamartoma;
Ant colony optimization;
Cosine similarity;
Rough dependency;
Support vector machine

Abstract *Background:* Computer-aided diagnosis (CAD) systems for the detection of lung disorders play an important role in clinical decision making. CAD systems provide a second opinion to the physician in interpreting computed tomography (CT) images. In this work, a CAD system to diagnose pulmonary hamartoma nodules from chest CT images is proposed.

Methods: Segmentation of lung parenchyma from CT images is carried out using Otsu's thresholding method. Nodules are considered to be the region of interests (ROIs) in this work. Texture, shape and run length based features are extracted from the ROIs. Cosine similarity measure (CSM) and rough dependency measure (RDM) are used independently as filter evaluation functions with ant colony optimization (ACO) to select two subsets of features. The selected subsets are used to train two classifiers namely support vector machine (SVM) and Naive Bayes (NB) classifiers using 10-fold cross validation. All the four trained classifiers are tested and the performance measures are estimated.

Results: CT slices of patients affected with pulmonary cancer and hamartoma are used for experimentation. From the lung parenchymal tissues of 300 CT slices, 390 nodules are extracted. The feature selection algorithms, ACO-CSM and ACO-RDM are run for different feature subset sizes. The selected features are used to train SVM and NB classifiers. From the results obtained, it is inferred that SVM classifier with the feature subsets chosen by ACO-RDM feature selection approach yielded a maximum classification accuracy of 94.36% with 38 features.

Conclusion: From the results, it can be clearly inferred that selecting relevant features to train the classifier has a definite impact on the performance of the classifier.

1. Introduction

Pulmonary hamartomas are one of the common causes of solitary pulmonary nodules [1,2]. They are benign tumors which grow as a disorganized mass and account for 75% of all benign

* Corresponding author.
E-mail addresses: nehemiah@annauniv.edu (H.K. Nehemiah), kannan@annauniv.edu (A. Kannan).

Peer review under responsibility of Faculty of Engineering, Alexandria University.

http://dx.doi.org/10.1016/j.aej.2017.04.014

tumors of the lung [3]. They are asymptomatic and do not show any symptom. They are detected accidentally when computed tomography (CT) scan is done for some diagnosis [4]. Hamartomas are made of normal lung tissues including fat, connective tissue, smooth muscle and calcification, which grow abnormally without affecting surrounding tissues [5]. Hamartoma nodules appear like popcorns with fat and calcification [6]. In cases where the nodules contain little fat and no calcification, the radiologists find it difficult to discriminate benign and malignant nodules.

Based on the location of hamartomas in the lung, they are of two types namely peripheral parenchymal type and central endobronchial type. Peripheral parenchymal type hamartomas arising from small bronchi are asymptomatic in nature whereas central endobronchial hamartomas arising from large bronchi are associated with symptoms of obstruction [7].

Hamartoma nodules are small and ranges from 1 to 3 cm in size. They may grow above this range in certain exceptional cases [5,6] and sometimes they may be multiple [8]. As they grow in size, the nodules may imitate bronchogenic carcinoma and hence accurate interpretation of imaging and diagnosis is important [9]. Fine needle aspiration (FNA) helps in the diagnosis of hamartoma nodules and differentiates them from other pulmonary nodules [10].

In chest radiographs pulmonary hamartomas are identified as coin lesions [11]. Chest radiographs fail to detect hamartomas when multiple nodules are present and in the absence of calcification and fat [12]. The central lucency caused by adipose tissue may be considered as air within a cavity necessitating differential diagnosis [12]. CT images are superior to chest X-rays and are more sensitive in detecting fat and popcorn-like calcifications. Hamartoma nodules are smooth and appear with sharp margins in CT images. CT scan examination is usually suggested to confirm the disease and its similarity toward other nodules of lung [5]. Li et al. [13] and Awai et al. [14] suggest that a CAD scheme can help the radiologists to make conclusions about the nodules.

In this work, a computer aided diagnosis system to detect the presence of pulmonary hamartomas from lung CT images is proposed. The lung tissues are segmented using a threshold based segmentation approach. Nodules considered as ROIs are extracted from the segmented lung and from which textural, run length and geometrical features are extracted. A filter approach that combines ant colony optimization algorithm is used to select the relevant features. The selected features are used to train support vector machines (SVM) and Naive Bayes (NB) classifiers to mark the presence or absence of hamartoma nodules. Cosine similarity measure (CSM) and rough set dependency measure (RDM) direct the ant colony algorithm in feature subset selection. The rest of the article is organized as follows: Literature survey related to the work carried out is presented in Section 2. System framework is explained in Section 3. Section 4 contains experimental results and discussions and Section 5 presents the conclusion and future scope of this work.

2. Literature survey

2.1. Computer aided diagnosis systems

Elizabeth et al. developed a diagnostic system [15] to mark the presence or absence of lung cancer in chest CT images. They applied optimal thresholding, convex hull and Canny's edge operator in sequence to segment lungs and reconstruct its edges. A probabilistic neural network was trained using GLCM features that were extracted from the nodules. They evaluated their approach using 100 diseased images and 100 normal lung images and obtained an accuracy of 97%.

Shiraishi et al. developed a CAD system [16] to discriminate benign and malignant solitary pulmonary nodules on chest X-rays. The nodules were marked by the radiologists and segmented by means of the difference image technique [17]. Their training dataset included 53 chest X-rays containing solitary pulmonary nodules smaller than 3 cms with no calcification. They located 22 benign nodules and 31 malignant nodules from which seventy-five image features were extracted. A linear discriminant analyzer (LDA) [17] was used for feature selection which selected six image features. Likelihood measure of malignancy was used to classify the data. They tested their approach on a test set that included 5 chest X-rays containing 3 cancerous nodules and 2 benign nodules. The test set was obtained from the Japanese Society of Radiological Technology (JSRT) image database. Their system obtained an accuracy of 80% when the likelihood measure of malignancy was set to 50%.

Li et al. developed a CAD system using likelihood estimate with LDA classifier [13] and evaluated whether the output of a CAD system helps the radiologists in differentiating the lung nodules. Their dataset consisted of high resolution computed tomography (HRCT) slices containing 28 primary lung cancer nodules and 28 benign nodules. The images were analyzed by 16 radiologists without and with the computer output to mark their confidence level regarding the malignancy of a nodule. The area under the ROC curve of the CAD scheme was 0.831 for discriminating benign and malignant nodules. The ROC value was obtained when the radiologists used the CAD system improved from 0.785 to 0.853 by a statistically significant level of $p = 0.016$.

Han et al. proposed a CAD system [18] based on hierarchical vector quantization (VQ) to detect pulmonary nodules in the early stage. High level VQ was used to segment the lungs from CT images and low level VQ was used to detect and segment the nodules. Rule based filtering was carried out to select features for training SVM classifier. They validated their approach on the CT scan images of 205 patients having juxtapleural nodule annotation, taken from Lung Image Database Consortium (LIDC). Geometric, intensity, Gradient and Hessian features were extracted from 475 nodules and were used to train SVM classifier in different combinations. Their system obtained a maximum of 89.2% sensitivity at 4.14 false positives per scan when intensity and gradient features were used to train the classifier.

Elizabeth et al. proposed an approach to identify the most promising slice to diagnose lung cancer from chest CT images [19]. From the segmented lungs, the ROIs that existed in the same location in three adjacent slices were extracted using region growing approach and considered for analysis. The best slice among the three adjacent slices was identified and features were extracted from the ROIs that were present in the best slice. The labels of these ROIs and their features were used to train a radial basis function neural network. They trained their system with 1564 chest CT slices and tested with 150 slices. Their system was able to detect the cancerous nodules with 94.44% accuracy.

Choi and Choi in their work [20] toward nodule detection from CT images segmented the lung volume using optimal multiple thresholding, 3D connected component labeling and rule based pruning. 2D and 3D features extracted from the nodules were used to train Genetic Programming based classifier. They evaluated their system using the scan images of 32 patients containing 5453 annotated CT slices obtained from LIDC. Among the 1716 candidate nodules that were extracted, only 76 were cancerous nodules. Hence to create a balance in the dataset used for training the classifier, along with the 76 nodules, only 76 non-nodules were considered. 2D, 3D geometric and intensity based features were extracted from these ROIs. From this dataset, 80% of the features were used to train GP based classifier and the remaining 20% were used to test. Their system was able to achieve an accuracy of 89.6% with 94.1% sensitivity and 5.45 FPs/scan.

Messay et al. proposed a CAD system to detect lung nodules [21]. Initially they preprocessed the CT images by orienting, down-sampling, and performing local contrast enhancement (LCE) and lung segmentation. Candidate nodules were extracted using intensity thresholding and morphological processing. They trained their system with the CT scan images obtained from the medical branch of University of Texas and evaluated their approach using CT scans of 84 patients taken from LIDC. The training set contained 606 nodules from which 245 features were generated. Features were selected using sequential forward approach with two distinct classifiers: Fisher Linear Discriminant (FLD) classifier and quadratic classifier. FLD selected 40 features with a training sensitivity of 97.52%. From the LIDC dataset, a total of 143 nodules were extracted and FLD was able to detect 92.8% of all nodules in the dataset.

2.2. Feature subset selection

Boroczky et al. developed a CAD system to discriminate true cancer nodules from nodule like structures such as blood vessels [22]. Their approach detected the volume of interest (VOI) enclosing a nodule, segmented and labeled as nodule, background or lung wall. From each VOI, twenty-three 2D and 3D gray level distribution and shape features were extracted. SVM driven genetic algorithm based feature selection was carried out in their work to select features in which sensitivity was used as the fitness parameter. The selected features were used to train SVM classifier. Their database contained 52 true nodules and 443 false nodules obtained from different multi-slice CT scans. Their method was able to generate ten optimal features from the 23 features yielding 100% sensitivity and 56.4% specificity using leave-one-out cross validation.

Kindie et al. suggested a two-step approach using rough sets for feature selection [23]. After handling the missing values either by rejecting or by imputing records, an indiscernibility relation based on rough sets was derived to select the reducts. These reducts were used to train a back propagation neural network. Their system yielded an accuracy of 97.3% with 13 features when applied to hepatitis dataset, 98.6% with 7 features when applied to Wisconsin breast cancer dataset and 90.4% with 6 features when applied to Statlog heart disease dataset taken from University of California Irvine (UCI) repository.

Chen et al. proposed a hybrid ACO feature selection algorithm combining F-score measure and SVM classifier [24]. Their approach updated the pheromone values based on the classification accuracy of SVM and the size of feature subset. In every iteration, a specified number of features with high pheromone values were added to the subset. The subsets were evaluated using SVM classifier. The subset that yielded maximum classification accuracy was considered as the best subset. They tested their approach on Corel image databases, from which 80 images containing 19 features with 4 class labels were used. The algorithm selected 9 features on an average using 10-fold cross validation. The recall and precision values of the system are found to be 96.45% and 97.08%.

Tabakhi et al. proposed a feature selection approach using ant colony optimization (ACO) [25], in which cosine similarity measure (CSM) was used as the filter evaluation function. Features with highest pheromone values and lowest similarity values to the existing features in the partially constructed subsets were added to the subsets. Pheromone updation was based on the frequency of the features selected in all the subsets. After all the specified iterations, the features were sorted based on the pheromone values and the required features were considered to form the subset to train the SVM classifier. They evaluated their approach on Wine, Hepatitis, Ionosphere, Dermatology, Spam-base, Arrhythmia, Madelon and Arcene datasets taken from UCI repository. Their approach gave an average classification error rate of 19.84%.

Liu et al. investigated nine common CT imaging signs to find the correlation between CT findings and the lung diseases [26]. The lung tissues with the imaging signs were considered as ROIs. From these ROIs, they extracted 180 features collectively using histogram of gradient (HOG) features, wavelet features, local binary pattern features and the features generated from histogram. Optimal features were selected using a genetic optimization algorithm in which FIsher criterion was used as the fitness measure. All the feature subsets yielding fitness above a threshold are used in the evaluation of a k-NN classifier. The subset giving maximum classification accuracy was used to train SVM, Bag, Naive Bayesian, k-NN and Adaboost classifiers. They tested their approach on a set of 511 ROIs using fivefold cross-validation and obtained 92 features at 0.5 threshold value. They obtained an average accuracy of 80.26% with SVM, 77.88% with Bag, 77.84% with Naive Bayesian, 73.58% with k-NN and 75.70% with Adaboost classifiers.

Christopher et al. used filter approach to select features for the CAD system which they developed to diagnose allergic rhinitis [27]. Their system used information gain and Pearson's correlation to select 40 features from a set of 91 features. When evaluated on 872 samples collected from an Allergy testing centre at Chennai, India, their system was able to achieve an accuracy of 88.31%.

The following three inferences are arrived from the review of the works carried out by other researchers. First, the nodular CT sign of different pulmonary diseases may look similar and present difficulty for the clinician to interpret accurately [26,28]. Second, introducing feature selection in a diagnosis system which deals with numerous image features, increases the performance of the classifier and decreases the computation time [24,29]. Third, meta-heuristic algorithms suggest better solutions compared to traditional optimization algorithms in medical diagnosis as it involves features in large scale [30]. Hence in this proposed work, a dedicated computer aided diagnosis system to detect the presence of hamartoma is devel-

oped in which features are selected using ant colony optimization approach. To find the optimal subset, dependency measure based on rough sets and cosine similarity measures are used as filter evaluation functions in ant colony algorithm. To the best of our knowledge, this is the first work to develop a CAD system to diagnose pulmonary hamartoma and apply ACO in feature selection.

3. System framework

The framework of the proposed CAD system is presented in Fig. 1. The major subsystems of this framework are segmentation subsystem, ROI extraction subsystem, feature extraction subsystem, feature selection subsystem, classification subsystem and two databases namely image database and feature database.

3.1. Segmentation subsystem

Otsu's segmentation algorithm is used to separate the lung tissues from the CT slice by finding a suitable threshold. Airways, disease patterns and sometimes image noise may be seen as holes in the segmented binary image [31]. In a CT image, the intensity values of the lung pixels are the same as the background pixels [32] and hence they are also removed to get the segmented lung fields.

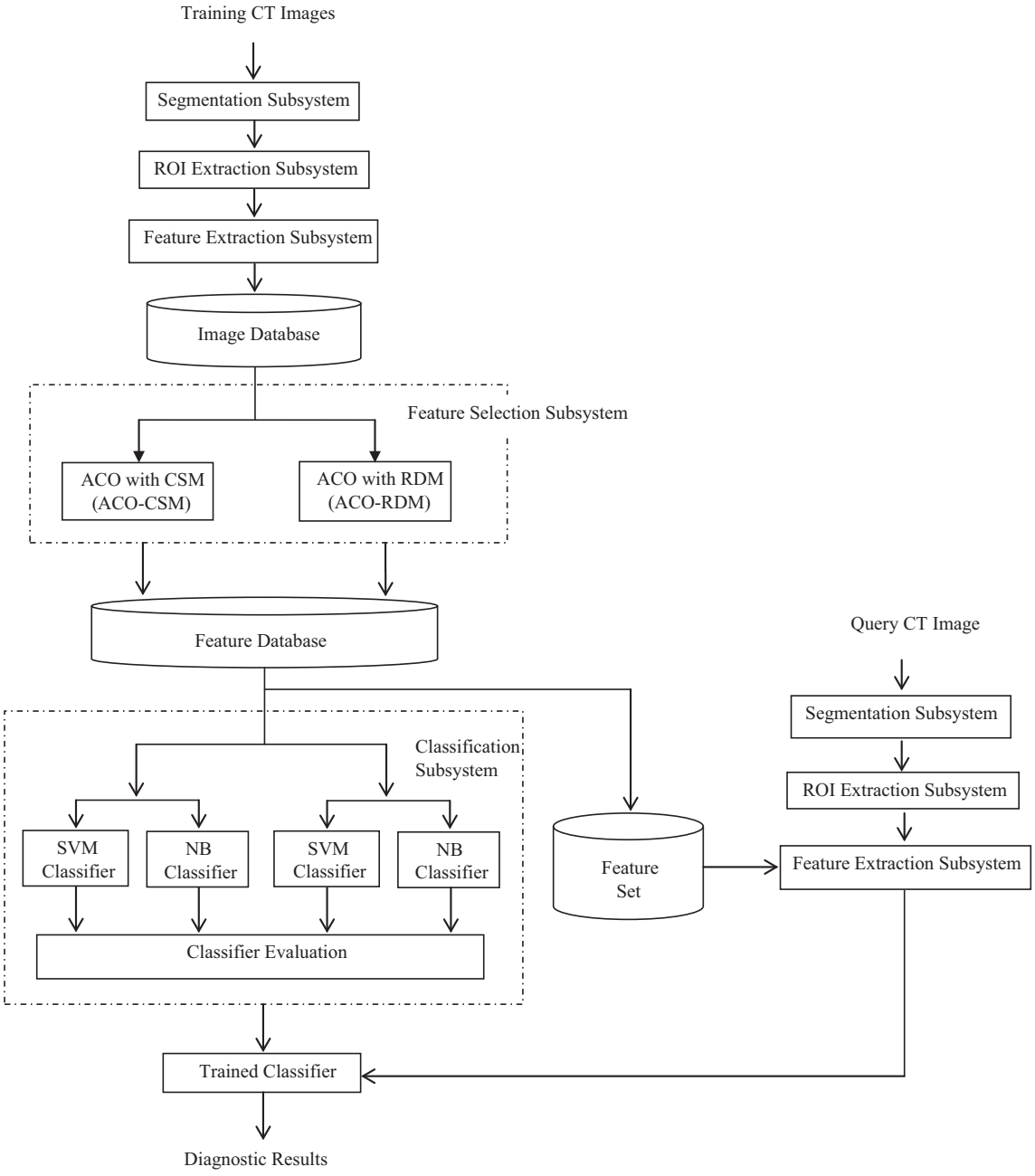

Figure 1 System framework.

Input: Chest CT slice

Step 1: Apply Otsu's thresholding algorithm to the input chest CT slice to obtain the binary image [33].
Step 2: Fill the holes or gaps present in the binary lung image with the intensity level of the pixels surrounding the holes using image morphological operations.
Step 3: Remove the background of the image using morphological operations to obtain the lung fields [31].
Output: Segmented lungs.

3.2. ROI extraction system

The pathology bearing regions are considered as ROIs in this work. Hamartoma nodules are well-defined, smooth and round exhibiting fat and calcification [5]. The size of the nodules may range from 1 to 3 cm [3,4].

Input: Segmented lungs.
Process: Extract all the nodular pathology bearing regions using pixel based segmentation, by finding a suitable threshold level and obtain their class labels from an expert.
Output: Region of Interests with their class labels.

3.3. Feature extraction subsystem

Input: ROIs.
Process:
Step 1: Compute the twenty-one GLCM features for each ROI in four orientations 0°, 45°, 90°, and 135° given in [34]. The features are auto-correlation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, two measures of homogeneity, maximum probability, variance, sum average, sum entropy, sum variance, difference variance, difference entropy, two information measures of correlation, normalized inverse difference and normalized inverse difference moment.

Step 2: Compute the run length features [35] suggested by Tang namely short run emphasis, long run emphasis, gray level non-uniformity, run length non-uniformity, run percentage, low gray level run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis and long run high gray level emphasis in 0°, 45°, 90° and 135° orientations.

Step 3: Compute the shape features namely area, major axis length, minor axis length, eccentricity, elongation, circularity ratio, centroid, orientation, filled area, convex area, Euler number, equiv-diameter, solidity, smoothness, extent and perimeter from each ROI [31].

Step 4: Combine the eighty-four features obtained from step 1, forty-four features from step 2 and sixteen features from step 3 to form the feature vector of an ROI.

Step 5: Perform min-max normalization [36] to the features in the feature vectors and convert the values to the range [0,1] using Eq. (1).

$$X' = \left(\frac{X - X_{min}}{X_{max} - X_{min}}\right)(new\_X_{max} - new\_X_{min}) + new\_X_{min} \qquad (1)$$

where X is the actual value of that feature in the feature vector, X' is the normalized value that is to be substituted for X in the feature vector, X_{min} is the minimum value of the feature and X_{max} is the maximum value of that feature.

Output: Training dataset containing the normalized feature vectors of each ROI.

3.4. Image database

The CT scan images of the lung, the slices corresponding to each CT scan image, the ROIs extracted from each CT slice along with their class labels, and the features extracted from the ROIs corresponding to each CT slice are stored as six relations in the image database.

The relations are CT_IMG (Img_ID, CTimage) with Img_ID as primary key, CTSLICE (Img_ID, SliceId, Slice) with {Img_ID, SliceId} as the primary key, ROI (Img_ID, SliceId, ROI_ID, ClassLabel) with {Img_ID, SliceId, ROI_ID} as the primary key, ROI_GLCM_FEATURES (Img_ID, SliceId, ROI_ID, FName, Orientation1, Orientation2, Orientation3, Orientation4) with {Img_ID, SliceId, ROI_ID, FName} as the primary key, ROI_RUNLENGTH_FEATURES (Img_ID, SliceId, ROI_ID, FName, Orientation1, Orientation2, Orientation3, Orientation4) with {Img_ID, SliceId, ROI_ID, FName} as the primary key and ROI_SHAPE_FEATURES (Img_ID, SliceId, ROI_ID, Area, MajorAxisLength, MinorAxisLength, Eccentricity, Elongation, CircularityRatio, Centroid, Orientation, FilledArea, ConvexArea, EulerNumber, Equiv-Diameter, Solidity, Smoothness, Extent, Perimeter) with {Img_ID, SliceId, ROI_ID} as the primary key.

3.5. Feature selection subsystem

The objective of this subsystem is to select a subset of relevant features to construct a classifier model. In this work, a filter based ant colony optimization is used in which features are selected based on the intrinsic characteristics of the features. A learning model is not used in the process of feature selection. This makes the filter approaches faster to implement thereby increasing its computational efficiency [25].

3.5.1. Ant colony based feature selection

ACO is a population based meta-heuristic approach [37] suggested by Dorigo et al. which is used to select a subset of features based on the behavior of ants searching for food. If an ant needs to choose between paths, it prefers the path with high pheromone level which indicates that a promising food source is available in that path. Over time, the pheromone trail starts to evaporate, thus reducing its attractive strength and avoids the convergence to a locally optimal solution. There would not be any exploration of new paths, if evaporation is not present.

The idea of ant colony optimization is to mimic this behavior of ants with simulated ants. The simulation environment is represented as a fully connected undirected graph $G = \langle V, E\rangle$ where V, the set of vertices $v_1, v_2 \ldots v_n$ corresponds to the features $F = \{f_1, f_2 \ldots f_n\}$ present in the feature database, E denotes the set of edges connecting vertices and 'n' represents the total number of vertices in the graph. Each feature is

mapped onto a vertex and hence the number of vertices in the graph is the same as the number of features in the image database. In this work, the number of ants N_{ant}, to explore the feature space is the same as the number of features 'n'. Each ant is initially placed on a vertex in the graph. Each ant visits a set of vertices thereby choosing features independently using cosine similarity [25] and rough dependency measures [38].

3.5.1.1. Cosine similarity measure.
Cosine similarity measure identifies the cosine value of the angle between two features. This measure gives information about the orientation of two features without considering their magnitude [39]. It is computed using Eq. (2).

$$sim(f_i,f_j) = \left| \frac{\sum_{a=1}^{m}(f_{ia}f_{ja})}{\sqrt{\left(\sqrt{\sum_{a=1}^{m}f_{ia}^2}\right)\left(\sqrt{\sum_{a=1}^{m}f_{ja}^2}\right)}} \right| \tag{2}$$

where f_i,f_j are any two features in 'm' feature vectors.

If two features have the cosine similarity value as 1, then they are said to be in same orientation.

Two features at 90° orientation have a similarity value 0 and features diametrically opposed have a similarity of 1 considering only the magnitude. Features with low similarity value are chosen to form the feature subsets.

3.5.1.2. Rough dependency measure.
A decision system is represented as $I_S = \{U, A, C, D\}$ where U is a non-empty finite set of objects called Universe, A is a non-empty set of features, C is the set of conditional features and D is the decision feature. Also, C and $D \subseteq A$. For any feature subset $S \subset A$ an indiscernible relation denoted by $IND(S)$ is defined as given in Eq. (3).

$$IND(S) = \{(x,y) \in UXU : \forall a \in S, a(x) = a(y)\} \tag{3}$$

where $a(x)$ gives the value of feature 'a' of object x. If $S \subseteq A$, for any $X \subseteq U$ then the lower and upper approximations of X with respect to S are defined as given in Eqs. (4) and (5).

$$\underline{S(X)} = \{X \in U : [x]_{IND(S)} \subseteq X \tag{4}$$

$$\overline{S(X)} = \{X \in U : [x]_{IND(S)} \cap X \neq \varnothing \tag{5}$$

where $[x]_{IND(S)} = \{y \in U : a(y) = a(x) \forall a \in S\}$ is the equivalence class of x in U/IND(S).

Positive region is the set of all objects from U which are classified with certainty to one class of U/IND(S), employing features from 'C'. It is computed using Eq. (6).

$$POS_{(S)}(D) = \bigcup_{X \in U/IND(P)} S(X) \tag{6}$$

Dependency of 'D' on S is defined as given in Eq. (7).

$$\gamma_s(D) = \frac{|POS_s(D)|}{|U|} \tag{7}$$

where $|U|$ is the cardinality of the objects in the universe.

In the proposed method, the decision system is defined by $I_s = \{U, A, C, D\}$ where U represents the training dataset containing the feature vectors with their class labels; A refers to $\{f_1, f_2 \ldots f_n$, class label$\}$; C refers to $\{f_1, f_2 \ldots f_n\}$ and D refers to {Class label = yes/No}. For any two features $f_i, f_j \in$ C, the heuristic information is given as in Eq. (8).

$$\gamma_{f_i,f_j}(D) = \frac{|POS_{f_i,f_j}(D)|}{|U|} \tag{8}$$

If the dependency measure of two features is equal to 1, then they are dependent on each other and if independent, the measure gives a value 0.

3.5.2. Algorithms for feature selection
Let $\mathbb{F}$ be a set of feature subsets; $\mathbb{F} \ni \{F_1, F_2 \ldots F_N\}$ where $0 < N \leqslant N_{ant}$. The number of feature subsets is represented using 'N' and the number of ants is represented using N_{ant}. Each feature subset contains a set of features; $F_i \ni \{f_1, f_2 \ldots f_{n_{max}}\}$ where $0 < n_{max} < n$; n represents the total number of extracted features. The pheromone value (τ) associated with every feature is set to a constant initially. Heuristic values are associated with ACO algorithms, which are prior known values used in tuning the algorithm to find an optimal solution. Features are added to feature subsets using either exploration or exploitation. Exploration is the ability to avoid local optima in feature search, thereby preventing the ants from selecting the same set of features into the subset. Exploitation is where the ants exploit the promising path based on their experiences. The exploration and exploitation balance is achieved using the parameters q_{rand} and $q_{\text{const}} \in [0,1]$. If $q_{\text{rand}} \leq q_{\text{const}}$, the ants exploit the known paths and the next feature is added to the subset using Eq. (9) when cosine similarity is used as the selection measure and Eq. (10) when rough dependency measure is used; if $q_{\text{rand}} > q_{\text{const}}$, the feature space is explored using the transition probability given in Eqs. (11) and (12) respectively. The feature with higher probability is added to the subset that is being generated.

$$f_j = \max([\tau_u][\eta(f_i,f_u)]^{\beta}) \tag{9}$$

$$f_j = \max([\tau_u][\gamma_{f_i,f_u}(D)]^{\beta}) \tag{10}$$

$$P_j(i,j) = \frac{[\tau_j][\eta(f_i,f_j)]^{\beta}}{\sum_{u \in j_i}[\tau_u][\eta(f_i,f_u)]^{\beta}} \tag{11}$$

$$P_j(i,j) = \frac{[\tau_j][\gamma_{f_i,f_j}(D)]^{\beta}}{\sum_{u \in j_i}[\tau_u][[\gamma_{f_i,f_u}(D)]^{\beta}]} \tag{12}$$

where τ_j is the pheromone level of feature f_j, $\eta(f_i,f_j)$ is the heuristic information between the two features, $\gamma_{f_i,f_j(\mathrm{D})}$ is the dependency measure of features f_i,f_j on the class label D, $\tau_u$is the pheromone level of the considered feature that is not yet included and β is the parameter to increase the significance of the heuristic information in selecting features. In this work, β is set to 1 and q_{const} is set to 0.7 [25].

3.5.2.1. Notations used.
Dataset: Training dataset with m feature vectors and n features
n_{max}: number of features to be selected in the final subset
N_{max_iter}: number of iterations
$sim(f_i, f_j)$: cosine similarity between features f_i and f_j
$\eta(f_i, f_j)$: heuristic information between features f_i and f_j based on cosine similarity
$\gamma(f_i, f_j)$: heuristic information between features f_i and f_j based on rough dependency
τ (f_i): Pheromone level of feature i

Count[f_i]: Selection count of feature i by different ants in different iterations
q_{rand}, q_{const}: parameters in the range [0,1] to decide on exploration and exploitation
β: parameter to increase the effect of heuristic information in selecting features
ρ: pheromone evaporation rate
N_{ant}: Number of ants.

3.5.2.2. Subset Generation using cosine similarity measure (ACO-CSM).

Algorithm ACO-CSM (*Dataset*)
begin

1. For all the features in *Dataset* compute cosine similarity $sim(f_i,f_j)$ between features using Eq. (2);
2. Set the heuristic information $\eta(f_i,f_j) \leftarrow 1/sim(f_i,f_j)$;
3. For each feature in *Dataset* do
4. Assign initial pheromone level, $\tau[f] = 1/n$
5. Set its selection count, *Count*[*f*] to 0.
6. EndFor
7. Set $q_{const} \leftarrow 0.7$ and $\beta \leftarrow 1$; /* Parameters to decide exploration or exploitation */
8. Set $\rho \leftarrow 0.2$; /* Initialization of pheromone evaporation rate */
9. Set $N_{ant} \leftarrow n$; /* Number of ants equal to the number of features in *Dataset* */
10. For k $\leftarrow$ 1 to N_{max_iter} do /* Repeat subset selection for N_{max_iter} iterations */
11. For i $\leftarrow$ 1 to N_{ant} do /* Every ant selects n_{max} features into the subset */
12. For j $\leftarrow$ 1 to n_{max} do
13. Generate q_{rand} ϵ [0,1];
14. If $q_{rand} \leq q_{const}$ /* Exploitation */
15. Select a feature *f* that is not yet added to the subset using Eq. (9).
16. Else /* Exploration */
17. Select a feature *f* using transition probability given in Eq. (11).
18. EndIf
19. Increment *Count*[*f*]; /* Count of the selected feature is increased by 1*/
20. EndFor /* index j */
21. EndFor /* index i */
22. For each feature *i* in *Dataset* do
23. $\tau_i(k+1) \leftarrow (1-\rho).\tau_i(k) + \frac{Count[i]}{\sum_{j=1}^{n} Count[j]}$; /* Pheromone updation */
24. EndFor
25. EndFor /* index k */
26. Sort *Count* in descending order and select the topmost 'n_{max}' features to build the subset
27. End ACO-CSM

Output: Feature Subset Set_1

3.5.2.3. Subset Generation using rough dependency measure (ACO-RDM).

The ACO_RDM subset generation algorithm is presented briefly in this section. The heuristic information between two attributes is computed using rough dependency measure as given in Eq. (8). The heuristic information used in this algorithm is proportional to the degree of dependency. The algorithm parameters such as initial pheromone level and evaporation rate are assumed to be the same as ACO-CSM algorithm. Every ant iteratively adds a feature to its feature subset by either exploitation or exploration using Eqs. (10) and (12). The trade-off between exploration and exploitation is balanced by the parameters q_{rand} and q_{const}. At the end of the maximum number of iterations, each ant corresponds to a feasible solution (feature subset). The feature subset is then used for evaluating the performance of the classification approaches.

Output: Feature Subset Set_2

3.6. Feature database

The feature vectors of the ROIs present in the training set are pruned according to the features selected by ACO-CSM and ACO-RDM algorithms. The resultant sets are stored as two relations in this database.

3.7. Classification subsystem

The selected feature subsets are used to train SVM and Naive Bayes classifiers independently. Training is carried out using tenfold cross validation.

Input: Feature subsets Set _1 and Set_2

Process:

Step 1: Train SVM and Naive Bayes classifiers using feature subset Set_1 selected by ACO-CSM algorithm.

Step 2: Train SVM and Naive Bayes classifiers using feature subset Set_2 selected by ACO-RDM algorithm.

Step 3: Evaluate the classifiers obtained in the previous steps and identify the classifier model that yields the highest accuracy as the best trained classifier for this system.

Step 4: Identify the feature subset used by the best trained classifier for the diagnosis of the disease and store it in feature set.

Output: Trained classifier

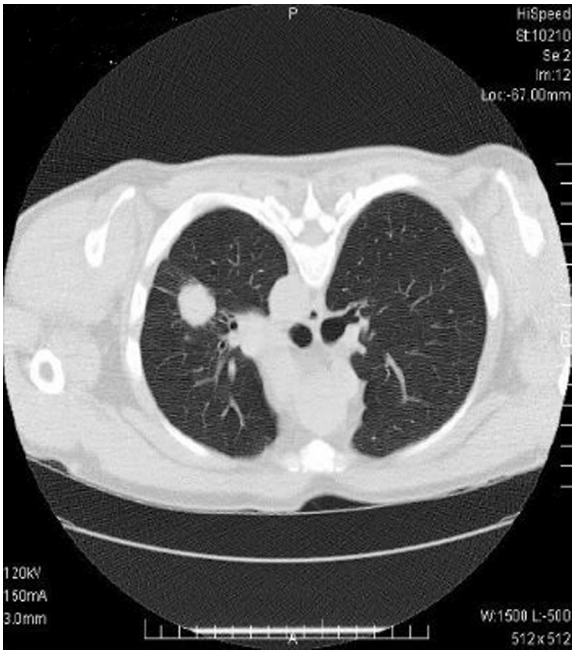

Figure 2a Input image.

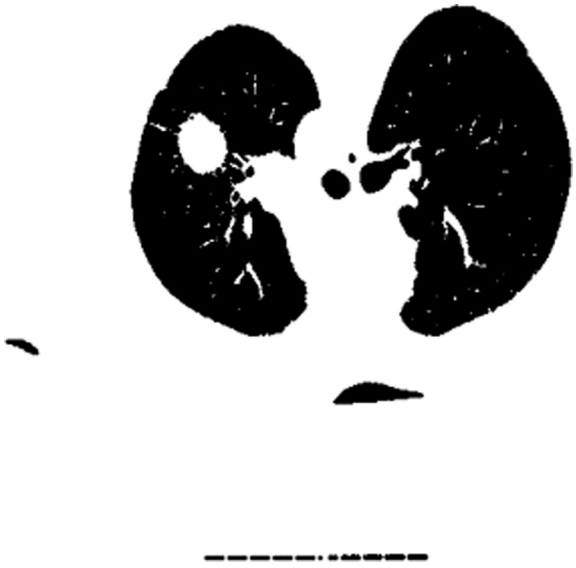

Figure 2b Segmented output.

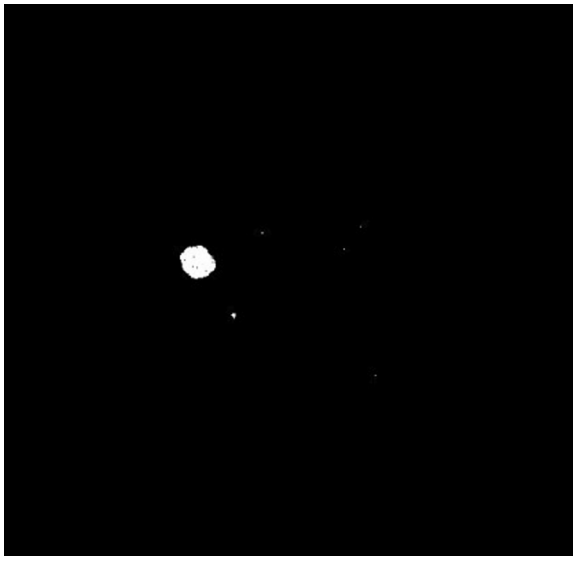

Figure 2c Extracted ROIs.

3.8. Diagnosis phase

Input: Test CT slice

Process:

Step 1: Segment the lung tissues from the input CT slice using the steps given in Section 3.1.

Step 2: Extract the ROIs from the segmented lung as given in Section 3.2.

Step 3: Extract only the selected features from the ROIs by referring to the features present in feature set.

Step 4: Present the extracted features to the trained classifier model to obtain the classification results.

Output: Diagnostic results.

4. Results and discussions

The dataset used for experimentation is CT scan slices of patients affected with pulmonary hamartoma and lung cancer. Lung fields are segmented from 300 lung CT slices and 390 nodules of all sizes are extracted and given to a radiologist for labeling. Of these, 181 nodules are hamartoma nodules and the remaining nodules are cancerous. The system is trained and tested using tenfold cross validation. Two input CT slices containing Hamartoma, their segmentation outputs and their ROIs are shown in Figs. 2a, 2b, 2c, 3a, 3b, 3c respectively.

In this work, the feature selection algorithms are run for 25 times with 144 ants, as the number of ants equals the number of features extracted from the ROIs. The parameters ρ, β and q_{const} are set to 0.2, 1 and 0.7 respectively [25]. The algorithms are run for different values of 'n_{max}' ranging from 20 to 70 [25]. Better results are obtained when 'n_{max}' is set to a value in the

Figure 3a Input image.

Figure 3b Segmented output.

Figure 3c Extracted ROIs.

range 30–46. The performance measures [18] are computed using equations (13) through (16) and are given in Tables 1, 2, 3 and 4 respectively, where TP is the number of hamartoma nodules identified correctly by the system, FP is the number of cancer nodules labeled as hamartoma nodules, TN is the number of cancer nodules identified correctly by the system and FN is the number of hamartoma nodules labeled by the system as cancer nodules.

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \tag{13}$$

$$Specificity = \left(\frac{TN}{TN + FP}\right) \tag{14}$$

$$Precision = \left(\frac{TP}{TP + FP}\right) \tag{15}$$

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) \tag{16}$$

The feature subsets obtained from ACO_CSM and ACO_RDM algorithms are used with J48 decision tree classifier also and the results obtained are given in tables 5 and 6 respectively.

From the tables it can be inferred that, SVM classifier trained with the 38 features selected using ACO_RDM yielded a maximum accuracy of 94.36% whereas SVM trained with 38 features selected from ACO_CSM yielded an accuracy of 85.64%. Using the 38 features selected by ACO-RDM, Naive Bayes and decision tree classifiers yielded a maximum accuracy of 91.02% and 90% respectively. With the 38 features selected by ACO-CSM Naive Bayes and decision tree classifiers yielded only 83.07% and 84.87% respectively. It can be inferred that ACO_RDM feature selection algorithm performed better in all the cases, as rough dependency measure handles the indiscernibility that exists between the features more efficiently than cosine similarity measure. Rough dependency measure finds the informative features based on their dependency to the tar-

Table 1 Performance of SVM classifier when ACO-CSM is used.

No. of features to be selected 'n_{max}'	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
46	88.46	91.16	86.12	85.05
42	87.18	89.5	85.17	83.94
38	85.64	88.39	83.25	82.05
34	84.32	85.56	83.20	81.59
30	83.33	85.64	81.33	79.89

Table 2 Performance of NB Classifier when ACO-CSM is used.

No. of features to be selected 'n_{max}'	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
46	85.64	88.39	83.25	82.05
42	84.35	85.64	83.25	81.59
38	83.07	84.53	81.81	80.10
34	81.59	82.87	80.48	78.53
30	80.51	81.77	79.43	77.49

Table 3 Performance of SVM classifier when ACO-RDM is used.

No. of features to be selected 'n_{max}'	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
46	93.11	95.08	91.39	90.63
42	94.36	96.69	92.35	91.6
38	94.36	96.69	92.35	91.6
34	93.07	95.58	90.91	90.10
30	92.3	93.75	91.30	88.24

Table 4 Performance of NB classifier when ACO-RDM is used.

No. of features to be selected 'n_{max}'	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
46	91.002	93.28	87.98	87.24
42	91.02	93.92	88.52	87.63
38	91.02	93.92	88.52	87.63
34	90.25	92.26	88.52	87.43
30	88.83	89.18	88.52	87.30

Table 5 Performance of J48 Decision Tree classifier when ACO_CSM is used.

No. of features to be selected 'n_{max}'	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
46	86.41	88.39	84.68	83.33
42	85.38	87.29	83.73	82.29
38	84.87	87.29	82.77	81.44
34	83.32	86.56	82.20	81.04
30	82.05	82.87	81.33	79.36

Table 6 Performance of J48 Decision Tree classifier when ACO_RDM is used.

No. of features to be selected 'n_{max}'	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
46	92.31	93.92	90.9	89.94
42	91.02	93.92	88.52	87.62
38	90.00	92.59	87.56	87.5
34	89.74	92.27	87.55	86.5
30	88.97	91.71	86.6	85.56

get class label. Also the uncertainty (unclear boundary values) present in the medical image datasets is approximated effectively by rough sets, which in turn improves the classification accuracy. On the other hand, cosine similarity measure considers only the orientation between the features. If two features have the cosine similarity value as 1, then they are said to be in same orientation. Two features at 90^{o} orientation have a similarity value 0. Features with low similarity value are chosen to form the feature subsets. Hence features selected using rough dependency measure with ACO improved the accuracy in diagnosis.

5. Conclusion and future work

In this work, a computer diagnosis system to detect Hamartoma nodules from CT scan images is proposed. The features selected by ant colony optimization algorithm with cosine similarity and rough dependency measures are used to classify the nodules using NB and SVM classifiers. The selected features are also used with J48 decision tree classifier. From the results, it can be seen that SVM classifier yielded better results when rough dependency measure is used with ant colony optimization to select features. The performance of the classifier model can be improved further, if the segmentation algorithm is able to detect the nodules even at periphery.

References

[1] B. Amini, S.Y. Huang, J. Tsai, M.F. Benveniste, H.H. Robledo, E.Y. Lee, Primary lung and large airway neoplasms in children: current imaging evaluation with multi-detector computed tomography, Radiol. Clin. North Am. 51 (4) (2013) 637–657.

[2] B. Trotman-Dickenson, Cystic lung disease: achieving a radiologic diagnosis, Eur. J. Radiol. 83 (2014) 39–46.

[3] R.I. Whyte, J.S. Donington, Hamartomas of the lung, Semin. Thorac. Cardiovasc. Surg. 15 (3) (2003) 301–304.

[4] T. Umashankar, A.K. Devadas, G. Ravichandra, P.J. Yaranal, Pulmonary hamartoma: cytological study of a case and literature review, J. Cytol. 29 (4) (2012) 261–263.

[5] K. Furuya, K. Yasumori, S. Takeo, I. Sakino, N. Uesugi, S. Momosaki, T. Muranaka, C.T. Lung, Part 1, Mimickers of lung cancer—spectrum of CT Findings with pathologic correlation, Am. J. Roentgenol. 199 (2012) 454–463.

[6] S.S. Siegelman, N.F. Khouri, W.W.J. Scott, Pulmonary hamartoma: CT findings, Radiology 160 (1986) 313–317.

[7] Bateson, Relation between intrapulmonary and endobronchial cartilage-containing tumours (so called hamartomata), Thorax 20 (1965) 447–461.

[8] T.E. King, K.L. Christopher, M.I. Schwarz, Multiple pulmonary chondromatous hamartomas, Hum. Pathol. 13 (1982) 496–497.

[9] S. Jacob, D. Mohapatra, M. Verghese, Massive chondroid hamartoma of the lung clinically masquerading as bronchogenic carcinoma, Indian J. Pathol. Microbiol. 51 (1) (2008) 61–62.

[10] A. Halvani, H.R.J. Darjani, S. Taghipour, Endobronchial chondroid hamartoma, Tanaffos 6 (3) (2007) 68–70.

[11] M. Shiau, E. Portnoy, S.M. Garay, Management of Solitary Pulmonary Nodules, Clinically Oriented Pulmonary Imaging, in: J.P. Kanne (Ed.), 2012, pp. 19–27.

[12] B. Lazovic, R. Jakovic, S. Dubajic, Z. Gataric, Pulmonary hamartoma – case report and review of literature, Arch. Oncol. 19 (1–2) (2011) 37–38.

[13] F. Li, M. Aoyama, J. Shiraishi, H. Abe, Q. Li, Q. Suzuki, R. Engelmann, S. Sone, H. MacMahon, K. Doi, Radiologists' performance for differentiating benign from malignant lung nodules on high resolution CT using computer estimated likelihood of malignancy, Am. J. Roentgenol. 183 (2004) 1209–1215.

[14] K. Awai, K. Murao, A. Ozawa, M. Komi, H. Hayakawa, S. Hori, Y. Nishimura, Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance, Radiology 230 (2004) 347–352.
[15] D.S. Elizabeth, H.K. Nehemiah, C.S.R. Raj, A. Kannan, A novel segmentation approach for improving diagnostic accuracy of CAD systems for detecting lung cancer from chest computed tomography images, ACM J. Data Inform. Qual. 3 (2012), article 4.
[16] J. Shiraishi, H. Abe, F. Li, R. Engelmann, H. MacMahon, K. Doi, Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists performance – initial experience, Radiology 227 (2003) 469–474.
[17] M.L. Giger, K. Doi, H. MacMahon, Image feature analysis and computer-aided diagnosis in digital radiography. III. Automated detection of nodules in peripheral lung fields, Med. Phys. 15 (1988) 158–166.
[18] H. Han, L. Li, F. Han, B. Song, W. Moore, Z. Liang, Fast and adaptive detection of pulmonary nodules in thoracic CT images using a hierarchical vector quantization scheme, IEEE J. Biomed. Health Infor. 19 (2) (2015) 648–659.
[19] D.S. Elizabeth, H.K. Nehemiah, C.S.R. Raj, A. Kannan, Computer-aided diagnosis of lung cancer based on analysis of the significant slice of chest computed tomography image, IET Image Proc. 6 (2012) 697–705.
[20] W.J. Choi, T.S. Choi, Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images, Inf. Sci. 212 (2012) 57–78.
[21] T. Messay, C. Russell, S.K..Rogers. Hardie, A new computationally efficient CAD system for pulmonary nodule detection in CT imagery, Med. Image Anal. 14 (2010) 390–406.
[22] L. Boroczky, L. Zhao, K.P. Lee, Feature subset selection for improving the performance of false positive reduction in lung nodule CAD, IEEE Trans. Inf Technol. Biomed. 10 (3) (2006) 504–511.
[23] K. B Nahato, H.K. Nehemiah, A. Kannan, Knowledge mining from clinical datasets using rough sets and backpropagation neural network, Comput. Math. Methods Med. (2015). Article ID 460189.
[24] B. Chen, L. Chen, Y. Chen, Efficient ant colony optimization for image feature selection, Signal Process. 93 (2013) 1566–1576.
[25] S. Tabakhi, P. Moradi, F. Akhlaghian, An unsupervised feature selection algorithm based on ant colony optimization, Eng. Appl. Artif. Intell. 32 (2014) 112–123.
[26] X. Liu, L. Ma, L. Song, Y. Zhao, X. Zhao, C. Zhou, Recognizing common CT imaging signs of lung diseases through a new feature selection method based on fisher criterion and genetic optimization, IEEE J. Biomed. Health Infor. 19 (2015) 635–646.
[27] J.J. Christopher, H.K. Nehemiah, A. Kannan, A clinical decision support system for diagnosis of allergic rhinitis based on intradermal skin tests, Comput. Biol. Med. 65 (2015) 76–84.
[28] I. Sluimer, A. Schilham, M. Prokop, B.V. Ginneken, Computer analysis of computed tomography scans of the lung: a survey, IEEE Trans. Med. Imag. 25 (2006) 385–405.
[29] V. Haleh, D.J. Kenneth, Genetic algorithms as a tool for feature selection in machine learning, Artif. Intell. (1992) 102–109.
[30] L.F. Chen, C.T. Su, K.H. Chen, P.C. Wang, Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis, Int. J. Neural Comput. Appl. 21 (2012) 2087–2096.
[31] D.S. Elizabeth, A. Kannan, H.K. Nehemiah, Computer aided diagnosis system for the detection of bronchiectasis in chest computed tomography images, Int. J. Imag. Syst. Technol. 19 (2009) 290–298.
[32] A. El-Bazl, A.A. Farag, R. Falk, R.L. Rocca, Automatic identification of lung abnormalities in chest spiral CT scans, Proc. Acoust. Speech Signal Process. 2 (2003) 261–264.
[33] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[34] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 3 (1973) 610–621.
[35] X. Tang, Texture information in run-length matrices, IEEE Trans. Image Process. 7 (1998) 1602–1609.
[36] J. Han, M. Kamber, Data Mining: Concepts and Techniques, second ed., The Morgan Kaufmann Series, 2006.
[37] M. Dorigo, L.M. Gambardella, Ant colony system: a cooperative learning approach to the traveling salesman problem, IEEE Trans. Evol. Comput. 1 (1) (1997) 53–56.
[38] Z. Pawlak, Rough sets, Int. J. Comput. Inform. Sci. 11 (5) (1982) 341–356.
[39] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Addison Wesley, Boston, 2006.