Journal of Industrial and Systems Engineering
Vol. 11, Special issue: 14th International Industrial Engineering Conference
Summer (July) 2018, pp. 163-175

(IIEC 2018)
TEHRAN, IRAN

Detection of lung cancer using CT images based on novel PSO clustering

Fatemeh Sadeghi[1] , Abbas Ahmadi[1*]

[1]*Department of Industrial Engineering and Management Systems, Amirkabir University of Technology,Thehran,Iran*
f.sadeghi2007@yahoo.com, abbas.ahmadi@aut.ac.ir

Abstract
Lung cancer is one of the most dangerous diseases that cause a large number of deaths. Early detection and analysis can be very helpful for successful treatment. Image segmentation plays a key role in the early detection and diagnosis of lung cancer. *K*-means algorithm and classic PSO clustering are the most common methods for segmentation that have poor outputs. In this article, we propose a new that of *K*-means and classic PSO clustering. The obtained results show that the new PSO clustering has better results as compared to the other methods. Comparison between the proposed method and classic PSO, in terms of fitness function and convergence of fitness function indicate that the proposed method is more effective in detecting lung cancer.
Keywords: Lung cancer, image clustering, PSO clustering

1-Introduction

In medical imaging, detection of lung cancer is one of the most challenging tasks. Many studies have been applied to develop optimal method in this area. Clustering is one of the most common techniques in image segmentation and especially in medical imaging and tumor detection. One of the major reasons for non-accidental death is cancer. It has been proved that lung cancer is the topmost cause of cancer death in men and women worldwide. The general detection of lung cancer is poor because detectors are not able to find the disease until it is at an advanced stage. Studies have shown that early detection can reduce the risk of death. Detection of lung tumor is performed by various imaging modalities and computed Tomography (CT) with low cost and optimal quality and robustness is the best choice (Nithila and Kumar, 2017). *K*-means is one of the popular clustering algorithms that can be used for color-based segmentation of brain MR images and tracking brain tumors (Wu et al. 2007). As you may know tracking the tumor in a color image is much easier than the gray scale image. Juang and Wu (2010) have used *K*-means algorithm to segment color-converted brain MR images and to track brain tumors. However, the outcome of *K*-means algorithm cannot detect a spherical image sufficiently. *K*-means algorithm can only detect clusters with the spherical shapes and structures.

PSO algorithm can find spherical shapes and produce better outputs. Cui et al. used PSO algorithm to find optimal centers of clusters in document clustering. Ahmadi et al., (2010) showed that PSO algorithm generate better performance in clustering tasks.

*Corresponding author
ISSN: 1735-8272,

Omran et al., (2005) used PSO in image clustering and improved the fitness function to track tumors in MR images.

They changed the fitness function so that it can minimize the intra-cluster distance and quantization error and simultaneously maximize the distance between clusters. Afshar et al., (2016) used GK algorithm to cluster the CT images but they have to use fuzzy concept and define membership functions. In this article, we use PSO algorithm to track the tumors in CT images. We modify the fitness function by re-definition of the intra-distance. Moreover, we employ image structures such as intensity to detect the true shape of tumors. Also, many studies use numerical data such as UCI data set (Wisconsin breast cancer data set), but in this article, we use CT images that are common and available. In order to better perform of algorithm, we separate lung from background by snake optimization method. This method with internal and external energy terms without effecting general framework can be changed (Akgul and Kambhamettu, 2003).

The rest of the article is organized as follows. Section 2 explains the related concepts. In section3, the proposed tumor detection approach is provided. In section 4, the experimental results are given. Finally, in section 5, the conclusion and future research directions are highlighted.

2-Data clustering for image segmentation

There are two main approaches to image classification: supervised and unsupervised ones. In the supervised approach, the number and the characteristics (mean and variance) of the classes in the image are known in advance and they are used in training step followed by a classification step. There are several popular supervised algorithms such as the minimum-distance-to-mean and Gaussian maximum likelihood classifiers (Vafaie and Jong, 1992). For unsupervised approaches, classes are unknown and the algorithm starts by partitioning the image into groups (or clusters), according to a similarity measure, which can be compared by an analyst to available reference data (Lillesand et al. 2014). Accordingly, the unsupervised classification is also referred to as a clustering problem. In general, the unsupervised approach has several advantages over the supervised approach (Davies, 1997) namely:

- For unsupervised approaches, there is no need for an analyst to specify in advance all the classes in the image data set. The clustering algorithm will automatically find distinct classes, which dramatically reduce the work of the analyst.
- The characteristics of the objects being classified can vary with time; the unsupervised approach is an excellent way to monitor these changes.
- Some characteristics of object may not be known in advance. Unsupervised approaches will automatically flag these characteristics (Omran et al. 2005).

The focus of this article is in the unsupervised approach. Next, we explain unsupervised well-known *K*-means algorithm.

In the area of image segmentation, each image is considered as a data set, and each pixel is a data point. Different features can be considered for each pixel such as intensity and coordinates. Using clustering, pixels belonging to the same object will probably go into the same cluster, so objects can be separated. *K*-means clustering explained in the following subsection can only detect clusters with the same shapes (Afshar et al. 2016).

2-1-K-meansclustering

One of the popular clustering techniques that is easy to understand. The algorithm aims to find the best centroids for the clusters and make the squared error between them and data points the lowest amount. The equation (1) shows the associated objective function:

$$J = \sum_{k=1}^{K} \sum_{H_i \epsilon C_k} \| H_i - M_k \|^2. \tag{1}$$

In this equation *K* indicates the number of clusters M_k is the center of cluster C_k and H_i shows the i^{th} data point. Algorithm (1) defines the steps of K-means algorithm (Jain, 2010).

Table 1. Algorithm *K*-means

Algorithm 1.*K*-means clustering
1- Determine initial centroids randomly. Repeat step 2 and 3 until a pre-defined criterion is met.
2- Assign each point to its nearest cluster.
3- Update new centroids

The *K*-means clustering has many features that make it a popular approach. It is easy to understand and easily applicable. It also converges to a final solution very quickly. The *K*-means algorithm is highly sensitive to initial solution and it may converge to local optimal solutions (Ahmadi et al. 2010).

3-Proposed lung cancer detection approach

In this article in order to get the best accuracy for lung cancer detection, at first we separate the lung from background, because timorous pixels are similar to those of the background. Hence, we use snake optimization method to separate the lung from the background. Then, we employ PSO algorithm by the improved fitness function to segment the underlying lung images to detect tumors successfully. The main steps of the proposed approach are summarized as follows:

1- Separate lung from the background using snake optimization method
2- For all the images obtained in step 1, execute the PSO algorithm
3- Execute the PSO algorithm with the improved fitness function.

3-1-Data structure

Each pixel uses a data set and we define four features for them which are intensity, location and mean of the intensity of the neighbouring pixels, respectively. Data structure of the pixel (x,y) is given below:

$$H(x,y) = \big(I(x,y), x, y, N(x,y)\big). \tag{2}$$

According to the relation, x and y are the x coordinate and y coordinate. I(x,y) and N(x,y) indicate the intensity and the mean of the intensity of the neighbouring pixels of pixel (x,y), respectively.

3-2- Snake optimization method

As previously mentioned, we separate the lung from background by using snake optimization method. The reason for choosing this method is that the lung contour of one section can be used as an initial contour for other sections (Afshar et al. 2016). A snake model is a curve which can evolve from an initial position to object boundaries. This method needs user interaction; thus it is assumed as a semi-automatic method (Jacob and Blu, 2004). There is variety of snake models, in this approach the point-based snake model used and a contour is considered as a set of discrete points in such models.

The energy function which should be minimized is a combination of internal and external forces. Internal energy is related to the shape of contour. However, external energy depends on image features. Assuming x(s) and y(s), as coordinates in contour direction, and defining s as the snake variable between zero and one, a contour is described using the variable V(s) as follows (Sonka et al. 2014).

$$V(s) = [x(s), y(s)]. \tag{3}$$

The energy function to be minimized is defined as follows:

$$\begin{aligned} E^*_{snake} &= \int_0^1 E_{snake}\big(V(s)\big)ds \\ &= \int_0^1 \Big(E_{int}\big(V(s)\big) + E_{image}\big(V(s)\big) + E_{con}\big(V(s)\big)\Big)ds \end{aligned} \tag{4}$$

In these equations, E_{int} is the internal energy and E_{image} is the image force. E_{con} stands for constant external forces.

Internal energy can be calculated as follows:

$$E_{int} = \alpha(s)|dv/ds|^2 + \beta(s)|d^2v/d^2s|^2 \quad (5)$$

In this equation, coefficients are elasticity and stiffness respectively. The second energy term which is related to image data, is typically defined as a function of image gradient:

$$E_{image} = -|\nabla I(x,y)|^2 \quad (6)$$

I(x,y) is the intensity of the pixel located in (x,y). The reason for putting a negative sign at the beginning of the equation is that image gradient is typically higher in object boundaries and the goal of the minimization energy function is to find object edges.

There are several ways to optimize the objective function mentioned before including gradient decent method, genetic algorithm and dynamic programming. In this article we adopt the Viterbi algorithm which is a popular dynamic programming method (Forney, 1973).

3-3- Particle swarm optimization

Particle swarm optimization (PSO) is a population-based optimization algorithm simulating the social behaviour of birds in a flock (Eberhart and Kennedy, 1995) and (Eberhart et al. 2001).The PSO includes a swarm of solutions that can be candidates as a solution of optimization problem. Each particle plays the role of solution and according to the variables of problem, particles show the dimensional point in the search space. For example, in problem with n variables, the particles are n-dimensional.

The quality and fitness of a particle can be measured by using a fitness function. By doing that, it can quantify how close a particle is to the optimal solution (Omran et al. 2005).

Each particle is flown through the search space, having its position adjusted based on its distance from its own personal best position and the distance from the best particle of swarm (Shi and Eberhart, 1998).

Due to its abilities, PSO has been used in other applications such as classification and clustering (Xiao et al. 2003), (Xiao et al. 2004), (Omran et al. 2005), (Cui et al. 2005), (Cui et al. 2006) and (Holden and Freitas, 2005).

In this article the main use of PSO is the single swarm clustering and the image clustering based on PSO. For the sake of brevity, we just explain the algorithm of single swarm clustering. In the following, we explain the image clustering based on PSO using the modified fitness function.

Table2. Algorithm single swarm clustering

Single swarm clustering algorithm
Initialize a swarm size of n
Repeat
For each particle $i \epsilon [1, \dots, n] do$
Update position and velocity
If $F\big(M_i(t+1)\big) < F\left(M_i^{pb}(t)\right) then$
$M_i^{pb}(t+1) \leftarrow M_i(t+1)$
End if
End for
$M^*(t+1) \leftarrow argmin\{F\left(M_i^{pb}(t)\right) i \in [1, \dots, n]\}$

According to the algorithm each particle is represented as$M_i = (m^{(1)}, \dots m^{(k)})_i$,where $m^{(k)}$ denote the center of cluster k. After defining the fitness function and particle presentation, a single swarm can be used to obtain the solution of the clustering problem. The search commences from an initial population in the solution space and proceeds to find a near-optimal solution (Ahmadi et al. 2010).

3-4-Image clustering

In this section we explain the use of PSO clustering in the medical image specially CT images of lung. The algorithm of image clustering is the same as PSO swarm clustering, but in this study we use the proposed method by Omran et al. (2005) and improved the fitness function of the algorithm. Their proposed method is similar to that of Cui et al (2005 and 2006), but the main difference is how they define the fitness function. They desire to cluster images such that intra-cluster distance between clusters is maximized (Ahmadi et al. 2010). The notations of PSO-based image clustering are given below:

- $N_b$denote the number of spectral classes
- N_p denote the number of image pixels
- N_c denote the number of spectral classes (as provided by the user)
- H_p denote the N_b components of pixel p
- M_j denote the mean of cluster j

Different measures can be used to express the quality of image clustering algorithms. The most general measure of performance is the quantization error, defined as (Omran et al. 2005).

$$U_e = \frac{\sum_{j=1}^{N_c} \left[\sum_{\forall H_p \in C_j} d(H_p, M_j)\right] / \left|C_j\right|}{N_c} \tag{7}$$

$$d\left(H_p, M_j\right) = \sqrt{\sum_{k=1}^{N_b} (H_{pk} - M_{jk})^2} \tag{8}$$

In equation (7), quantization error includes Euclidean distance between centroids and $\left|C_{ij}\right|$is the cardinality of the set $C_{ij}$and number of clusters is shown by N_c.
These two equations used in the improved fitness function, which the description will come in the following:

$$\bar{d}_{max}(H_i, X_i) = max_{j=1,\ldots,N_c}\{\sum_{\forall H_p \in c_{ij}} d(H_p, M_{i,j}) / \left|C_{ij}\right|\} \tag{9}$$

Equation above is the maximum average Euclidean distance of particles to their associated classes, and

$$d_{min}(x_i) = min_{\forall j_1, j_2, j_1 \neq j_2}\{d\left(M_{ij1}, M_{ij2}\right)\} \tag{10}$$

The above equation is the minimum Euclidean distance between any pair of clusters. This fitness function has as objective to simultaneously

- Minimize the intra-distance between pixels and their cluster means ,as quantified by $\bar{d}_{max}(H_i, x_i)$,and
- Maximize the inter-distance between any pair of clusters, as quantified by $d_{min}(x_i)$.

According to these definitions, the fitness function is as follows:

$$f(x_i, H_i) = W_1 \bar{d}_{max}(H_i, x_i) + W_2(H_{max} - d_{min}(x_i)) + W_3 U_{e,i} \tag{11}$$

In this article we change the fitness under consideration the distance between particles centroids are constant so we change it with Euclidean distance between max far point in one cluster and min close point in one cluster and the combination with classic PSO fitness function gets the result that comes in section experimental result. PSO based image clustering by improved fitness function is given in algorithm 3.

Table3. PSO based image clustering by improved fitness function

Algorithm 3. PSO based image clustering by improved fitness function

For k=1 to k_{max}

(a) For each particle i

i. For each pixel H_p

- Calculate d(H_p, M_{ij})for all clusters C_{ij}
- Assign $H_p\ to\ C_{ij}$ where

d$(H_p, M_{ij}) = min_{\forall c=1,\ldots,N_c}\{d(H_p, M_{ic})\}$

ii. calculate the fitness , $f(x_i, H_i) = W_1\bar{d}_{max}(H_i, x_i) + W_2(H_{max} - d_{min}(x_i)) + W_3U_{e,i}$

(b) Find the global best solution $\hat{y}(t)$

(c) Update the cluster centroids using equation (12), (13)

$$V_i(t+1) = WV_i(t) + C_1r_1(t)\big(y_i(t) - x_i(t)\big) + c_2r_2(t)(\hat{y}(t) - x_i(t)) \tag{12}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{13}$$

In this article in order to get the best result we define the fitness function as the first part maximize the distance between far-point from one cluster and the far-point from another cluster and the second part maximize the distance between centroids and the third part as the same, and the equation should be minimized. Also according to the Data structure each pixel has four feature that the first and forth one respectively show the intensity of pixel and the mean of the intensity of the neighbouring pixels of pixel so we can use this two features in calculating the distance in the process of PSO clustering. The impact of these two features makes the algorithm to get the desired result

4-Experimental results

In this article, we use 10 images from Lung image database consortium (LIDC) (Grove et al. 2015). Also, we separate lung from background by using snake optimization method, the results of the method come respectively. As you see, the snake optimization method by separating lung from background has removed the additions especially the points with the same colour of tumor, so it makes the best picture of lung and shows the boundary of lung. After snake optimization, each clustering method, K-mans, classic PSO and PSO with improved fitness function tested on CT images. Numbers of clusters are two and the results are compared based on values of fitness function. For this purpose, 30 particles were trained for 80 iterations,$w = 0.72$,$c_1 = 1.39, c_2 = 1.49$ and for the fitness function $w_1 = 0.2, w_2 = 0.3$, $w_3 = 0.5$. As it can be seen in the table 4,5,6,7 the result of PSO with improved fitness function is better than other approach and can segment the pictures ideally.

Table4. Comparing the result of segmentation of methods

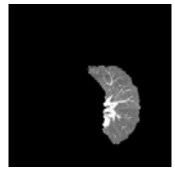
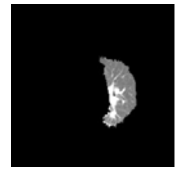
Original image			
Lung separated with snake method			
Tumor segmented with k-means			
Tumor segmented with PSO			
Tumor segmented with proposed PSO			

Table5. Comparing the result of segmentation of methods

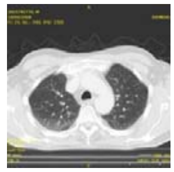
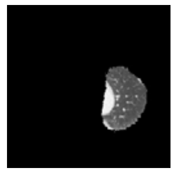
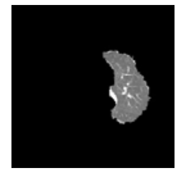
Original image		
Lung separated with snake method		
Tumor segmented with k-means		
Tumor segmented with PSO		
Tumor segmented with proposed PSO		

Table6. Compare the result of segmentation of methods

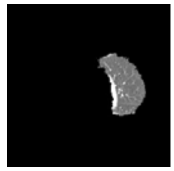
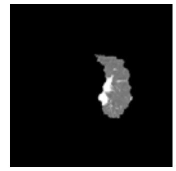
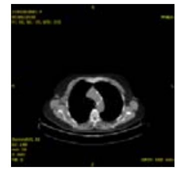
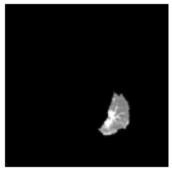
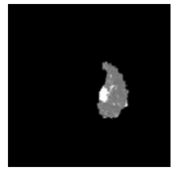
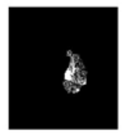
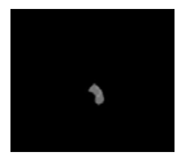
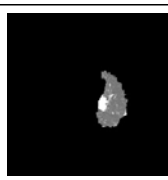
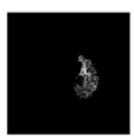
Original image		
Lung separated with snake method		
Tumor segmented with k-means		
Tumor segmented with PSO		
Tumor segmented with proposed PSO		

Table7. Compare the result of segmentation of method

Original image			
Lung separated with snake method			
Tumor segmented with k-means			
Tumor segmented with PSO			
Tumor segmented with proposed PSO			

In the following we show the clusters that made by proposed PSO method in table 8

Table8. compare the result of segmentation of methods

Separate by snake	Proposed PSO cluster1	Proposed PSO cluster2
		

Also we compare the amount of fitness function between the PSO and PSO with proposed fitness function for4 iteration in table 9.

Table9.comparing the amount of fitness function

	Iteration30	Iteration50	Iteration80	Iteration100
Fitness function PSO	12.024	8.6234	8.5532	8.2231
Fitness function proposed PSO	8.8166	8.1155	7.5642	7.5600

As regards the fitness function is minimizing, we can see that the proposed PSO gets the lower fitness function. Figure 1 and 2 show the convergence of classic PSO and he proposed PSO. In figure 2, we can see that for iterations 80 to 100, the fitness function is converged to 7.5600.

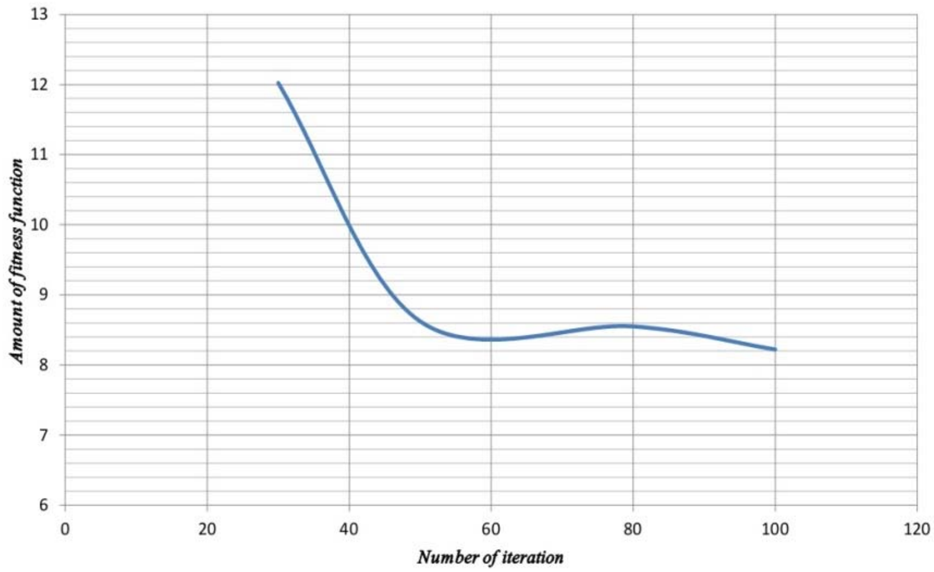


Fig1. The convergence of fitness function for classic PSO clustering

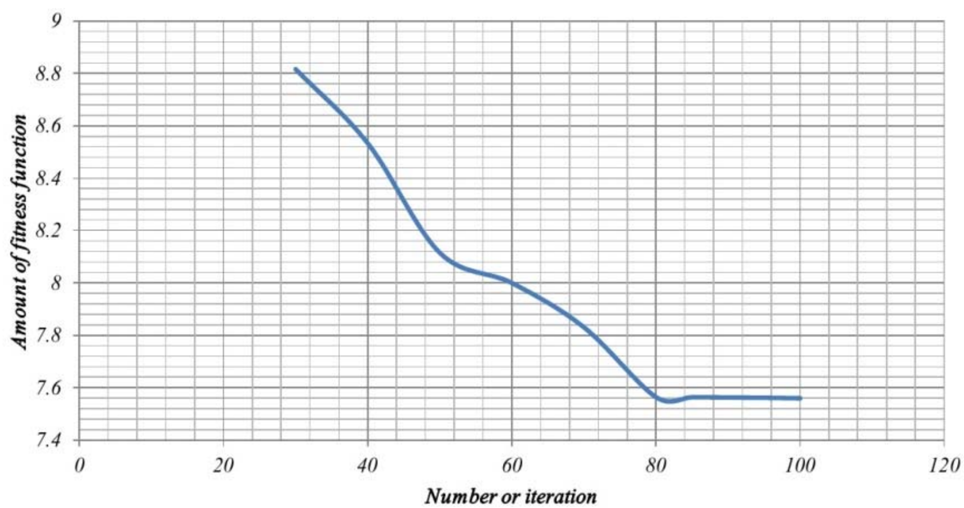


Fig2. The convergence of fitness function for the proposed PSO clustering

5-Conclusions and future work

Early detection of cancer plays important role in treating lung cancer successfully. Here, a novel PSO based clustering algorithm was proposed to detect lung cancer using CT images by introducing a new fitness functions. In this article, we used 10 CT images and separated them from back ground by using snake optimization method. Then, we clustered the image by PSO based clustering by improved fitness function. Then, we compared the result of segmentation by the proposed PSO method with classic PSO and *K*-means with 2 numbers of clusters. As for future work; one can use Mahalanobis distance instead of Euclidean distance in PSO algorithm. Moreover, we hope that we can use the result of tumor segmentation to get the 3D picture of tumor so we improve our imagination about the shape of tumor to treat the lung cancer correctly.

References

Ahmadi, A., Karray, F., & Kamel, M. S. (2010). Flocking based approach for data clustering. *Natural computing*, *9*(3), 767-791.

Afshar, P., Ahmadi, A., & Zarandi, M. F. (2016, July). Lung tumor area recognition in CT images based on Gustafson-Kessel clustering. In *Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on* (pp. 2302-2308). IEEE.

Akgul, Y. S., & Kambhamettu, C. (2003). A coarse-to-fine deformable contour optimization framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(2), 174-186.

Cui, X., Gao, J., & Potok, T. E. (2006). A flocking based algorithm for document clustering analysis. *Journal of systems architecture*, *52*(8-9), 505-515.

Cui, X., Potok, T. E., & Palathingal, P. (2005, June). Document clustering using particle swarm optimization. In *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE* (pp. 185-191). IEEE.

Eberhart, R. C., Shi, Y., & Kennedy, J. (2001). *Swarm Intelligence (Morgan Kaufmann series in evolutionary computation)*. Morgan Kaufmann Publishers.

Eberhart, R., & Kennedy, J. (1995, October). A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on* (pp. 39-43). IEEE.
,Perth,Australia,1942-1948

Davies, E. R. (1997). Laws' texture energy in TEXTURE. *Machine vision: theory, algorithms, practicalities*.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268-278

Grove, O., Berglund, A. E., Schabath, M. B., Aerts, H. J., Dekker, A., Wang, H., ... & Eikman, E. (2015). Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PloS one*, *10*(3), e0118261.

Holden, N., & Freitas, A. A. (2005, June). A hybrid particle swarm/ant colony algorithm for the classification of hierarchical biological data. In *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE* (pp. 100-107). IEEE.

Jacob, M., Blu, T., & Unser, M. (2004). Efficient energies and algorithms for parametric snakes. *IEEE transactions on image processing*, *13*(9), 1231-1244.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

Juang, L. H., & Wu, M. N. (2010). MRI brain lesion image detection based on color-converted K-means clustering segmentation. *Measurement*, *43*(7), 941-949.

Lillesand, T., Kiefer, R. W., & Chipman, J. (2014). *Remote sensing and image interpretation*. John Wiley & Sons.

Nithila, E. E., & Kumar, S. S. (2017). Automatic detection of solitary pulmonary nodules using swarm intelligence optimized neural networks on CT images. *Engineering science and technology, an international journal*, *20*(3), 1192-1202.

Omran, M., Engelbrecht, A. P., & Salman, A. (2005). Particle swarm optimization method for image clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, *19*(03), 297-321.

Shi, Y., & Eberhart, R. (1998, May). A modified particle swarm optimizer. In *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on* (pp. 69-73). IEEE.

Sonka, M., Hlavac, V., & Boyle, R. (2014). *Image processing, analysis, and machine vision*. Cengage Learning.

Vafaie, H., & De Jong, K. (1992, November). Genetic algorithms as a tool for feature selection in machine learning. In *Tools with Artificial Intelligence, 1992. TAI'92, Proceedings., Fourth International Conference on* (pp. 200-203). IEEE.

Xiao, X., Dow, E. R., Eberhart, R., Ben Miled, Z., & Oppelt, R. J. (2004). A hybrid self-organizing maps and particle swarm optimization approach. *Concurrency and computation: practice and experience*, *16*(9), 895-915.

Xiao, X., Dow, E. R., Eberhart, R., Miled, Z. B., & Oppelt, R. J. (2003, April). Gene clustering using self-organizing maps and particle swarm optimization. In *Parallel and Distributed Processing Symposium, 2003. Proceedings. International* (pp. 10-pp). IEEE.

Wu, M. N., Lin, C. C., & Chang, C. C. (2007, November). Brain tumor detection using color-based k-means clustering segmentation. In *iih-msp* (pp. 245-250). IEEE.