Using Soft Clustering and Dempster-Shafer Method for Liver Cancer Diagnosis

Babak Fouladi Nia[a1], Abbas Karimi[b1], Faraneh Zarafshan[c2], Manochehr Kazemi[d3]
[1]Department of Computer Engineering, Arak Branch, Islamic Azad University, Arak, Iran
[2]Department of Computer Engineering, Ashtian Branch, Islamic Azad University, Ashtian, Iran
[3]Department of Mathematics, Ashtian Branch, Islamic Azad University, Ashtian, Iran

Abstract
One of the most practical ways to improve the accuracy is to use data fusion methods in data mining techniques. The use of clustering techniques by considering the theories that cover the uncertainty in these techniques can lead to solving clustering problems, especially in cancer diagnosis. One of the common cancers that cause many deaths is liver cancer. Unfortunately, in recent years, the probability of getting this cancer has increased greatly. To identify the liver tumor, image segmentation technique is performed on CT scan images. Successful treatment of liver cancer requires accurate diagnosis of liver abnormalities. To achieve this goal, techniques based on automatic and semi-automatic detection are effective. The method proposed in this article has high accuracy and convergence speed.
Keywords: Liver cancer, Dempster-shafer method, Evidence theory, Fuzzy system, K-Means clustering, Medical image segmentation

Tob Regul Sci. ™ 2022;8(1): 3179-3197
DOI: doi.org/10.18001/TRS.8.1.243

Introduction

Incorrect methods in human lifestyle, consumption of non-organic foods and increased stress caused by daily work in stressful environments have led to an increase in the number of patients with incurable diseases. Cancers and tumors are the majority of these diseases. Today, with the advancement of medical science, timely and correct diagnosis of all types of diseases, a wide range

[a] DrFouladi@yahoo.com
https://orcid.org/0000-0003-3496-6121
[b] Corresponding author
akarimi@iau-arak.ac.ir
https://orcid.org/0000-0003-0120-2803
[c]fzarafshan@aiau.ac.ir
https://orcid.org/0000-0003-0327-5176
[d]m.kazemi@aiau.ac.ir
https://orcid.org/0000-0001-8392-6690

of these diseases can be prevented and even treated. Human errors and mistakes are always common in the field of treatment and have always caused such problems. Sometimes mistakes are irreparable and lead to the death of the patient [1]. For this reason, eliminating human errors has been prioritized by researchers. In the past decade, medical expert systems that combine artificial intelligence and machine vision have been introduced to minimize human errors. Medical equipment is divided into two types, which are: diagnosis and treatment. Most of the first type (diagnostic) devices use imaging [1]. X-rays can be passed through the patient's body or sound reflection and other things can be used to form these images. In recent years, the manufacturing companies of this equipment have provided diagnostic software to the market. Analysis of medical images has a high computational complexity and is divided into automatic and semi-automatic systems. Image analysis methods are used to detect abnormalities in body organs by scanning images. To identify and diagnose tumors in liver cancer using CAD[2], liver texture segmentation is used as a prerequisite step in medical imaging [2]. Due to the similarity of the structure of the edges of the liver and other nearby organs such as the heart, stomach and muscles, segmentation of the liver is very difficult and has been of great interest to researchers in recent research. Also, the low contrast between the liver and the adjacent organs creates problems in accurate segmentation. Also, in cases where the image of the liver is presented in different dimensions, problems arise in diagnosis and segmentation. Among the standard tools in the diagnosis of liver diseases such as cirrhosis, cancer and failure, imaging methods such as CT[3], MRI[4] and PET[5] can be mentioned. Among the mentioned methods, specialists usually prefer CT scan images, because they provide more comprehensive and accurate information about the liver texture. The main goal of this research is to improve the segmentation accuracy of liver CT scan images. Considering the problems mentioned about the liver texture and adjacent organs, the accurate segmentation of CT scan images requires the use of a method that can correctly differentiate the cancerous, healthy, and ignorance parts of the liver. In this research, a unique method is presented to detect the parts of ignorance. This issue has not been investigated in the articles that have been published so far. The cancer detection system consists of two parts: 1-image segmentation 2- image classification [3]. Liver segmentation steps start with a pre-processing step and continue by dividing it into constitutive regions. There are different techniques for segmenting the liver. Some of these techniques are semi-automatic and some are fully automatic. Semi-automatic techniques require user interaction and the desired area is identified and selected by the user and used for subsequent processes, while in fully automatic techniques segmentation is done without the need for user interaction. Considering that manual segmentation of liver CT scan images is very time-consuming, in this research, techniques based on automatic and semi-automatic segmentation are considered. The structure of the paper is as follows: In Section 2, various techniques available in

[2] Computer-Aided Diagnosis
[3] computed tomography
[4] magnetic resonance imaging
[5] positron emission tomography

the preprocessing and segmentation of liver cancer are discussed. Section 3 describes the techniques of liver cancer diagnosis and also compares these methods with each other. Conclusions are presented in Section 4. In the next section, the existing techniques in the diagnosis of liver cancer are reviewed and the advantages and disadvantages of the methods are discussed.

Literature review

In the following, some previously published articles in this field have been reviewed. For example, it is mentioned in the article [1] that the pre-processing step is a necessary step in order to obtain high accuracy in liver segmentation. CT scan images contain salt-and-pepper noise, which reduces accuracy. In the method presented in [2], three-dimensional anisotropic diffusion method is used as a pre-processing step to remove image noise. Also, this method preserves important parts of the image such as edges, lines or other details that are important for image interpretation. In the research [4], various machine learning algorithms such as KNN[6], MLP[7] neural network, and SVM were used. Among these algorithms, the SVM method has a better performance and has an accuracy of 95.29% in cancer detection. In research [5], a semi-automatic algorithm has been proposed that has an effective process in CT image segmentation. Among the most important advantages of this method, we can mention three cases of reducing decision making for cancer detection in the region, reducing computing time and resistance to noisy images. In research [6], two different algorithms have been used to increase the quality of CT scan images, which are: CLAHE[8] and limited variable histogram equalization CVHE[9]. CLAHE improves the quality of tumor region and detects normal liver in this method. Also, liver cancer is diagnosed by CVHE method. Also, SVM algorithm was used to classify liver CT scan images the results show that the accuracy of this method is 97%. In the article [7], a technique is presented that in the segmentation part of MRF, the process of assigning labels to image pixels is performed. In this way, the pixels that are more similar to each other are labeled with the same value (preprocessing). In Ahmadi et al.'s research [8], liver blood vessel segmentation was performed using a combination of FCM[10] and GA[11] methods on 20 CT scan images, and the accuracy, sensitivity, specificity, and CPU time values were 91%, 83.62%, 94.11% and 27.17 respectively. In the article [9] by Fouladi et al., a hybrid segmentation algorithm using FCM and Cuckoo optimization algorithm was presented and compared with FCM-GA hybrid algorithm. The results of this research showed that the values of sensitivity, specificity and accuracy were 84.24%, 94.30% and 91.36% respectively. A semi-automatic technique consisting of three steps is used in paper [10] to overcome the problems of leakage and over-segmentation, where the input images are pre-processed by a series of techniques to obtain binary images. In the research ([11]), as a pre-processing step, the image

[6] k-nearest neighbors
[7] multilayer perceptron
[8] contrast limited adaptive histogram equalization
[9] Constrained Variational Histogram Equalization
[10] Fuzzy C-means
[11] genetic algorithm

noise is removed and then the edge of the input image is sharpened using a frequency-based edge sharpening technique. In the image segmentation stage, the output of the previous stages is considered as the input of the OPBS[12] method for segmentation. In the research [12], the combination of selecting the seed point and the connected component with the watershed method has been used in order to achieve better accuracy and robustness in CT scan images. The results of this research indicate a high accuracy of 98.6%. In research [13], a new external force technique called GVFOM[13] is used. The model uses the GVF vector as a two-dimensional manifold embedded in the four-dimensional Euclidean space, which leads to the generalization of the Laplace operator in the Euclidean space, and two external force components are used to improve the properties of the GVF snake. In another research in 2021, a CNN[14] method was introduced to segment the liver in abdominal CT images, which has significant accuracy in segmenting the images. In this research, 180 abdominal CT scan images were used for training and validation. Also, researchers have extended their network using a self-supervised contour method [14]. In paper [15], researchers proposed a DAR-net[15] segmentation method to accurately determine the boundaries of liver images. Modeled on the U-Net network, they have proposed a dynamic adaptive pooling strategy based on interpolation optimization to process all features in the pooling context. In this research, the 3DIRCADB dataset was used, and the average DICE score was 96.13%, which was 13.02% higher than the unprocessed prediction result. In the article [16], by adapting and comparing advanced deep learning frameworks studied in different fields, it presents the most efficient deep learning architectures for liver segmentation. These frameworks have been implemented in a commercial software called "LiverVision". Experimental results show that "U-Net" and "SegNet" have been superior to other techniques in terms of time, cost and effectiveness. The paper [17] presented a new architecture called Dense-UNet feature selection, which performs better to separate the liver from abdominal CT images when the target region is small or partitioned. HU[16] values in a range are used to remove irrelevant organs and preprocessed data is used to train the proposed model. According to ground truth, the Dice score ratio can reach more than 94.9%. In the article [18], a technique for automatic segmentation of liver lesions in CT images according to the variety of lesions in terms of contrast, shape, size and location is presented. The proposed technique was evaluated using a set of 131 CT images from the LITS dataset, and MCC[17], sensitivity, specificity, Dice coefficient, VOE[18], and RVD[19] were 83.62, 83.86, 99.96, 82.99, 27.89 %, and 1.69% respectively. In research [19], the developed k-means clustering algorithm is used for liver segmentation. The k-means algorithm is one of the simplest

[12] Outline Preservation Based Segmentation
[13] gradient vector flow over manifold
[14] convolutional neural network
[15] dynamic adaptive residual network
[16] Hounsfield Units
[17] Matthews correlation coefficient
[18] volumetric overlap error
[19] relative volume difference

unsupervised learning algorithms that categorizes a certain image into a certain number of clusters. The disadvantage of this clustering technique is that the cyst region is not correctly extracted. To improve the performance, morphological operators (erosion and dilation) are used in the output of the K-means algorithm. In segmenting cyst regions in liver images, K-means clustering algorithm performs better than region growing algorithm. FCM[20] clustering technique is not very effective for segmenting liver tumors with noisy or outlier points and different clusters of different sizes. To overcome these problems, AFCM[21] clustering method has been used [6]. AFCM modifies similar pixels that are repeated at the center of clusters for defined iterations which makes segmentation more effective. In research [20], MIL[22] method for liver cancer diagnosis using abdominal CT images is presented, which is based on IO[23] and SVD[24] parameters by a hybrid algorithm of PSO[25] and Local optimization is emphasized.

FCM Clustering Method

Clustering algorithms include dividing a set of objects into groups(clusters) in such a way that similar objects are in the same group and are less similar to objects in other groups. Euclidean or Manhattan distance criteria are commonly used to determine the similarity of two sets from the point of view of the distance scale.

With a definite number of branches k and a number of data N, the matrix

$$U_{\mathrm{K*N}} = [\mathrm{u_{ik}}], \mathrm{k} = 1, \dots, \mathrm{C}. \text{ and } \mathrm{i} = 1, \dots, \mathrm{N} \quad (1)$$

Displays the data set, as $\mathrm{u_{ik}}$ describes the $\mathrm{x_i}$ data in the c_k cluster. If $\mathrm{u_{ik}}$ has only two values 1 (belongs to) or 0 (does not belong to) that are set by Boolean membership functions so Clustering is hard, and if $\mathrm{u_{ik}}$ has values between zero and one that is valued by continuous membership functions so clustering is fuzzy. If v_k is the center of the ck cluster then v_k is calculated as follows:

$$v_{\mathrm{k}} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{u_{ik}}\ x_i / \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{u_{ki}}, \mathrm{k} = 1, \dots, \mathrm{C} \quad (2)$$

The most famous hard clustering technique is the K-Means method. In this technique, the u_{ik} membership value must meet the following expression:

$$\forall k, \forall i, u_{ik} = \{0,1\}, \forall i, \sum_{k=1}^{c} u_{ik} = 1, and\ \forall k, 0 < \sum_{i=1}^{N} u_{ki} < N. \quad (3)$$

Using the search of a d_{ki} distance function for example, the Euclidean distance is equal to:

$$d_{ik} = \left|\left|X_i - V_k\right|\right| \quad (4)$$

Then we need to find a $\boldsymbol{U}$ that satisfies the membership constraint in Equation (5) and minimizes the following function.

[20] Fuzzy C-means
[21] an alternative FCM
[22] a Multiple Instance learning
[23] instance optimization
[24] singular value decomposition
[25] particle swarm optimization

$$J(U,V) = \sum_{k=1}^{c} \sum_{x_i \epsilon c_k} d_{ki}^2 . \quad (5)$$

A common way to find $\boldsymbol{U}$ is to use the iterative technique proposed by Lloyd. The FCM clustering method is an extension of the k-Means clustering algorithm. This technique allows data points to belong to more than one cluster, which indicates a degree of membership confidence in each cluster (as opposed to hard membership in k-Means).

u_{ik} membership must meet:

$$\forall k, \forall i, 0 < u_{ik} < 1, \forall i, \sum_{k=1}^{c} u_{ik} = 1, \text{and } \forall k, 0 < \sum_{i=1}^{N} u_{ki} < N. \quad (6)$$

Where $\{\mathbf{v_k}\}_{\mathbf{i=1}}^{\mathbf{c}}$ are the centers or initial instances of the clusters, and the array $\{\mathbf{u_{ki}}\}(= \mathbf{U})$represents a satisfactory matrix clustering:

$$U \in \left\{ u_{ik} \in [0,1] \,\middle|\, \sum_{k=1}^{c} u_{ik} = 1, \forall K \text{ and } 0 < \sum_{i=1}^{N} u_{ik} < N, \forall i \right\} \quad (7)$$

So that $\boldsymbol{m} \geq \mathbf{1}$ controls the degree of fuzzyness. If $\boldsymbol{m}$ is close to 1, then it gives the cluster with the closest center to the point more weight than the other clusters. The smaller $\boldsymbol{m}$ is considered, we will reach the fuzzy state, and vice versa the larger $\boldsymbol{m}$, we will move away from the fuzzy state and approach a hard state. Membership amounts and cluster centers are calculated as follows (Unlike hard membership in k-means):

$$u_{ik}^{(t+1)} = \frac{d_{ik}^{-2/(m-1)}}{\sum_{j=1}^{c} d_{ij}^{-2/(m-1)}} \quad (8)$$

$$c_k^{(t+1)} = \frac{\sum_{i=1}^{n} (u_{ik}^{(t+1)})^m \vec{x}_i}{\sum_{i=1}^{n} (u_{ik}^{(t+1)})^m} \quad (9)$$

The FCM clustering algorithm does not work well for noisy data. In the following, we will review the proposed method.

PCM Clustering Method

In this paper, a fuzzy PCM technique is proposed to improve the FCM method against outlier data and noise. The objective function of the algorithm is defined as relation (10).

$$J_f(X, U_f, C) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m d_{ik}^2 + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - u_{ij})^m \quad (10)$$

Adding the second part to this formula prevents all memberships from being zeroed. Because by decreasing the value of $\boldsymbol{u_{i,j}}$ in the first expression, it's inverse, ie $\boldsymbol{(1-u_{ij})}$ in the second expression will increase. In the above relation $\boldsymbol{\eta_i > 0, (i = 1, \dots, c)}$ is a constant that must be defined for each cluster to establish a balance between the two expressions. That is, $\boldsymbol{\eta_i}$ specifies the last fuzzy boundary point of each cluster. Therefore, the degree of cluster expansion can be controlled by this parameter. To determine the shape of the cluster, the value of $\boldsymbol{\eta_i}$ must be estimated:

$$\eta_i = \frac{\sum_{i=1}^{n} u_{ik}^m d_{ij}^2}{\sum_{i=1}^{n} u_{ik}^m} \tag{11}$$

In the PCM technique, the data belonging to i depends only on the distance between the same data and the j cluster and is independent of the distance between the data and the other clusters (by removing the normalization constraint). The amount of data belonging to each cluster is also determined by the ηi parameter. In the proposed method presented in the article, we will use this method.

KPFCM[26] Clustering Method

In the previous sections, FCM and PCM clustering techniques were examined. These techniques fall into the category of coreless clustering methods. The following is an overview of kernel-based clustering techniques, which is an important category in machine vision techniques. The use of kernel-based techniques in fuzzy clustering algorithms leads to improved algorithms in noise elimination and better data analysis. The KPFCM technique is an extension of the FCM and PCM methods. The KPFCM technique is essentially the same as the kernel-based fuzzy PCM algorithm. The objective function of the PFCM[27] algorithm is as follows.

$$J_{m,\eta}(\mathrm{u},\mathrm{T},\mathrm{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n}(au_{ik}^m + bt_{ik}^n)D_{ik}^2 + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-t_{ik})^m \tag{12}$$

$$\begin{aligned}\|\boldsymbol{\Phi(x_k)} - \boldsymbol{\Phi(v_i)}\|^2 &= (\boldsymbol{\Phi(x_k)} - \boldsymbol{\Phi(v_i)})^T(\boldsymbol{\Phi(x_k)} \\ &\quad - \boldsymbol{\Phi(v_i)}) \\ &= \boldsymbol{\Phi(x_k)^T\Phi(x_k)} - \boldsymbol{\Phi(v_i)^T\Phi(x_k)} \\ &\quad - \boldsymbol{\Phi(x_k)^T\Phi(v_i)} + \boldsymbol{\Phi(v_i)^T\Phi(v_i)}. \\ &= \boldsymbol{K(x_k, x_k)} + \boldsymbol{K(v_i, v_i)} \\ &\quad - \boldsymbol{2K(x_k, v_i)}\end{aligned} \tag{13}$$

Only the Gaussian kernel is used in the fuzzy clustering technique, because with the condition K (x, x) = 1, a Gaussian kernel can be defined as:

[26] kernel possibilistic fuzzy c-means model
[27] Possibilistic Fuzzy C-Means

$$\|\Phi(x_k) - \Phi(v_i)\|^2 = 2(1 - K(x_k, v_i)) \quad (14)$$

By placing the kernel function (14) in formula (12) we will have the following formula:

$$J_{m,\eta}(\mathrm{U},\mathrm{T},\mathrm{V}) = 2\sum_{i=1}^{c}\sum_{k=1}^{n}\left(au_{ik}^{m} + bt_{ik}^{\eta}\right)(1 - k(x_k, v_i)) + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1 - t_{ik})^{\eta} \quad (15)$$

In this formula $\boldsymbol{a, b > 0}$ as well as $\boldsymbol{m, \eta > 1}$ and $\boldsymbol{\gamma_i > 0}$ with given constraints, $\boldsymbol{J_{m,\eta}^{\phi}}$ can be solved using Lagrange multiplication. In this paper, the relation is omitted due to Article space constraints. Therefore, the reduction of relation (15) leads to the following necessary conditions, therefore, by reducing the relation (15) we will have the following essential conditions which of course are not sufficient, and for $\boldsymbol{J_{m,\eta}^{\phi}}$ a local minimum is obtained according to the following Formulas.

$$u_{ik} = \frac{(1 - K(x_k, v_i))^{-1/(m-1)}}{\sum_{j=1}^{c}(1 - K(x_k, v_j))^{-1/(m-1)}} \quad \forall i, k \quad (16)$$

$$t_{ik} = \left[1 + \frac{2b}{\gamma_i}\left(1 - k(x_k, v_i)\right)^{-1/(m-1)}\right]^{-1} \quad \forall_i, k \quad (17)$$

And the formula for calculating the centers is as follows:

$$v_i = \frac{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta})\, k(x_k, v_i) X_k}{\sum_{k=1}^{n}\left(au_{ik}^{m} + bt_{ik}^{\eta}\right) k(x_k, v_i)} \quad (18)$$

Finally, the value of $\boldsymbol{\gamma_i}$ in the KPFCM objective function of Equation (15) is calculated as follows.

$$\gamma_i = \frac{2\sum_{k=1}^{n}(1 - k(x_k, v_i))}{\sum_{k=1}^{n} u_{ik}^{m}} \quad (19)$$

Proposed algorithm

Fuzzy logic-based k-means clustering based on Dempster Shafer's evidence theory

Before explaining this, we will briefly describe the Dempster Shafer Theory.

Dempster Shafer Theory

One method of data fusion is Dempster Shafer Theory Presented by Dempster in 1980, and in 2003 Robert developed the Dempster Shafer algorithm. Classical probability theories are not possible to show ignorance, while most of the phenomena around us are associated with ignorance. Therefore, considering that in the field of medical image segmentation, we often encounter the phenomenon of ignorance, we can use Dempster Shafer theory to solve this problem.

Using Dempster Shafer theory, cluster mass functions can be combined in different ways.

Dempster shafer theory is based on the belief that results from evidence. So that the belief structure of this theory is related to the classical probability model. Dempster shafer theory is an important way to measure and quantify information system uncertainties. An important advantage of shafer theory is that it can be used to quantify uncertainty.

In this theory, rules for combining information from different sources are presented, the most famous of which is Dempster Composition Law. Mathematical representation of uncertainty is another advantage of this method. This theory is used as a tool to analyze uncertainties in inaccurate probability theory.

Diagnosis framework

Assume $\boldsymbol{\Theta}$ is a finite set of elements, an element can be a hypothesis, a goal, or a case of the state of a system. $\boldsymbol{\Theta}$ Is called the Diagnosis framework. An empty set $\emptyset$ indicates the state of a perfect system.

The hypothetical space is considered $\boldsymbol{\Theta\{H_1,H_2,\dots,H_n\}}$ in such a way that the following condition applies.

$$(H_i \cap H_j = \emptyset, \quad \forall\, i \neq j) \qquad (20)$$

The focal space of the hypothesis space is considered as follows:

$$2^{\Theta} = \{\emptyset, H_1, H_2, \dots, H_n, H_1 \cup H_2, \dots, H_i \cup H_j, H_k \cup H_l \cup \dots \cup H_n, \dots \Theta\} \qquad (21)$$

Two or more mass functions can be combined. The relationships of the hypotheses are as follows:

$$m = m_1 \oplus m_2 \dots \oplus m_n \quad (22)$$

The mass function m is called a Basic Probability Assignment function. The orthogonal sum $m = m_1 \oplus m_2$ represents the combination of m_1 and m_2 and contains common information from two sources.

$$(H_i \cap H_j = \emptyset, \quad \forall\, i \neq j) \qquad (23)$$

$$m(A) = K.\sum_{A_i \cap B_j = A} m_{s_i}(A_i).m_{s_i}(B_j) \quad (24)$$

$m(A)$ Represents the share ratio of set A of all relevant and available evidence and supports a claim that belongs to a particular element of Θ and belongs to set A, (belongs to set A and not to a specific subset of A). In examining a defective system, $m(A)$ can be considered as a degree of belief obtained by observing a specific defect. Different information or evidence may give different degrees of belief than a given defect. Each subset A of Θ is called a focal element so that $m(A) > 0$, also $C = \bigcup_{m(A)\neq 0} A$ is a core element of the mass function at Θ.

$$K = \frac{1}{1-k} \qquad (25)$$

In general, for n number of mass function $\mathbf{m_1}$, $\mathbf{m_{2,\dots,}m_n}$ the size of the incompatibility of k is as follows:

$$k = \sum_{A_i \cap B_j = A} m_{s_i}(A_i).m_{s_i}(B_j) \qquad (26)$$

The k provides the basic probability mass of incompatibility between sources of evidence. According to Equation 26, k is obtained from the sum of multiplication of mass functions of all subsets that have no subscription. k is generally interpreted as a measure of incompatibility between information sources. A larger value of k indicates greater source incompatibility. The denominator of $1-k$ in relation 25 is the normalization factor. m Is also a mass function within the same Diagnosis frameΘ.

This section examines decision-making methods using the Dempster Shafer algorithm. There are three types of space in Dempster Shafer's theory:

1- Close world

2- Open world

3- Open developed world

Each space has a specific application that must be selected according to the model of the problem. In situations where a new probability is not to be added to the problem space or we want to ignore the unknown assumption due to the simplicity of the modeling space, it is better to use a closed world that is close to the laws of probability. When there is a possibility of unknowing existence in the problem space but no new hypothesis is added to the problem, it is better to use the open world space and when in addition to the possibility of unknown assumption it is possible to add a new hypothesis to the original hypothesis space, an open developed world is used [9-12]. The open developed world has been used to model uncertainty.

$$m_{12}(H_2) = m_2(H_2).(1 - m_1(H_1)) \quad (27)$$

$$m_{12}(\overline{H_1}) = m_{12}(H_2) = m_1(\overline{H_1}).m_2(\Theta) \quad (28)$$

$$m_{12}(\overline{H_2}) = m_{12}(H_1) = m_2(\overline{H_2}).m_1(\Theta) \quad (29)$$

$$m_{12}(*) = m_1(\overline{H_1}).m_2(\overline{H_2}) \quad (30)$$

$$m_{12}(\Theta) = m_1(\Theta).m_2(\Theta) \quad (31)$$

$$m_{12}(\emptyset) = m_1(H_1).m_2(H_2) \quad (32)$$

$$\Theta = \{H_1, H_2, H_3, \dots, H_i, \dots, H_n, *\} \quad (33)$$

$$m(H_i \cup H_j) = m_i(\Theta).m_j(\Theta) \prod_{\substack{a=1 \\ a\neq i \\ a\neq j}}^{n} m_a(\overline{H_a}) \quad (34)$$

$$m(H_i) = m_i(H_i) \prod_{\substack{a=1 \\ a\neq i}}^{n} (1 - m_a(H_a)) \quad (35)$$

$$m(H_i \cup *) = m_i(\Theta).m_j(\Theta) \prod_{\substack{a=1 \\ a\neq i \\ a\neq j}}^{n} m_a(\overline{H_a}) \quad (36)$$

$$m(H_i \cup H_j \cup H_k \cup *) = m_i(\Theta).m_j(\Theta).m_k(\Theta) \prod_{\substack{a=1 \\ a\neq i \\ a\neq j \\ a\neq k}}^{n} m_a(\overline{H_a}) \quad (37)$$

$$m(* \bigcup_{i=2}^{n} H_i) unit\ n-1 \quad (38)$$

$$m(H_i \cup H_j \cup \dots \cup H_l \cup *) = m_i(\Theta).m_j(\Theta) \dots m_l(\Theta) \prod_{\substack{a=1 \\ a\neq i \\ a\neq j \\ \dots \\ a\neq l}}^{n} m_a(\overline{H_a}) \quad (39)$$

$$m(\overline{H_\iota}) = m_i(\overline{H_\iota}) \prod_{\substack{a=1 \\ a\neq j}}^{n} m_a(\Theta) \qquad (40)$$

$$m(\Theta) = \prod_{a=1i}^{n} m_a(\Theta) \qquad (41)$$

$$m(*) = \prod_{a=1}^{n} m_a(\overline{H_a}) \qquad (42)$$

$$m(\emptyset) = 1 - \left[\prod_{a=1}^{n}(1 - m_a(H_a)) + \sum_{a=1}^{n} m_a(H_a) * \prod_{\substack{b=1 \\ b\neq a}}^{n}(1 - m_b(H_b))\right] \qquad (43)$$

The resulting clustering data should be converted to a format that can be used for the data fusion unit. Raw data needs to be converted to probability format (confirmation or rejection of a hypothesis) in order to be used in the proposed method. In the following, the methods of preparing raw data and converting it into a probabilistic format for use in data fusion algorithms are discussed. It is worth mentioning that all simulations is done with MATLAB 2020a software and in a system with Intel (R) -Core (TM) -i7 CPU-Q740-1.73GHz and 8 GB RAM and 64-bit Windows. In this paper, in order to combine clustering, a proposed fuzzy algorithm is used in which the reliability of each hypothesis is determined according to the design of the fuzzy system.

In this method, the probability of belonging to each cluster is considered as a parameter to reject or confirm the hypothesis and fuzzy system design is done. For this purpose, the following fuzzy system is used to trust the probability of occurrence of a hypothesis. The output of a fuzzy system is the probability of being rejected, confirmed, or ignored in relation to a hypothesis. The designed fuzzy system is shown in the figure below.

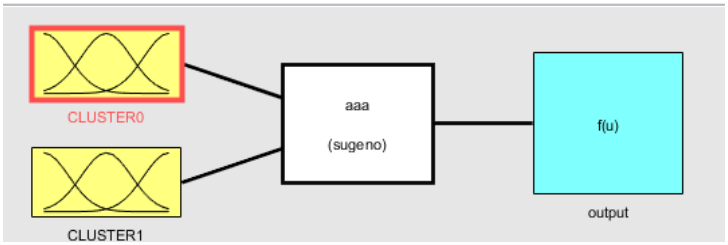


Figure 1: The Proposed fuzzy system

In the proposed fuzzy system, the inputs of the fuzzy system (Figure 4) are obtained according to the probability percentage of each cluster. The output of the fuzzy system (Figure 3) determines the proof or rejection of a hypothesis and can use the raw data obtained from the basic clustering to determine the degree of belonging of each pixel to each cluster based on the probability percentage between 0 and 100, so that The probability of what each observed pixel is related to which hypothesis and also the degree of this probability is determined. The output values of the fuzzy system are considered as the input of the Dempster Shafer algorithm. In Dempster Shafer algorithm, according to relations 27 to 41, the values of data fusion as well as unknown assumptions and new assumptions are made.

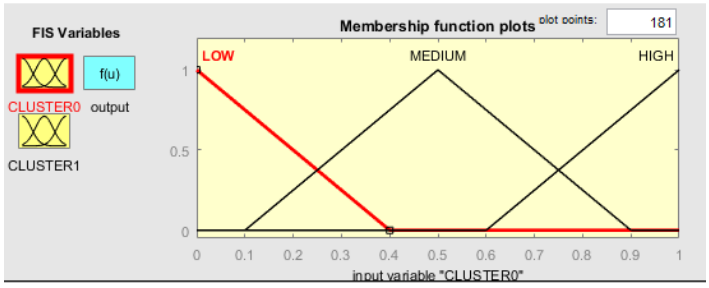


Figure 2: The first input of fuzzy system membership functions

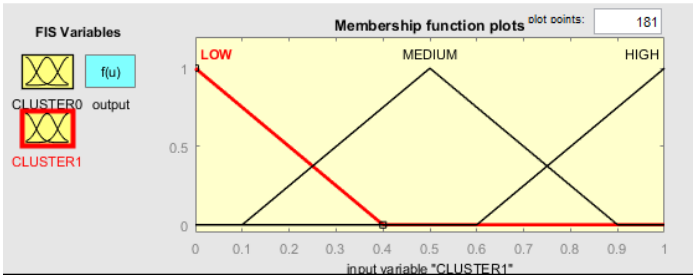


Figure 3: The Second input of fuzzy system membership functions

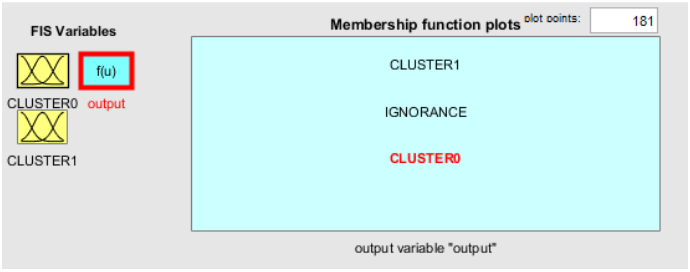


Figure 4: Output fuzzy system membership functions

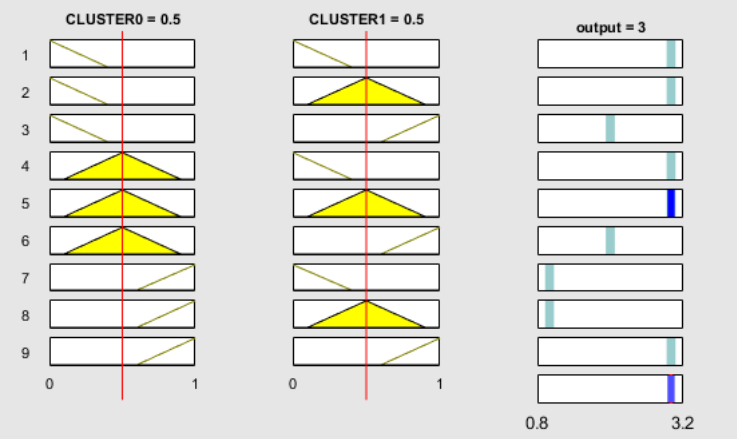


Figure 5: Fuzzy system rules

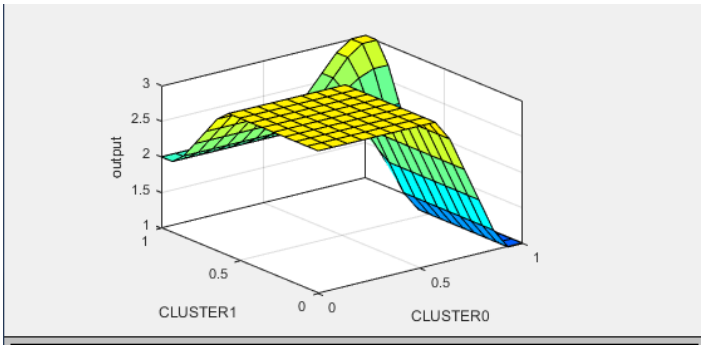
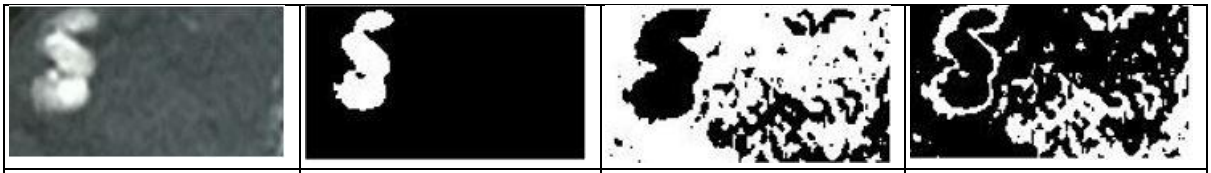


Figure 6: Fuzzy system graph



1)Main image	2)Tumor region	3)Healthy regions	4)Unknown regions

Figure 7: Sample image of "Cancer part clustering", "Healthy part clustering" and ""ignorance assumptions part clustering

Practically, we can refer to the problem of clustering to diagnose liver tumors, whose common assumptions are that each pixel is cancerous or healthy. With the help of the proposed method, the unknown assumptions can be added to the space of the initial hypotheses. To validate the proposed algorithm, the dataset images available on the website (https://sliver07.grand-challenge.org/) have been used.

Table 1- Comparison of different articles in the diagnosis of liver cancer

Reference	year	Method	features	constraints
27	2014	region growing method, pool segmentation, SVM segmentation	robust, convenient access to organ measurements	expert user intervention is required, high complexity
28	2014	Intelligent CDS method, using different classifiers	accuracy rate 29/95%	High complexity

29	**2014**	EM/MPM method	Reduce error in misdiagnosis of tumor regions	High complexity, only spherical structure is considered
30	**2011**	Pool-like algorithm	Reduce the number of decisions and reduce processing and calculation time	,Semi-automatic Low accuracy rate (%87)
31	**2015**	Use SVM classifier, haar wavelet transform	Reduce the number of decisions, reduce processing time and user dependency	user dependency
32	**2015**	Using CLAHE and CVHE algorithms, SVM classifier	accuracy rate %97	Computational complexity
33	**2015**	Using MRF algorithm, SVM classifier	Increase accuracy in diagnosing cancerous mass	Increase computational time
proposed method	**2022**	Use of fuzzy system and KMEANS	Increase the accuracy rate(91.66%) and possibility of data fusion in clustering	

Hypothesis extraction for clustering

We propose the use of KMEANS clustering information with Euclidean and Manhattan distance criteria for final clustering. Thus, in addition to combining the values obtained from the probability of dependence on each cluster head, which is done using the Dempster Shafer method, the results of the clustering section, which is done using different methods, can be used to improve the final clustering accuracy. Each clustering method can perform clustering by extracting feature characteristics. Raw data obtained from different clustering methods must be converted to a format that can be used for the data fusion unit. Raw data needs to be converted to probability format (confirmation or rejection of a hypothesis) in order to be used in the proposed method. In the following, the methods of preparing raw data and converting it into a probabilistic format for use in information fusion algorithms are discussed.

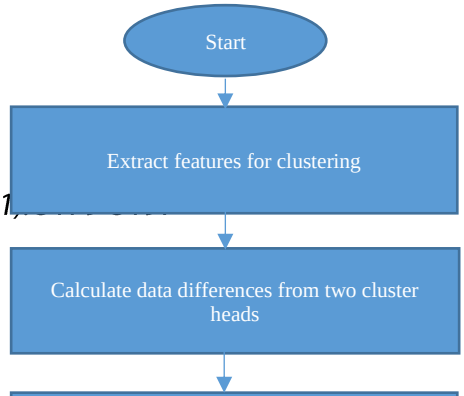


Figure 8: The Proposed flowchart for extracting probabilities and combining clustering information

It is assumed that clustering is done in two ways, kmeans with Manhattan and Euclidean distances. The clustering hypothesis space has three members, which are "Belonging to a cluster 1", "Belonging to a cluster 2"and "unknown or ignorance ", respectively. Each of the clustering methods calculates the probability values for the three hypotheses using the introduced fuzzy method. It can be shown that the probability of all three hypotheses is calculated as follows: For this purpose, first it is necessary to extract the mass function of each input based on the variance of the clustering method. Relations (20-41) are used to construct mass functions. To use the dempster shafer algorithm, two hypotheses of "Existence of value "and" not Existence of value " are considered:

$$\Theta = \{\text{Cluster head 1} \quad \text{Cluster head 2}\} \quad (44)$$

Therefore, the Θ Diagnosis framework consists of four subsets;

$$2^{\theta} = \{ (\text{Cluster head 1}) \text{ ، } (\text{Cluster head 2}) \text{ ، } (\text{ Existence or not Existence of value }) \text{ ، } (\text{ ignorance}) \} \quad (45)$$

Given that existence and non-existence in hypotheses cannot occur at the same time, this probability is considered to be zero, as can be seen. The value of k is calculated as follows:

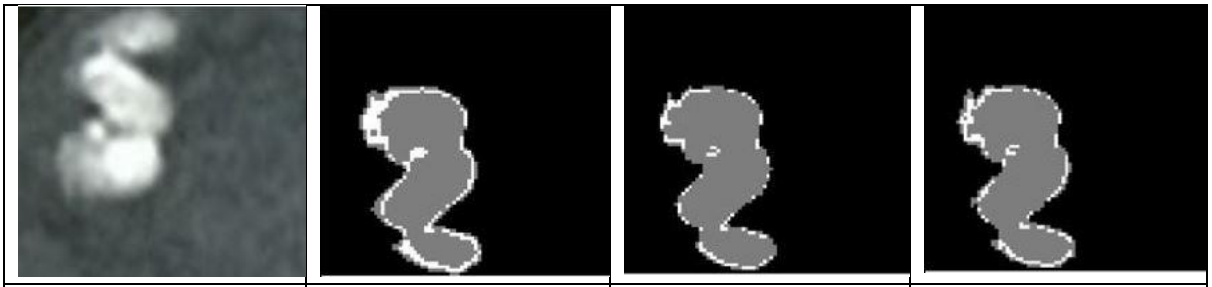
$$k = \frac{1}{1-k} = \frac{1}{1-\sum_{a_i \cap b_{j=\emptyset}} [m_A(a_i) * m_B(b_j)]} \quad (46)$$

k Is equal to "the degree of conflict between different information sources". The closer the value of k is to the number 1 indicates the contradiction of information sources and the closer the value of k is to the number 0 indicates the compatibility between them. In this research, the dempster Shafer algorithm for the feature layer has been improved by considering the maximum probability in the mass function.

$$K = \frac{1}{1-k} = \frac{1}{1-\sum_{a_i \cap b_j=\emptyset} = 1-[(\max(m)*(1-\max(m1)))+(\max(m1)*(1-\max(m)))]} \quad (47)$$

$$m_{j,t,l}(A) = K * \left[m_{j,t,1,l}(A) * m_{j,t,2,l}(A)\right], \quad (48)$$

$$m_{j,t,l}(N) = K * \left[m_{j,t,1,l}(N) * m_{j,t,2,l}(N)\right] \quad (49)$$



1)Main image	2)Kmeans method with Euclidean distance	3)Kmeans method with Manhattan distance	4)Kmeans method with Euclidean-Manhattan combined distance

Figure 9: 1) Main image, 2) Kmeans method with Euclidean distance, 3) Kmeans method with Manhattan distance, 4) Kmeans method with Euclidean-Manhattan combined distance.

The black part means that the texture is **healthy**; the gray part means **unhealthy**, the white part means **ignorance.**

Evaluation by other techniques

In the following, tests have been performed on two types of real clinical and virtual images. To validate the proposed method, the proposed technique is compared with six fuzzy clustering algorithms and listed in the table. The accuracy of the classification is assessed according to the JS criterion as follows:

$$JS(S_1, S_2) = \left|\frac{S_1 \cap S_2}{S_1 \cup S_2}\right| \quad (50)$$

Where S_1 represents a set of pixels belonging to a class that are found by a particular technique. And S_2 represents a set of pixels in a class very similar to a reference segmented image.

The proposed technique can be used for other clustering algorithms. Other distance criteria can also be used in the kmeans algorithm and the clustering technique can be performed based on it. In this paper, as an example, kmeans clustering algorithms with Manhattan and Euclidean distances are compared by FCM method. The Gaussian noise used in this paper was 2%.

Table.2. Compare the accuracy of segmentation of the proposed technique with other methods

Method	White matter	Gray matter	Cerebro Spinal Fluid	Average accuracy
K-Means	65	60	63	62.66
FCM	69	65	66	66.66
PCM	67	68	60	65.00
PFCM	74	76	71	73.66
SFCM[28]	92	86	73	83.66
KPFCM	95	88	83	88.66
Proposed	95	91	89	91.66

The important point of the proposed method compared to other clustering techniques is to reduce the uncertainty and clustering error in medical images. According to the results, the accuracy of segmentation of the proposed method was 91.66%. The proposed method can also be used in hybrid clustering techniques and techniques that require data fusion. This technique is especially effective for the Dempster–Shafer method because it requires an unknown hypothesis in case of uncertainty.

Conclusion

Correct and efficient diagnosis of liver cancer results in successful treatment. In this paper, fuzzy clustering methods based on evidence theory in the diagnosis of liver cancer are studied. The method proposed in this paper, in addition to having high accuracy, also has a high convergence speed. In semi-automatic methods, interaction with the user is required and the desired region is identified by the user and used for subsequent image processing processes. While in fully automatic methods, segmentation is done without user interaction. Finally, the features and disadvantages of these methods are evaluated. Also, kmeans clustering with Euclidean and Manhattan distance criteria was used for the final hybrid clustering based on Dempster-Shafer theory, which led to improved clustering accuracy in the Uncertainty conditions of the issue under discussion.

[28] Shape Fuzzy C -Means

Conflict of interest

Babak Fouladi Nia declares that he has no conflict of interest. Abbas Karimi declares that he has no conflict of interest. Faraneh Zarafshan declares that he has no conflict of interest. Manochehr kazemi declares that he has no conflict of interest.

References

1. Akram, M.U., A. Khanum, and K. Iqbal, An automated system for liver ct enhancement and segmentation. Graphics, Vision and Image Processing GVIP, **2010**. 10(4): p. 17-22.
2. Nural, W., J. Yussof, and H. Burkhardt, 3D anisotropic diffusion for liver segmentation. World Acad Sci. Eng Technol, **2009**. 57.
3. Qi, Y., et al. Semi-automatic segmentation of liver tumors from CT scans using Bayesian rule-based 3D region growing. in MICCAI workshop. **2008**.
4. Ali, L., et al. Intelligent image processing techniques for cancer progression detection, recognition and prediction in the human liver. in 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE). **2014**. IEEE.
5. Anisha, P., C.K.K. Reddy, and L.N. Prasad. A pragmatic approach for detecting liver cancer using image processing and data mining techniques. in 2015 International Conference on Signal Processing and Communication Engineering Systems. **2015**. IEEE.
6. Krishan, A. and D. Mittal, Detection and classification of liver cancer using CT images. International Journal on Recent Technologies in Mechanical and Electrical Engineering, **2015**. 2(5): p. 93-98.
7. Raj, A. and M. Jayasree, Automated liver tumor detection using markov random field segmentation. Procedia Technology, **2016**. 24: p. 1305-1310.
8. Ahmadi, K., A. Karimi, and B. Fouladi Nia, New Technique for Automatic Segmentation of Blood Vessels in CT Scan Images of Liver Based on Optimized Fuzzy C-Means Method. Comput Math Methods Med, 2016. **2016**: p. 5237191.
9. Nia, B.F. and A. Branch, A hybrid automatic segmentation method of blood vessels in CT scan images of liver. Journal of Medical Imaging and Health Informatics, **2017**. 7(4): p. 799-804.
10. Xu, L., et al., Liver segmentation based on region growing and level set active contour model with new signed pressure force function. Optik, **2020**. 202: p. 163705.
11. Sakthisaravanan, B. and R. Meenakshi, OPBS-SSHC: outline preservation based segmentation and search based hybrid classification techniques for liver tumor detection. Multimedia Tools and Applications, **2020**. 79: p. 22497-22523.
12. JE, A.L.K. and S.V. Jinny. Automatic Extraction of Lesions and Hepatic Structures in Liver using Segmentation Techniques. in **2020** Second International Conference on Inventive Research in Computing Applications (ICIRCA). 2020. IEEE.
13. Zhang, Z., et al., GVFOM: a novel external force for active contour based image segmentation. Information Sciences, **2020**. 506: p. 1-18.

14. Chung, M., Lee, J., Park, S., Lee, C. E., Lee, J., & Shin, Y.-G. (**2021**). Liver segmentation in abdominal CT images via auto-context neural network and self-supervised contour attention. Artificial Intelligence in Medicine, 113, 102023.
15. Xie, X., Zhang, W., Wang, H., Li, L., Feng, Z., Wang, Z., ... & Pan, X. (**2021**). Dynamic adaptive residual network for liver CT image segmentation. Computers & Electrical Engineering, 91, 107024.
16. Sengun, K. E., Cetin, Y. T., Guzel, M. S., Can, S., & Bostanci, E. (**2021**). Automatic Liver Segmentation from CT Images Using Deep Learning Algorithms: A Comparative Study. arXiv preprint arXiv:2101.09987.
17. Liu, Z., Han, K., Wang, Z., Zhang, J., Song, Y., Yao, X., ... & Sheng, V. S. (**2021**). Automatic liver segmentation from abdominal CT volumes using improved convolution neural networks. Multimedia Systems, 27(1), 111-124.
18. Araújo, J. D. L., da Cruz, L. B., Ferreira, J. L., da Silva Neto, O. P., Silva, A. C., de Paiva, A. C., & Gattass, M. (**2021**). An automatic method for segmentation of liver lesions in computed tomography images using deep neural networks. Expert Systems with Applications, 180, 115064.
19. Kaur, R., L. Kaur, and S. Gupta, Enhanced K-mean clustering algorithm for liver image segmentation to extract cyst region. IJCA Special Issue on Novel Aspects of Digital Imaging Applications, **2011**. 1: p. 59-66.
20. Jiang, H., et al., A novel multiinstance learning approach for liver cancer recognition on abdominal CT images based on CPSO-SVM and IO. Computational and mathematical methods in medicine, 2013. **2013**.