Indian Journal of Science and Technology, Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/104642, November 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Detection and Classification of Lung Cancer MRI Images by using Enhanced K Nearest Neighbor Algorithm

P. Thamilselvan* and J. G. R. Sathiaseelan

Department of Computer Science, Bishop Heber College, Tiruchirappalli - 620017, Tamil Nadu, India;
thamilselvan1987@gmail.com, ejgrsathiaseelan@gmail.com

Abstract

Objectives: To detect and classify the malignant cancer tissues and benign cancer tissues in MR lung cancer images by using k nearest neighbor mining algorithm. **Methods/Statistical Analysis:** In this paper, the Enhanced K Nearest Neighbor (EKNN) algorithm is executed to identify the lung cancer images. The k nearest neighbor technique is an important method of data mining algorithms. **Findings:** This work implicates four stages such as pre-processing, feature extraction, classification and detection of cancer tissues. In preprocessing stage, morphological process is used to filter the irrelevant noisy data in images. In the second phase, statistical and discriminator algorithm is used to extract the images. In the last stage, the improved k Nearest Neighbor (EKNN) method has been used to classify and identify the cancerous tissues in MRI images. The detection of cancer tissues and classification is done by executing four steps of Enhanced k Nearest Neighbor method which are measuring the Euclidean distance value, determining the k value, calculating the minimum distance and detecting the cancerous cells. **Improvements/Applications:** The experimental study with enhanced k nearest neighbor method shows better and promising classification result for classifying benign and malignant tissues.

Keywords: Geometrical and Statistical Properties, Image Classification, Image Mining, MRI Images, k Nearest Neighbor, Morphological Method

1. Introduction

Image mining is used to determine the knowledge from the image dataset by using the techniques like image clustering, image classification[1], image characterization based summarization and association rule mining[2]. Mining is a very predominant activity for researchers[3]. The image classification technique is represented by color histogram, feature transformation, describing the characterization of images and medical image analysis[4]. It plays a vital role in computer visualization and applications such as image quality assessment[5], image retrieval[6], and image annotation[7]. In the last few years, several new image classification techniques have been implemented in machine learning and artificial intelligence. This research work is aimed to improve the performance of EKNN to identify the cancer tissues in MRI images and to improve classification accuracy.

The k nearest neighbor method is one of the top 10 algorithms in data mining[8]. The many communal lazy learning algorithms are nearest neighbor classification, which classify the image by the class of its related images to the database. The nearest neighbor similarities deliver additional semantic reliable and noise free description of the similarity between images when comrade with the image similarities. In this paper, we proposed an enhanced k nearest neighbor method to detecting mass tissues and classify the benign and malignant tissue, which is analyzed in our earlier work[9]. The nearest neighbor technique is based on distance function and the classification process measured based on Euclidean

*Author for correspondence

distance calculation. The classification is done by executing four steps of nearest neighbor which are intention of allocating of majority class, Euclidean distance objective function, determining the k value, exposure of class based majority ranking and finding the minimum distance. The accuracy and then arithmetic measurement is calculated by using Euclidean distance measurement analysis.

The preprocessing techniques are also used in this work to obtain exact results. In the pre-processing phase, the morphological method has been used to remove the unwanted data from the image. The pre-processing technique is used to remove and filter the annoying noisy information from lung cancer images. The feature extraction technique that used to minimize the unique dataset by manipulative some converted features. To find feature extraction of image geometrical and statistical properties method have been used to extract images. Feature extraction is a progression of transmitting the input data into a matching format for image mining task. A number of features can be removed from images by applying statistical methods, image transform and texture based methods[10]. When handling large amount of image repository much of image data that becomes in valuable image next the process of feature extraction. It helps in enlightening the performance of many classifiers, minimizing the modeling time and makes the humble and laid-back to results. This work has been mainly focused on detecting and classification of lung cancer medical images and also to find the minimum distance of the nearest neighbor method.

In the related works several works were performed to develop conventional and classification based on methods for various types of medical images[11–18]. The neutral of this prediction and classification techniques is to detect and classifying cancerous and non-cancerous images. In[19] hyper spectral image classification technique based on the supervised random forest technique and it also Random Forest Classifier (RFC) algorithm has been applied airborne prism experiment images. The random forest method has method implemented for hyperspectral image classification and increase the speed of classification process. Finally, the classification process is carried out based on conversation of labeled to unlabeled data in the process of adaboost algorithm. The overall classification accuracy is 82.63% and kappa coefficient is 0.78% by using adaboost method.

To find the better results,[20] trained the collective of kernel regression trees which method the desired similarity process as an order of fuzzy decision bases. The incentive behind choosing kernel regression tree method is primarily their accuracy, natural ability and accuracy to handle uncertain and conditional information effectively. The hierarchical reasoning seems close to human distinguish between generally similar to visual patterns. The proposed kernel decision tree method tested with several image datasets and it demonstrates ease of interpretation, high classification accuracy, scalability and the feature representation. The automatic classification of lung cancer in magnetic resonance images is a significant analytical tool in computer aided diagnosis system[21]. The proposed method has been tested with CT (Computed Tomography) images to find the classification accuracy of lung cancer images by using Multilayer Perceptron Neural Network (MLPNN). In the feature extraction stage genetic algorithm has used extract the images. This MLPNN technique works in detection of lung cancer images with high accuracy, sensitivity and specificity.

In[22] reports the difficulty of urban scene classification using adaboost method in high-resolution synthetic satellite images. This adaboost method the concerned author has represented three methods, i.e., pyramid representation adaboost, GR adaboost, BOF adaboost to find the classification accuracy. This proposed adaboost method shows 87.67% accuracy in image classification accuracy. The purpose of unsupervised linear discriminant analysis method to classify and segment as cerebral white matter, gray matter and spinal fluid from the multispectral MR imaging of human brain[23]. In this technique ULDA method is composed of two progressions, Target Generation process and Linear Discriminant Analysis classification. The LDA method is classify and segment in MR images much more effectively.

The computational algorithms are useful for medical image analysis because they afford supplementary data that cannot be gained by humble interpretation of clinical performances and medical imaging[24]. This exertion intelligences the texture analysis of CT images and enlargement of back propagation neural network, probabilistic neural network and linear vector quantization neural network for classification of fatty liver computed tomography abnormal images. The proposed method shows that probabilistic neural network is a worthy classifier and it's giving an accuracy of 95% for classifying the cirrhosis liver. In[25] suggested Logistic Regression (LR) classifier method, to calculate the performance of classification by using leave-one-out cross validation method. The pattern recognition method and advance statistical

algorithm have been aggressively used to excavation the reproductive patterns during the process of premature stages of Alzheimer dementia[26]. The Logistic Regression (LR) algorithms prove that accuracy is 87.5% by using computed tomography brain images.

2. Proposed Work

This framework consists of three phases preprocessing, feature extraction, detection of cancer tissues and classification. In preprocessing level, the morphological techniques have been used to remove the irrelevant noisy data in images. In feature extraction level, geometrical and statistical techniques have been used to extract the images. In last level, EKNN techniques have been implemented to detecting the lung cancer tissues in MR images and classifying malignant or benign images. The stage of proposed framework is as enumerated as follows (Figure 1.)

This framework enlightening the contrast of the magnetic resonance images done using preprocessing technique is done by changing the input image to gray scale image. After enlightening the contrast of the MR images it is applied to geometrical and statistical method to extract the feature contrast. To detect and classify the lung cancer images nearest neighbor method has been implemented in this work.

2.1 Data Set

In this work, the performance of EKNN method has been evaluated using MR images of human sequence of benign and malignant. The set of several magnetic resonance images shown in Figure 2 is collected from a patient without lung cancer tissues and with lung cancer tissues. The MRI images are compressed to 256×256 before preprocessing stage and feature extraction stages. The EKNN technique tested around more than 50 images to identify and classify the lung cancer tissues in images and it has implemented by using Matlab10.

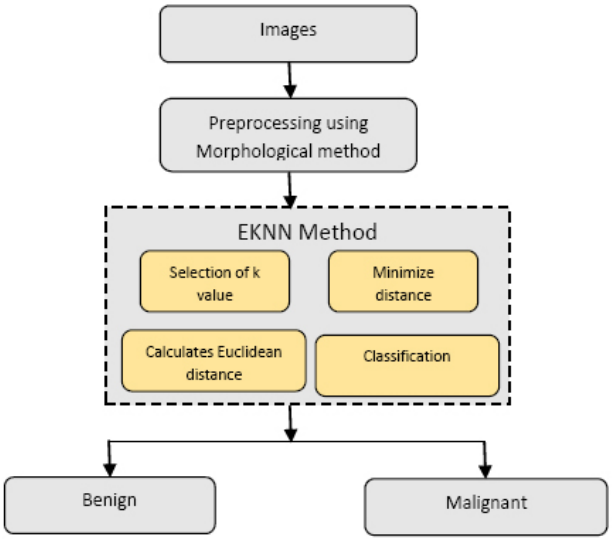

Figure 1. Stages of proposed framework.

2.2 Morphological Method

Morphological method is technique of image preprocessing based on the shape and form of objects. This technique applies a constructing element to generate an input image to output image at the equivalent size. The significance value of apiece pixel in the input image is based on conforming pixel in the input image with its neighbors[27]. The morphological basically contains four operations such as opening, closing, dilation and erosion. At this preprocessing stage we have used only erosion and dilation process it is also used to accomplish morphological image preprocessing analysis.

2.2.1 Dilation

The dilation is a process of transformation that provides an image that is the same shape as the original with different size. Dilation increases the width of maximum areas, so it can eliminate bad imprudent noises from the image. This function applies to the suitable pixels in the neighborhood and it allocates a value to the matching pixel in the output image. The result of dilation processed image is shown in Figure 3.

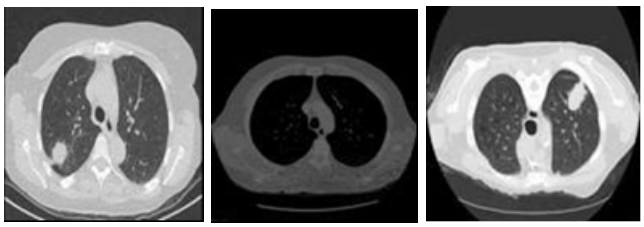

Figure 2. Sample image data.

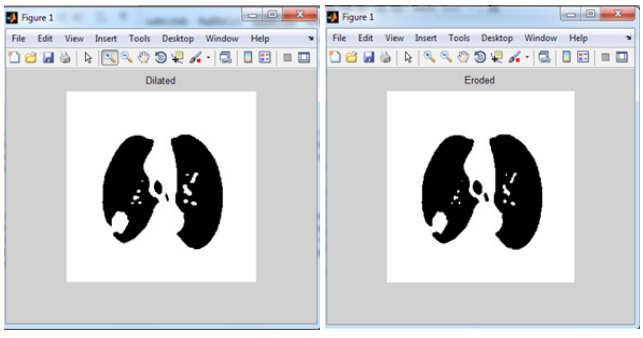

Figure 3. Dilation and erosion.

2.2.2 Erosion

The erosion process is used to minimize the object in the image and it reduces the width of smallest region, so it can confiscate positive noises from the image. The erosion process that applies to the proper pixels in the neighborhood and allocates the value to appropriate pixels. The result of erosion processed image is shown in Figure 3. The process of morphological method to reduce the noise from the conversion and rotation of image into white and black and it describes the development of preprocessing system.

2.3 Geometrical and Statistical Method

The geometrical and statistical structures perimeter, diameter, irregularity index and area have been estimated from the separated lung image nodules[28]. The number of pixels from image that having the values which gives the area of segmented cancer image. The values of images that gives circumstantial of the image which is black. Lung cancer image is categorized incompletely in its cancer border. For this investigation, the indiscretions in the cancer that are calculated by directory[29].

$$I = 4\,\pi\,A/P^{\wedge}\,24\,\pi\,A$$

Where, P is the boundary of cancer A is area of cancer in the pixels. The indiscretion directory is equivalent to only for circle and any other shape. The feature extracted process image by using geometrical and statistical method is shown in Figure 4.

2.4 Enhanced k Nearest Neighbor Method

The nearest neighbor method is one of the humblest and hoariest methods in supervised learning classification.

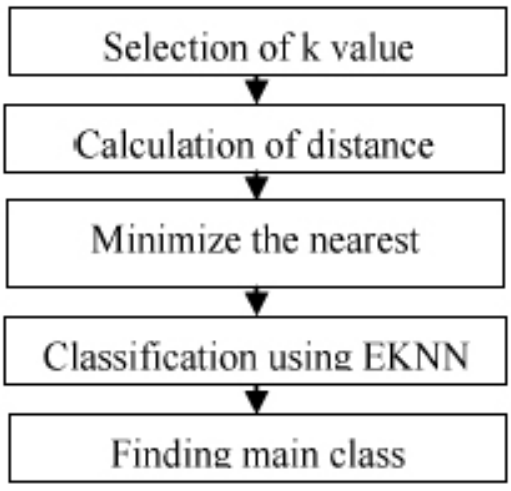

Figure 4. Steps of k nearest neighbor classification.

The aim of this technique is to identify nearest k shortest distance value between the pixels and classify the appeared example according to greatest comparable class. Basically familiarity is demarcated with Euclidean distance measurement calculation. The arbitrary instance x is defined by the feature vector.

$$< a_1(x), a2(x),\ldots a_n(x)$$

Here $a1(_x)$ represents the characteristic value of instance x. Then the distance among two instance x_i and x_j is distinct to be $d(x_i,x_j)$ as below

$$d(x_i, x_j) = \sqrt{\sum^{n}_{r=1}} (a_r(x_i) - a_r(x_j))^2$$

Subsequently, the illustration is allocated to most similar class from nearest neighbor method and it is also used to estimate the actual value for an unidentified samples. The suitable k value and distance dimension regulates the enactment of nearest neighbor classifier method. If the data values are not consistently circulated, then the fortitude of k value becomes difficult to find. Basically highest k values are selected in the event of unwanted dataset to make the borders horizontal between the classes.

The true k value can be designated by several experiential methods like clustering and cross validation. The special case of class value is predicted to be the class of closest training sample is called nearest neighbor algorithm. It is difficult to indicate similar k value for all various applications. Various challenges have done to novel methods to growth the performance of k nearest neighbor algorithm by using prior information such as circulation of data and feature range selection. In common, the five phases are accomplished for k nearest neighbor method (Figure 5).

Step 1: Choosing of k value: k value is completely up to handler. Essentially after coarsely some trials the k value will be chosen based on the result.

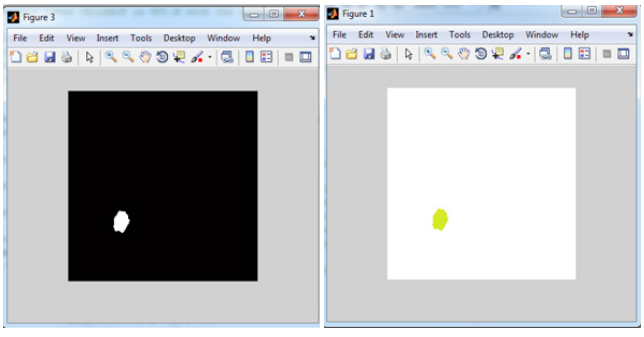

Figure 5. Identified malignant tissues using proposed method.

Step 2: Distance Calculation: The distance calculation will be measured in this level. The many distance calculated based on the Euclidean distances only.

Step 3: Distance arrangement in ascending order: nominated k value is also important in this level. The recognized distance will be arranged in the ascending order and the minimum k values are taken.

Step 4: Classification: classes of k nearest neighbors are documented.

Finding Main class: In the final stage, interrogated data can be categorized based on identified knn by using highest ratio. This ratio is measured for every class of knn with the number of information possessed by the class over k. Let P= {p1, p2, p3 ….. pc} is the set of knn possibilities for all class, here c is number of class. The highest ratio is calculated in below equation.

$$P_{max} = \max(P_i/k)$$

2.4.1 Pseudocode for EKNN

The following program that represents sample code of nearest neighbor classification technique that identifies the shortest neighbors in a set of points. To find the distance metric Euclidean distance measurement can used.

Usage:

1. Neighbors distances = k Nearest Neighbors (data matrix, query matrix, k), 2. Data matrix (N x D) - N vectors with dimensionality D (within which we search for the nearest neighbors), 3. Query matrix (M x D) - M query vectors with dimensionality D, 4. K (1 x 1) - Number of nearest neighbors desired.

```
Function [neighbor Ids neighbor Distances] = k Nearest Neighbors (data matrix, query matrix, k)
{
Neighbor Ids = zeros (size (query matrix, 1), k);
Neighbor distances = neighbor Ids;
Num data vectors = size (data matrix, 1);
Num query vectors = size (query matrix, 1);
For i=1: num queryvectors,
dist = sum((repmat(query matrix(i,:),num data vectors,1) – data matrix).^2,2);
[sortval sortpos] = sort (dist,'ascend');
Neighbor Ids (i, :) = sortpos(1:k);
Distance of Neighbors (i, :) = sqrt (sortval(1:k));
end
}
```

3. Result and Discussion

In this research, we effectively established for method for detecting of lung cancer cells using EKNN technique. The performance of this method is analyzed by using 50 of non-cancer and cancer MR images and it is shown in Figure 2 and Figure 5. The information is calculated by different size of malignant and benign onto normal cancer images. We calculated the performance of the improved nearest neighbor classifier in terms of classification accuracy. The classification accuracy process is correctly performed by following formula.

$$\text{Accuracy} = TP+TN/TP+TN+FP+FN$$

Where, TP is True Positive - properly classified positive cases, TN is True Negative - properly classified negative cases, FP is False Positive - wrongly classified negative cases and FN is False Negative – wrongly classified progressive cases. In this research, we have admirably enhanced a resolution for the identification of lung cancer tissues using image mining algorithm.

The Figure 5 shows the identification of lung cancer tissues in MRI images for analyzing the performance of proposed k nearest neighbor technique. The proposed technique is a supervised learning technique and it is also used to calculate the minimum distance of nearest neighbor. The classification results of this proposed k nearest neighbor techniques shows 96.57% and it finds the minimum distance of neighbor i.e., is 0.41876.

3.1 Performance Analysis

The assessment of proposed nearest neighbor technique in terms of classification accuracy, error rates and minimum distance of neighbors shows the better result. We compared our proposed nearest neighbor technique with existing nearest neighbor method[30] that is shown in Figure 6 and Table 1 to analyze the performance of this method.

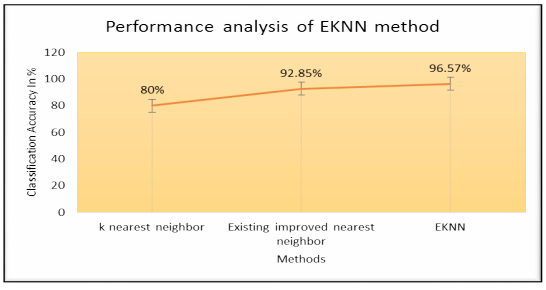

Figure 6. Relative analysis of EKNN method.

Table 1. Performance analysis of EKNN method

Technique	Classification accuracy	Error rates	Minimum neighbor distance
K nearest neighbor	80%	20	–
Existing improved nearest neighbor	92.85%	7.15	–
EKNN method	**96.57%**	**3.43**	**0.41876**

According to the experimental result, the proposed EKNN technique is also best method for identifying cancer tissues and efficient for classifying the lung images as benign or malignant. Based on the result, our proposed method produces the 96.57% accuracy in classification accuracy and as well as minimize the processing time.

Our results have compared to previous results of knn reported recently based on the medical MR image classification. The classification performance of this method reflects reasonable accuracy. This work has been carried out by using MATLAB version 10.0 environment.

4. Conclusion

In this research analysis, we have enhanced k nearest neighbor technique for identifying and classify lung cancer tissues in images as malignant and benign classes. The image classification system is designed by nearest neighbor that provides better results in classification of lung cancer images. Based on the experimental discussion and results, the proposed technique is efficient for technique for identify the cancer tissues in image and classify the benign and malignant image. This EKNN method achieves 96.57% accuracy in classification rates and it is also segments malignant cells exactly. The classification performance of this proposed method shows low error signal, better classification accuracy and its finds the minimum distance. This research work can be further extended to test in large amount of data to reduce processing time and increase the classification accuracy.

5. References

1. Zhang X, Hao S, Xu C, Qian X, Wang M, Jiang J. Image classification based on low-matrix recovery and naïve bayes collaborative representation. Neurocomputing. 2015; 169(2):110–18.
2. Gajjar TY, Chauhan NC. A review on image mining frameworks and algorithms. International Journal of Computer Science and Information Technologies. 2012; 3(3):4064–6.
3. Lei B, Tan EL, Chen S, Dong N, Wang T. Saliency driven image classification method based on histogram mining and image score. Pattern Recognition. 2015; 48(8):2567–80.
4. Zhu X, Xie Q, Zhu Y, Liu X, Zhang S. Multi sparsity kernel reconstruction for multi class image classification. Neurocomputing. 2015; 169(2):43–9.
5. Hong R, Pan J, Hao S, Wang M, Xuo F, Wu X. Image quality assessment based on matching pursuit. Information Science. 2014; 273:196–211.
6. Hong R, Tang J, Tan HK, Ngo CW, Yan S, Chua TS. Beyond searching: Event driven summarization for web videos. ACM Transactions on Multimedia Computing, Communications, and Applications. 2011; 7(4).
7. Hong R, Wang M, GaoY, Tao D, Li X, Wu X. Image annotation by multiple learning with discriminative feature mapping selection. IEEE Transactions on Cybernetics. 2014; 44(5):669–80.
8. Wu X, Kumar P, Quinlan JR, Ghosh J, Yang Q, Motoda H. Top 10 algorithm in data mining. Knowledge and Information Systems. 2008; 14(1):1–37.
9. Thamilselvan P, Sathiaseelan JGR. A comparative study of data mining algorithms for image classification. International Journal of Education and Management Engineering. 2015; 5(2):1–9.
10. Mangai JA, Wagle S, Kumar VS. An improved k nearest neighbor classifier using interestingness measures for medical image mining. International Journal of Medical Health Biomedical Bioengineering and Pharmaceutical Engineering. 2013; 7(9):236–40.
11. Chen CK. The classification of cancer stage in micro array data. Computer Method and Programs in Biomedicine. 2012; 108(3):1070–7.
12. Chen HI, Miser J, Kuan CF, Fang YA, Lam C, Li YC. Critical laboratory result reporting systems in cancer patients. Computer Methods and Programs in Biomedicine. 2013; 111(1):249–54.
13. Lee MY, Yang CS. Entropy based feature extraction and decision tree induction for breast cancer diagnosis with standardize thermograph images. Computer Methods and Programs in Biomedicine. 2010; 100(3):269–82.
14. Su S Q, Zhang C, Huang G, Zhu Y. An intelligent decision support method for diagnosis of colorectal cancer through serum tumor markers, Computer methods and programs in biomedicine, 2010,100(2), 100(2), pp.97-107.
15. Kalinli A, Sarikoc F, Akgun H, Ozturk F. Performance comparison of machine learning methods for prognosis of hormone receptor status in breast cancer tissue samples. Computer Methods and Programs in Biomedicine. 2013; 110(3):298–307.
16. Sartakhti, Zanooei MH, Mozafari K. Hepatitis disease diagnosis using a novel hybrid method based on support

vector machine and simulated annealing. Computer methods and programs in biomedicine. 2012; 108(2):570–9.
17. Amini S, Homayouni S, Safari A. Semi supervised classification of hyperspectral image using random forest method. IEEE International Geoscience and Remote Sensing Symposium; 2014. p. 2866–9.
18. Ruta A, Li Y. Learning pairwise image similarities for multi classification using kernel regression trees. Pattern Recognition. 2012; 45(4):1396–408.
19. Bhuvaneswari C, Aruna P, Loganathan D. A new fusion model for classification of the lung disease using generic algorithm. Egyptian Informatics Journal. 2014; 15(2):69–77.
20. Yin H, Cao Y, Sun H. Combining pyramid representation and adaboost for urban scene classification using high-resolution synthetic aperture radar images. IET Radar Sonar Navigation. 2011; 5(1):58–64.
21. Lin GC, wang WJ, wang CM, sun SY. Automated classification of multi spectral MR images using linear discriminant analysis. Computerized Medical Imaging and Graphics. 2010; 34(4):251–68.
22. Mala K, Sadasivam V, Alagappan S. Neural network based texture analysis of computed tomography images for fatty and cirrhosis liver classification. Applied Soft Computing. 2015; 32:80–6.
23. Zhang X, Hu B, Ma X, Xu L. Resting state whole brain functional connectivity networks for MCI classification using L2 regularized logistic regression. IEEE Transactions on Nano Bioscience. 2015; 14(2):237–47.
24. Davatzikos C, Fan Y, Shen D, Rensick SM. Detection of prodromal Alzheimer's disease via pattern classification of MRI imaging. Neurobiological Aging. 2008; 29(4):514–23.
25. Sreedhar K, Panlal B. Enhancement of images using morphological transformations. International Journal of Computer Science and Information Technology. 2012; 4(1):33–50.
26. Ali AH, Hadi EM, Mazhir SN. Diagnosis of liver from computed tomography images using unsupervised classification with geometrical and statistical features. International Journal of Advanced Research in Computer Science and Software Engineering. 2015; 5(3):28–38.
27. Patil SA, Kuchanur MB. Lung cancer image classification using image processing. International Journal of Engineering and Innovative Technology. 2012; 2(3):37–42.
28. Ramteke RJ, Monali YK. Automatic medical image classification and abnormality detection using k nearest neighbor. International Journal of Advanced Computer Research. 2012; 2(4):190–6.
29. Shenbagarajan A, Ramalingam V, Balasubramanian C, Palanivel S. Tumor diagnosis in MRI brain image using ACM segmentation and ANN-LM classification techniques. Indian Journal of Science and Technology. 2016; 9(1):1–12.
30. Kumar NS, Arun M. Enhanced classification algorithms for the satellite image processing. Indian Journal of Science and Technology. 2015; 8(15):1–9.