Journal of Human Centered Technology Vol. 1: No. 1 (2022) 10 – 17

===

Journal of Human Centered Technology

===

https://humentech.utm.my

Classification of lung cancer stages from CT scan images using image processing and k-Nearest Neighbor

Mohd Firdaus Abdullah[1*], Siti Noraini Binti Sulaiman[1] , Muhammad Khusairi Usman[1], Nor Khairiah A. Karim[2], Ibrahim Lutfi Shuaib[2], Muhamad Daniyal Irfan Alhamdu[1] , Adi Izhar Che Ani[1]

[1]Faculty of Electrical Engineering,
Universiti Teknologi MARA (UiTM), Permatang Pauh ,Pulau Pinang, 13500, Malaysia

[2]Regenerative Medicine Cluster, Advanced Medical and Dental Institute,
Universiti Sains Malaysia, Bertam, Kepala Betas, Pulau Pinang, 13200, Malaysia

*Corresponding Author f.abdullah@uitm.edu.my

Received 10 October 2021; Accepted 6 December 2021; Available online 31 january 2022
https://doi.org/10.11113/humentech.v1n1.6

Abstract:
Lung cancer is the prevalent cause of death among people around the world. The detection of the existence of lung cancer can be performed in a variety of ways, such as magnetic resonance imaging (MRI), radiography, and computed tomography (CT). Such techniques take up a lot of time and financial resources. Nevertheless, for the detection of lung cancer, CT provides a lower cost, fast imaging time, and increased availability. Early diagnosis of lung cancer may help physicians treat patients to minimize the number of deaths. This paper revolves around the categorization of lung cancer stages from CT scan images using image processing and k-Nearest Neighbor. The central objective of this study is therefore to establish an image processing technique for extracting features of lung cancer from CT scan images. Extracting the features from the segmented image can help to detect cancer inside the lung. The purposed method comprises the following steps by using image processing techniques: data collection, data pre-processing, features selection, and lung cancer classification. The pre-processing was done using a median filter to remove noise contained in the images. Three features need to be extracted which are area, perimeter, and centroid. Finally, the set of data with these features were used as inputs for lung cancer classification. By analysis results, the kNN method has a high accuracy of 98.15%.

Keywords: Lung cancer; Image processing; Filtering; k-Nearest Neighbors (kNN); Feature selection

1. Introduction

Lung cancer is one of the leading causes of cancer fatalities in the U.S. and also worldwide. [1]. Based on research from [2], cancer comes second as the main cause of mortalities in the world and contributed to 8.8 million deaths in 2015. Lung cancer became the primary cause of death for men in 2014 in Indonesia [3]. Cancer is the fourth leading cause of death that accounts for 12.6% of all deaths in public hospitals and 26.7% in private hospitals in Malaysia [2]. Furthermore, according to the latest report from the World Health Organization published in April 2011, lung cancer is the leading cause of death among Malaysian men [4]. The classifications are as follows; stage I is when the cancer is confined to the lung. The cancer is confined to the chest in stages II and III, but when the tumor grows larger and more invasive, the tumor is classified as stage III. Stage IV is when cancer spreads from the chest to other parts of the body [5]. Cancer can be carried away in blood or the lymph fluid. When cancer leaves its site and begins to move into a lymph node or another part of the body this process is called metastasis [6].

Detection of lung cancer can be done in several ways, such as using computed tomography (CT) and MRI [7]. For human bodies, organs such as the lungs, liver, pancreas, kidney, and bone, it is most appropriate to use CT scans because CT scans produce cross-section images of the body using computed technology and x-rays that play a key role in the diagnosis of medical conditions [7]. Besides, CT scan is more preferred because it is easy to use and provides accurate classification and foreign mass location [7]. The use of MRI is expensive, less available as compared to CT scans, and time-consuming diagnostics. CT is the most reliable method for early detection of cancer, and this modality is mostly used in treatment methods e.g. radiotherapy. In addition, CT has good classification detection, offers low cost, provides short imaging time, and has widespread availability.

Recently, medical image processing techniques have shown an increasing trend among researchers owing to the need to develop an advanced system of clinical examination and diagnosis [8]. Early diagnosis of lung cancer may help physicians treat patients to minimize the number of deaths. Feature extraction is one of the main paces in the computer vision algorithm and for medical imaging, it is important for segmenting the lung cancer, nodule measurement, and image display [9].

Kumar and Garg [10] used the image processing technique to remove the noise in the image by using filters and segmentation techniques to detect the abnormal region in x-ray images and extract the featured area, perimeter, and shape of the lung nodule. The features then acted as the input for the neural network. Amal et al. [11] in the research of feature-based nodule for the classification in low dose CT scanning, reported that the nodule was detected and the classification process of the type is needed. The SURF and LBP are used to develop the features for describing the texture of the nodule. Gomathi and Thangaraj [12] proposed computer-aided diagnose (CAD) using the Fuzzy Possibilities C Mean (FPCM) algorithm for the segmentation. The rule-based technique was applied to classify the cancer nodule. Extreme Learning Machine (ELM) was performed for a better classification.

Kernel Nearest Neighbors (kNN) algorithm is a nonlinear classification. This algorithm is one of the simplest types of ruled-based classifier and it contains a wide variety of variations. The accuracy of this algorithm is very high and it is applicable for various problems. The kNN algorithm assigns an unknown input sample of its reference sample. The algorithm looks up into the nearest k sample in the reference set instead of the reference set and makes the decision based on the sample which is the nearest k sample [9]. It is also one of the simplest ways of classifying the data [13]. The kNN uses the database by depending on the value k and classifies the new data set by training it to detect the new data to the nearest neighbors and predict the data [14]. It provides an alternative solution to increasing the computational power of the linear machine. The algorithm maps the data into a high-dimensional feature space. If an appropriate kernel is chosen to reshape the distribution of samples the performance may be improved.

Therefore, the objective of this research is to develop an image processing method to extract features of lung cancer on CT scan images. This project aims to determine the suitable features for the lung nodule and compute the dataset with these features as an input for lung cancer classification. The proposed method was developed using MATLAB to detect the cancer nodule in the lung. Section II consists of the related work for research methodology, which involves ethical approval and data collection, image segmentation, features extraction, and classifier. Meanwhile, Section III consists of the results and discussion. The final part is the conclusion for the overall results.

2. Methodology

The overall flowchart of image segmentation features extraction and classification processes is shown in Figure 1. This project was divided into four major parts which were first, data collection and preprocessing, second, image segmentation, third, features extractions, and fourth, classifier. Data collection was the first part, where a data sample was obtained from the databases of Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia. Before the data was collected, The ethic was approved by the Human Research Ethics Committee of USM (JEPeM) under the School of Medical Sciences, USM, IPPT, Bertam, Pulau Pinang. Next was formulating the framework of image segmentation, and, extracting the features and the final part was the classification of the stage of lung tumor using kNN. The details for the methodology are briefly described in the subsection below:

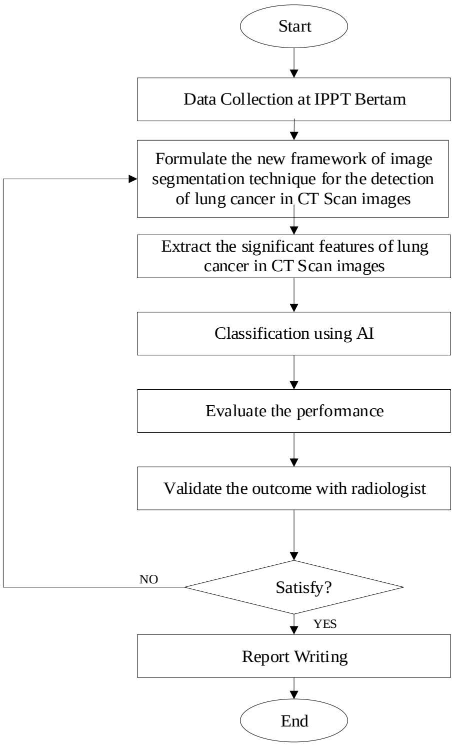
Start
Data Collection at IPPT Bertam
Formulate the new framework of image segmentation technique for the detection of lung cancer in CT Scan images
Extract the significant features of lung cancer in CT Scan images
Classification using AI
Evaluate the performance
Validate the outcome with radiologist
Satisfy?
NO
YES
Report Writing
End

Figure 1. Classification of lung cancer flowchart

2.1 Data Collection and Pre-Processing

This section explains the dataset used in this research. Data from subjects with underlying lung cancer were collected. CT scan images were collected for four subjects with abnormal lung disease and a total of 100 samples of images were obtained from each of the subjects. A Siemens SOMATOMS Definition Flash CT-scan machine was employed to collect images. The image was sorted in a file and loaded into MATLAB software. The format for the images file was in DICOM images. This format is the standard for the storage and transfer of medical images. These images were from CT scans where the CT scan images have low noise. The dimension of the images was 512 × 512 pixels. The pre-processing was done using median filters to remove noise contained in the images. Some unwanted features were removed in the pre-processed images.

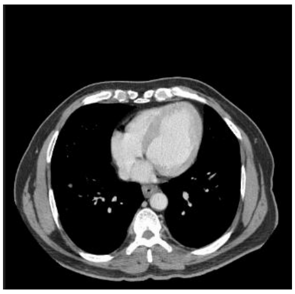

Figure 2. Image from database

2.2 Formulate the Framework of Image Segmentation

The original image from the database was segmented manually by using ImageJ and Adobe Photoshop CS6 software. This software was used in this study as the platform to manually segment the image. The image firstly was uploaded in the ImageJ software and inverted the color of the image. Then, the image was converted into a binary image by using the plugin in the software. The color of the image was the black and white color. When the color was in black and white the threshold of the image was adjusted by 99.70% of thresholding. Then, the image was uploaded into adobe photoshop to remove the background of the images. The background was removed so that the segmentation could be done within the inside of the lung only.

2.3 Features Extraction

After the manual segmentation was performed, the images of the segmented lung only consisted of blood capillaries and cancer. Features extraction was then performed on the images. Any important information extracted from the images was called a feature [5]. This provided more detailed information from the images. For features extraction, geometric features were extracted. Physical dimensional measurements were measured as the shape measurement. Area, perimeter, and centroid were the only features that were considered to measure.

1) Area

The total number of white pixels in the image was the total area of the region of interest [9]. The white pixel

$$\text{Total Pixel in Image} = \text{Width} \times \text{Height} = 512 \times 512$$
$$1 \text{ Pixel} = 0.264 \text{ mm}$$
$$\text{Size of ROI} = S = [P \times 0.264 \text{ mm}^2] \qquad (1)$$

P = total number of pixels in the region of interest

2) Perimeter

A perimeter was the number of pixels of the outline of the region in an image. The sum of the distance between boundary points:

$$P = |S_nS_1| + \sum_{i=1}^{n-1} |SiSi + 1| \qquad (2)$$

3) Centroid

Centroid consisted of 2 elements which were the horizontal coordinate or known as x-coordinate and the vertical coordinate or known as y-coordinate. The other elements were in order of dimension for the centroids.

2.4 Classifier

The kNN was used to classify the input pattern of two classes which were Stage 1, 2, 3 and 4. After the data for the features were extracted, kNN classified the data in the form of a vector. kNN is a statistically-based method for data classification. Euclidean, Hamming, and Mahalnobis are the type of data classifications that are usually used. The Eucilidean formula between the sample input and member is elaborated as below:

$$d_{st}^2 = (X_s - Y_t)(X_s - Y_t) \qquad (3)$$

The value for X_s was the member of the data while the Y_t was the target data. The sample data was declared as the input feature vector for the kNN to determine its closest members depending on the value for *k*. The *k* was the value of the member that the sample points will predict the result. *k* was depending on the parameters that were extracted. If the value of *k* is perfect it will produce a high accuracy result. For this research, 80% of the data was set as the training input while the other 20% was set as the sample data [10]. For the accuracy of this research the following equation was used:

$$Acc.(\%) = \frac{x}{y} X\ 100 \qquad (4)$$

where *x* refers to the total number of levels correctly classified, and *y* refers to the total number of input data.

3. Results and Discussion

This section shows the results and discussion in sequence. Three features were used for the image pre-processing procedure which was the area, perimeter, and centroid. The technique was applied on 100 CT images from each patient with lung cancer.

3.1 Results of Manual Segmentation

The most difficult part was to remove the touching object in the images. By implementing the manual segmentation for the images with ImageJ and Adobe Photoshop CS6, it was easier to segment the boundaries from the images. The image was converted into a binary image where the color only consisted of white and black. Figure 3 shows the original image was inverted into a binary image by using the ImageJ.

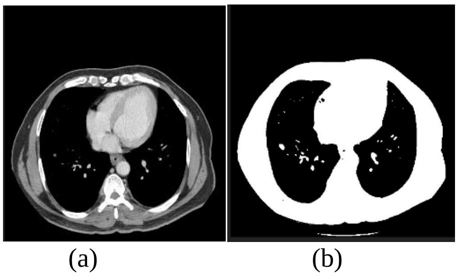
(a) (b)

Figure 3. (a) Original image and (b) binarize image

3.2 Results of Features Extraction

The image from the patient contained four images where lung cancer was present. The objects in each image from the patient were not the same because some of the objects in the images were the blood capillaries. The position of the blood capillaries changed from one image to the next, and in some cases, it did not appear. However, while lung cancer also appeared in some images but not others, the position of the lung cancer always remained in the same place.

This is why the centroid parameters were required to check the occurrence of the object in the image. If there were huge differences in the centroid from the first image that meant the object was the blood capillary while if the centroid were static and the difference of the centroid was only small that meant the object was a suspected cancer. Features extraction was one of the most important stages in this study to decide the stage of lung cancer. Three features were considered to be extracted from the image.

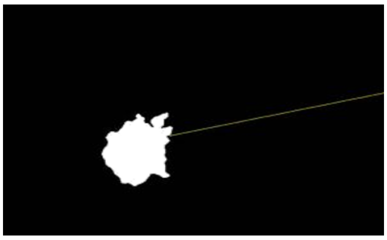

Figure 4. Segmented nodule

The white pixel in the image that was considered as 1 was extracted to give the area of the region. The total of the outline of the white pixel was the perimeter value and the center of the nodule was for the centroid. The estimated features for the sample images were as follow:

- Area: 141
- Perimeter: 41.3
- Centroid: 130.2099,296.2099

All values were extracted in terms of pixel values only. The three features continued to be extracted from the other images. Table 1 shows the values that were extracted from these parameters with different images. Area, perimeter, and centroid would act as a base for the classification process.

Table1. Results for feature extraction

Image	Features		
	Area	**Perimeter**	**Centroid**
Im 1	154	45.056	115,319
Im 2	101	34.189	115,319
Im 3	195	69.759	320,344
Im 4	247	76.102	320,344

3.3 Classification using kNN

The smallest growth in the lung was the lung nodules which were measured between 5mm to 25 mm in size [15]. The bigger size was called the malignant nodules where the size was greater than 25mm and it also tended to grow faster than the lung nodules. By using the segmentation the lung nodule was detected and the features were extracted to determine the stages of the lung cancer. The measurement of the extent to which cancer had spread was the measurement of the stage of lung cancer [16]. It is important to know the staging for cancer to know how particular cancer should be treated.

The four stages were listed in order of severity:

a. Stage1, the cancer is only in the lung

b. Stage 2 and 3, cancer has spread to other parts of the chest

c. Stage 4, cancer has spread to other parts of the body from the chest

The medical experts had decided to use the TNM system for the staging of non-small cell lungs. The 'T' was for the extent of the primary tumor, 'N' was for the regional lymph node, and 'M' for metastasis. Tables 2 and 3 show the criteria that were decided by the experts in the medical field for the classification of lung cancer stages [13]. Based on Table 3, the classification of lung cancer stages is all in stage 2 since all the subjects were chosen randomly.

Table 2. Stages of cancer

Primary Cancer	Criteria
C1	X < 3 cm
C2	< 3 cm × < 7 cm
C3	X > 7 cm
C4	Any size larger than above

Table 3. Classification results

Image	Area	Perimeter	Centroid	Stages
Im1	154	103.254	155,299	2
Im2	164	81.296	150,298	2
Im3	118	96.679	130,296	2

3.4 Evaluating the Performance

Based on the classification for the stage of the tumor, the system was evaluated by using the substitution losses in the classification. The kNN predicted that the incorrect classification was 37/2000 or 0.0185. So, therefore, the accuracy of the study was 98.15%.

4. Conclusion

This study was successfully carried out to classify the lung cancer stages using kNN. Three features were extracted which were area, perimeter, and centroid. The set of data with these features were used as inputs for lung cancer classification. Based on the analysis of the results, the kNN method has a high accuracy of 98.15%. The results show that this research has the potential to determine the stages of lung cancer. Future work should focus on the data collection so that it can classify different stages of lung cancer.

Acknowledgment

This study was financially supported by the Ministry of Higher Education Grant Scheme (FRGS) for the project, "Formulation of A New Watershed-based Segmentation Method for Features Extraction of Lung Cancer Nodule in CT-scan Images" (Ref: FRGS/1/2019/TK04/UITM/02/15) and the ethics from Universiti Sains Malaysia (USM/JEPeM/19040231). The authors are thankful to Advanced Control System and Computing Research Group (ACSCRG) and Research Group Advanced Rehabilitation Engineering in Diagnostic and Monitoring (AREDiM), Universiti Teknologi MARA Pulau Pinang for their contribution that led to the completion of this study

References

[1] P. Su, J. Yang, K. Lu, N. Yu, S. T. Wong and Z. Xue, A fast CT and CT-fluoroscopy registration algorithm with respiratory motion compensation for image-guided lung intervention, IEEE Transactions on Bio-medical Engineering, 2013, 60(7):2034–2041. https://doi.org/10.1109/tbme.2013.2245895

[2] A. A. M. Asmayani Khalib, Malaysian study on cancer survival (MySCan), National Cancer Registry National Cancer Institute, Malaysian Ministry of Health, 2018, 4. https://vdocument.in/malaysian-study-on-cancer-survival-mohgovmy-cancer-registry-national-cancer.html

[3] Y. F. Riti and H. A. Nugroho, Feature extraction for lesion margin characteristic classification from CT scan lungs image, 1st International Conference on Information Technology, Information Systems and Electrical Engineering, 2016, 54–58. https://doi.org/10.1109/ICITISEE.2016.7803047

[4] M. Assefa, I. Faye, A. S. Malik and M. Shoaib, Lung nodule detection using multi-resolution analysis, International Conference on Complex Medical Engineering, 2013, 457–461. https://doi.org/10.1109/ICCME.2013.6548290

[5] S. Singh, Y. Singh and R. Vijay, An evaluation of features extraction from lung CT images for the classification stage of malignancy, IOSR Journal of Computer Engineering, 2015, 78–83.

[6] C. Science and M. Studies, Lung cancer detection from CT image using image processing techniques, International Journal of Advance Research in Computer Science and Management Studies, 2015, 3(5):249–254.

[7] S. Anitha, L. Kola, P. Sushma and S. Archana, Analysis of filtering and novel technique for noise removal in MRI and CT images, 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2017, 1–3. https://doi.org/10.1109/ICEECCOT.2017.8284618

[8] A. Ravishankar, S. Anusha, H. K. Akshatha, A. Raj, S. Jahnavi and J. Madhura, A survey on noise reduction techniques in medical images, International Conference of Electronics, Communication and Aerospace Technology, 2017, 385–389. https://doi.org/10.1109/ICECA.2017.8203711

[9] K. M. M. Tun and A. S. Khaing, Feature extraction and classification of lung cancer nodule using image processing techniques, International Journal of Engineering Research & Technology, 2014, 3(3):2204–2210.

[10] V. Kumar and K. Garg, neural network based approach for detection of abnormal regions of lung cancer in X-ray image, International Journal of Engineering Research & Technology, 2012, 1(5):1–7.

[11] A. Farag, A. Ali, J. Graham, S. Elhabian, A. Farag and R. Falk, Feature-based lung nodule classification, in: G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. Chung, R. Hammound, M. Hussain, T. Kar-Han, R. Crawfis, D. Thalmann, D. Kao and L. Avila, Advances in Visual Computing, 2010, 79–88. https://doi.org/10.1007/978-3-642-17277-9_9

[12] M. Gomathi, P. Thangaraj, A computer aided diagnosis system for detection of lung cancer nodules using extreme learning machine, International Journal of Engineering Science and Technology, 2010, 2(10):5770–5779.

[13] R. Kaur, S. Guru and G. Sahib, Feature extraction and principal component analysis for lung cancer detection in CT scan images, Medicine, 2013, 15536402.

[14] N. M. Somari, M. F. Abdullah, M. K. Osman, A. M. Nazelan, K. A. Ahmad, S. P. R. S. Appanan, L. K. Hooi, Particles contaminations detection during plasma etching process by using k-Nearest Neighbors and fuzzy k-Nearest Neighbors, International Conference on Control System, Computing and Engineering, 2017, 512– 516. https://doi.org/10.1109/ICCSCE.2016.7893630

[15] A. Kulkarni and A. Panditrao, Classification of lung cancer stages on CT scan images using image processing, International Conference on Advanced Communications, Control and Computing Technologies, 2015, 978:1384–1388. https://doi.org/10.1109/ICACCCT.2014.7019327

[16] M. V. A. Gajdhane and P. D. L.M, Detection of lung cancer stages on CT scan images by using various image processing techniques, IOSR Journal of Computer Engineering, 2014, 16(5):28–35.