Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 45 No. 5
May 2024

Early Detection of Lung Cancer of CT Scans in Biomedical Image Processing Using Feature Extraction Methods and Support Vector Machine (SVM) Classification

Retz Mahima Devarapalli[1], Sajja Tulasi Krishna[2], Hemantha Kumar kalluri[3]
[1]Research Scholar,Department of Computer Science & Systems Engineering,Andhra University College of Engineering,Andhra University, Visakhapatnam, AP, India.
[2]Department of Computer Science and Engineering,Koneru Lakshmaiah Education Foundation,Green Fields, Vaddeswaram, AP, India, 522302
[3]Department of Computer Science and Engineering,SRM University, AP, India

Abstract
Introduction: Lung cancer remains a significant cause of mortality, underscoring the critical need for early detection to improve survival rates. Despite existing methods for lung cancer detection, accuracy enhancements remain a priority. In this paper, we propose a novel approach utilizing Fuzzy-C-Means (FCM) and Support Vector Machine (SVM) classification for lung cancer detection.
Objectives: The primary objective of this study is to develop a robust lung cancer detection method that outperforms existing approaches in terms of accuracy.
Methods: The proposed methodology comprises four key steps:
Pre-processing: Raw images undergo pre-processing using Median Filter (MF) to enhance their quality and reduce noise.
Segmentation: The pre-processed images are segmented using the Fuzzy-C-Means algorithm (FCM), which partitions the image into distinct regions, facilitating subsequent analysis.
Feature Extraction: Local Binary Pattern (LBP) is employed to extract discriminative features from the segmented images. LBP is known for its effectiveness in capturing texture information, making it suitable for our classification task.
Classification: Extracted features are fed into Support Vector Machine (SVM) for classification. SVM is chosen for its ability to handle high-dimensional data and its robustness in classification tasks.
Results: Experimental evaluations were conducted on two standard benchmark datasets: LIDC Dataset and SPIE-AAPM Dataset. The findings demonstrate the superiority of our proposed approach over state-of-the-art methods. Specifically, our method achieves higher accuracy in lung cancer detection, thereby validating its effectiveness in improving early detection rates.
Conclusions: The proposed approach combining Fuzzy-C-Means segmentation and SVM classification presents a promising solution for enhancing lung cancer detection accuracy. By leveraging advanced image processing techniques and machine learning algorithms, our method demonstrates superior performance compared to existing approaches. These findings underscore the potential of our method to contribute significantly to early detection efforts, ultimately leading to improved patient outcomes in the fight against lung cancer.

Keywords: Fuzzy-C-Means, Local Binary Pattern, Support Vector Machine, Median Filter

1. Introduction

Cancer is an abnormal growth of cells that prevents the functioning of the body and spreads innumerable diseases. Cancer is considered the deadliest disease in the world. Lung cancer is men's most particularly occurring common cancer and women's third most common cancer. There were 2 million cases in 2018 which contributes to 12.3% of all cancers [1]. Lung cancer is an uncontrollable development of anomalous cells in either one or both lungs. It prevents the respiratory organ from functioning correctly and blocks the air duct passage. Therefore, preventing the respiratory organ from the nourishment of the body with

oxygen fully doesn't occur. Those unusual cells can persist on replicating, and that results in the formation of a tumour. Those tumours may be benign. If the cells remain in one place or malignant, the cells disperse throughout the body via bloodstreams, which might prove to be life-threatening.

The burden of lung cancer could be decreased by early identification of cancer and treatment of patients with cancer. Many diseases have a high opportunity to cure if properly diagnosed early. Significant improvements could be made in the lives of cancer patients by early cancer detection and avoiding delays in care. With the accelerated advancements in technology development, an important domain in the research is to detect lung cancer.

2. Related work

To detect cancer, several phases have been suggested and implemented variously by researchers earlier. The general strategy outlined for detecting cancer typically included four steps, namely, 1) Pre-processing,.2) Segmentation, 3) Feature Extraction 4) Classification.

Tim Adams et al. [2] applied marker-based watershed segmentation (WST), i.e., to separate the lung tissues and image background. In the next stage, feature extraction, several texture features are extracted. Later, the extracted features are fed to Support Vector Machines (SVM) for classification. Asuntha et al. [3] applied image enhancement, gray level conversion in the first stage of pre-processing. Superpixel segmentation algorithm was used to perform segmentation. Later, feature attributes are drawn forth from the segmented lung image. . The SVM algorithm is used for classification, and 89.5% accuracy was reported. Suren et al. [4] applied median and Gaussian filters on the lung images in the pre-processing stage. Watershed segmentation is applied to perform segmentation. Eccentricity, centroid, area, mean intensity, and perimeter values are considered as features. In the classification phase, SVM is used as a classifier that classifies the given data into two classes that resulted in an accuracy of 92%.

Deep Prakash et al. [5] converted the original lung images to grayscale images in the pre-processing stage. Later, in the segmentation stage, Discrete Wavelet Transform (DWT) was used to extract the Region of Interest (ROI). GLCM was used to generate a feature vector consisting of energy, entropy, and mean. The extracted features are fed to SVM for classification. Swati et al. [6] applied image de-noising and optimal thresholding, which is the pre-processing stage. Morphological operations are used for segmentation. GLCM was used to extract features such as energy, correlation, area, homogeneity. The extracted features are fed to SVM classification. Diego et al. [7] used a two-step process of 2D segmentation, and connectivity analysis along with a 3D blob algorithm was used in the segmentation stage. Geometrical features and histogram measurements are extracted to generate the feature vector. The generated features are used to train the SVM classifier which resulted in an accuracy of 89.1%. Sasidhar et al. [8] proposed an active contour technique for the segmentation of lung images. The features are extracted from HARALICK texture features. Later, SVM was used for classification and reported as 92% classification accuracy. Devarapalli et al. [9] surveyed the detection of lung cancer. The researchers cover the different techniques for the detection of early stages of cancer.

3. Proposed work

The proposed method consists of four modules, firstly pre-processing, secondly, segmentation, thirdly feature extraction, and finally, classification. In the pre-processing module, the median-filter, digital filtering method, which is non-linear, is applied to eliminate the noise. In the segmentation module, to extract and preserve only lung partitions from the pre-processed images and to obliterate the unnecessary portions from the image, i.e., removing the surrounding information, the Fuzzy Logic C-Means algorithm is applied. In the feature-extraction phase, to generate the attribute-feature vector from the segmented lung portion of the image, the Local Binary Pattern (LBP) algorithm is applied. To train and evaluate with the SVM Classifier to classify benign or malignant images, these sampled LBP features are used [10]. The block diagram of the proposed work is shown in Figure 1. The brief description of these modules is explained in the following subsections:

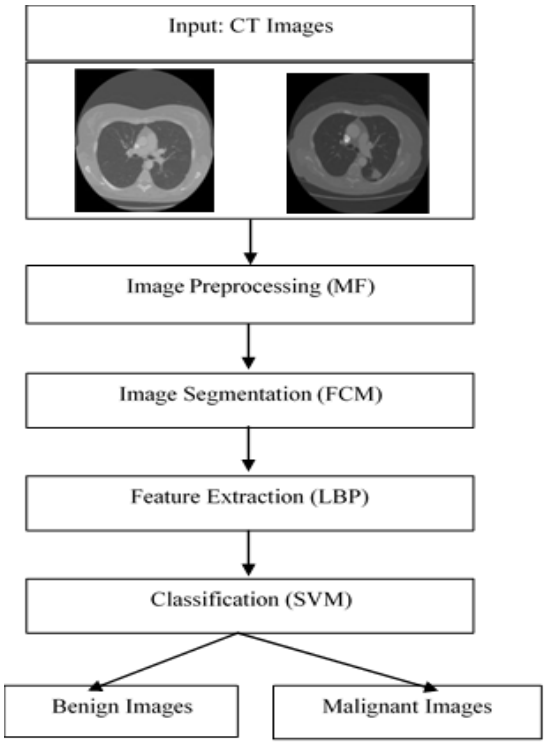


Figure 1. Proposed System Architecture

3.1 Image Pre-processing

Image Pre-processing is used to enhance the image quality, which reduces any noise present in the image [11], thereby providing an enhanced image for further effective processing. In this step, initially, all the images are converted to grayscale images; after that, Median Filter (MF) is applied to reduce the distortions during the acquisition of images. The median filter is an operation that is non-linear and reduces the noise while preserving the information of images [12]. Sample images before applying the median filter and after using the median filter are shown in Figure 2.

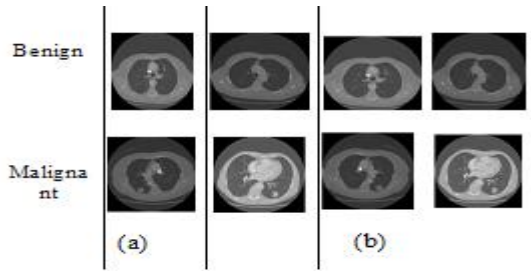


Figure 2. Sample images of Pre-processing stage (a) Original Images (b) After applying Median Filter

3.2 Segmentation

Image Segmentation is a technique used commonly for extracting the desirable part of an image into several regions, which are used for better identification [13]. The segmentation of medical images is widely used to detect the anomaly. In the proposed work, to do the segmentation Fuzzy C Means algorithm is applied on CT lung images. Fuzzy C Means segmentation algorithm retains much information and generates a C partition optimal by minimizing the weights within the group SSE objective function JFCM [14].

Basic C-Means Algorithm:

A C-means algorithm classifies objects into c disjoint subsets Gi where i = 1,2,...,c. Also called clusters. A cluster center is determined in each cluster (Vi).

Step 1: Consider c initial values for Vi cluster centers.

Step 2: To the nearest center of the clusters, allocate all objects Xk.

Step 3: Calculate new cluster centroids.

Step 4: Stop if the clusters are convergent; else, go to Step2.

A cluster is said to be convergent if no object changes its membership or no centroid changes its position.

Necessary Steps of Fuzzy C-Means:

Consider NxC Membership matrix U = (u_ki) where (1<= k <= N, 1<=i<=c) in which u_ki takes a real value. To obtain fuzzy memberships non-linearity for U is considered, given by:

$$J_{fcm}(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ki})^m D(x_k, v_i), (m > 1) \quad (1)$$

Stpe1: Generate c initial values for V_i cluster centers. $\bar{V} = (\overline{v_1}, \dots, \overline{v_c})$.

Step2: Calculate $\bar{U} = \min_{U \in U_f} J_{fcm}(U, \bar{V})$ and the solution is given by

$$\overline{u_{kl}} = \left[\sum_{j=1}^{c} \frac{D(x_k, \bar{v}_i)^{\frac{1}{m-1}}}{D(x_k, \bar{v}_j)} \right]^{-1} \quad (2)$$

Step 3: Calculate $\bar{V} = \min_{V} J_{fcm}(\bar{U}, V)$ and is given by,

$$\overline{v_l} = \frac{\sum_{k=1}^{N} (\bar{u}_{kl})^m x_k}{\sum_{k=1}^{N} (\bar{u}_{ki})^m} \quad (3)$$

Step 4: Stop, if $\bar{U}$ or $\bar{V}$ is convergent; else go to step2 [15].

Sample Images before applying the segmentation and after applying the segmentation are shown in Figure 3.

3.3 Feature Extraction

Feature Extraction retrieves the relevant information of an image which helps to fasten the classification process. In the proposed model, the Local Binary Pattern (LBP) extracts the features as vector form [16]. LBP thresholds are neighboring pixels with the present pixel value that is an efficient image descriptor texture. In this process, all the texture features of the images are extracted, and when these features are used for classification [17]. LBP feature extraction algorithm is to distinguish the gradation successive difference values of the pixel points in the central part of the image, the edge part, and the quarter part to represent the local texture, which is essential

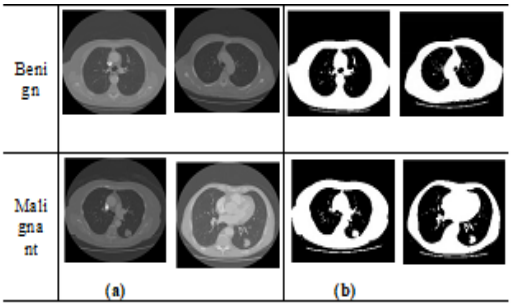


Figure 3. Sample images of segmentation stage (a) Before applying segmentation (b) After applying segmentation

feature information local texture, which is essential feature information of each part in the image. Then, a binary value string is generated, where the starting digit value on the binary string (I = 1, 2, … , 8) and is converted as the LBP value. A 3*3 window is slided to extract the information is done as by comparing the adjacent pixels with center pixel and, if the value being compared is greater than the center pixel value, then it is considered as 1 and if the current pixel value being compared is less than the center pixel value, then it is considered as 0. The LBP algorithm analysis is shown in Figure 4.

LBP Algorithm:

1.Choose P neighboring pixel (8 neighbors) at a distance R for every pixel (a, b) in an image.

2. Compute the intensity value difference between the current pixel (a,b) with neighboring pixels P.

3.Threshold the differences in intensities, so that 0 is allocated to all the adverse variations and all the favorable variations are allocated to 1, resulting in a bit vector.

4.Convert the vector of P-Bit to its respective decimal-value and substitute this decimal value for the intensity value at (a,b).

Thus, the LBP features for every pixel is given in Eq. 4.

$$LBP(P,R) = \sum_{P=0}^{P-1} f(gp - gc)2P \quad (4)$$

where gc indicates the intensity of the current pixel value and gp the intensity of the neighboring pixel values. At a radius R, P is the neighboring pixels selected.

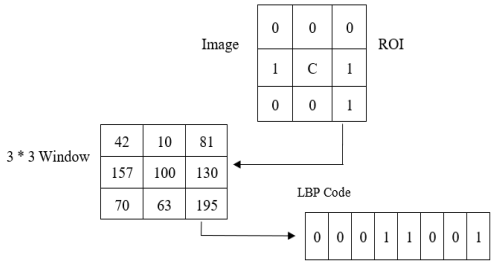


Figure 4. Local Binary Pattern algorithm process

3.4 Classification

The next and last step of the proposed system is the classification of the normal lung and abnormal lung images by using the SVM Classifier [18]. The SVM classification method is a binary, formed from the two classes which take input labeled data and generates a model for classifying the new labeled data into either one of the two classes.

The extracted features are provisioned to the SVM classifier for training which could be able to classify the data [19]. To separate the classes, any number of hyperplanes can be considered. Support Vector Machine finds a plane with the maximum distance between the classes which is considered as Maximum Margin, for better classification. Figure 5 depicts the Support Vector machines.

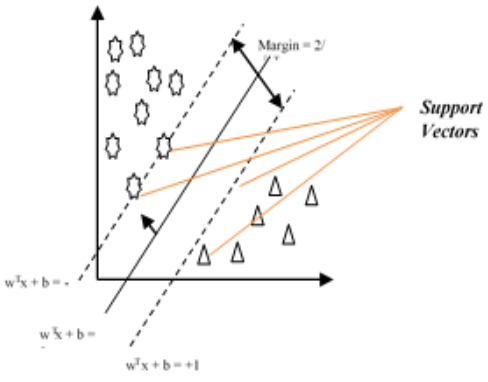


Figure 5. Support Vector Machine

The regular way is to separate binary classes of data with a straight line (i.e., one-dimension), flat plane (i.e., two-D) or an N-D hyperplane. In real-time, there are situations to handle non-linear regions to separate the groups. The kernel function of SVM regulates this type of situations. The non-linear data is transformed by the Kernel Function into a feature space of higher-dimensional into the linear separation. SVM has different kernels such as linear [20], non-linear, polynomial, Gaussian kernel, Radial basis function (RBF) [21], sigmoid, etc. In this paper the proposed system used RBF kernel. Eq. 5 shows the RBF kernel functionality.

$$K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \quad (5)$$

where $\|x-x'\|^2$ is the squared euclidean distance of two data points x and x'. gamma and C are two parameters of the RBF kernel. If the value of gamma is high, then the decision-boundary curve is also high. C is a constant which is preferably low, and then the classifier escaped from heavily penalized misclassified data.

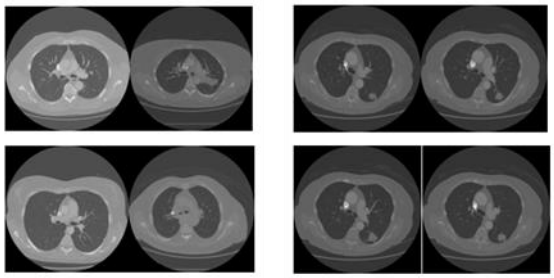
4. Dataset preparation

To test the effectiveness of the proposed approach, two benchmark datasets were used.

4.1 Dataset1

One is the Lung Image Database Consortium (LIDC) database [22] which is a public database. This database is a collection of Computed Tomography (CT) images of 1018 patients, where the images are in the Digital Imaging and Communication ins Medicine (DICOM) format, which is used as a standardized set in the medicine. This database along with the CT images, annotations are provided for each patient are marked by the radiologists. The annotations process is done in two stages, where in the early stage each individual radiologist marks the annotations individually. In the next stage, the results marked by the individual radiologists are again presented and these are re-analyzed, and the annotations were again marked individually. Each individual annotations marked by the radiologists are recorded in the XML file.

These dataset images need some pre-processing according to respective problems. As part of this, we converted the images of DICOM format into Joint Photographic Experts Group (.jpg) format and also found marked slices of each patient's scanned images by radiologists records through the XML file automatically. After this process, some of the slices were retrieved, among all, for each patient. Later, manually, by considering the XML files provided, separated all malignant and benign slices of every patient into different directories; which is done by the nodule characteristics. If the malignancy >3 those patients, scanned slices are considered as malignant or considered as benign. Finally, we retrieved 7328 CT images of 1006 patients. Sample collection of the benign slices and malignant slices are shown in Figure 6.

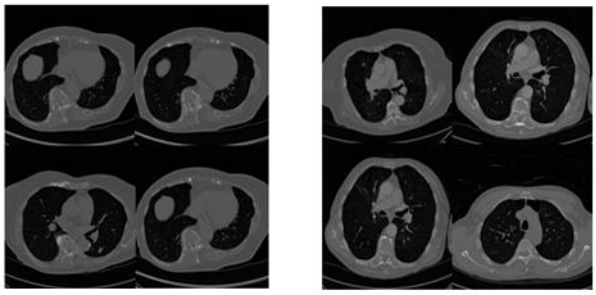


(a) (b)

Figure 6. Sample images of LIDC dataset (a) Benign Slices (b) Malignant Slices

4.2 Dataset2

The other dataset used in this work is the Lung CT challenge, which was sponsored by the Society of Photographic Instrumentation Engineers (SPIE) along with American Association of physicists in Medicine (AAPM) [23] and the National Cancer Institute (NCI). The dataset contains CT scans of 70 patients with a total of 22,489 images. The nodule locations and diagnoses were given in a special excel file. Sample collection of the respective dataset benign slices and malignant slices are shown in Figure 7.



(a) (b)

Figure 7. Sample images of SPIE–AAPM Lung CT Images (a) Benign Slices (b) Malignant Slices

Training	Testing	Existing System			Proposed System
		Author	Method	Accuracy	Accuracy
16 Images	5 Images	Suren et al. [4]	Watershed segmentation, SVM	92%	100%
80%	20%	Antonio et al. [25]	Genetic algorithm, SVM	93.19 %	100%
10-fold CV	10-fold CV	Henry et al. [24]	Support Vector Machine	84.85 %	96.02 %

Table 1. Experimental Results on LIDC Dataset

Training	Testing	Existing System			Proposed System
		Author	Method	Accuracy	Accuracy
10 Images	60 Images	Elmar et al. [26]	Support Vector Machine (SVM)	78.08%	91.66%
10 Images	73 Images	Tim Adams et al. [2]	Support Vector machine	58.09%	82.43%

Table 2. Experimental Results on SPIE – AAPM Lung CTx Challenge

5. Implementation and results

The first three stages of the proposed work (preprocessing, segmentation, feature extraction) is implemented in Matlab 2018a and the classification using Support Vector Machine is implemented in Python. The Support Vector Machine (SVM) classifier is used for training and testing on both the LIDC dataset and SPIE – AAPM Lung CT Challenge images.

5.1 Experiments on LIDC Dataset

The researchers Suren et al. [4] used watershed segmentation with SVM images as 16 training and 5 testing images, they got 92% accuracy. The proposed approach is also tested with 16 training and 5 testing images and achieved 100% classification accuracy. Antonio et al. [25] used Genetic algorithm with SVM for 80% training and 20% testing images, they got 93.19% accuracy. The proposed approach is also tested with 80% training and 20% testing and achieved 100% classification accuracy. Henry et al. [24] used 10-fold cross-validation (CV) on entire dataset, they got 84.85% accuracy. The proposed approach is also tested with 10-fold CV on entire dataset and got 96.02% accuracy. These results are placed in Table 1. The experimental results clearly shows that the proposed approach provides better results compared with earlier researcher results [4, 24, 25].

5.2 Experiments on SPIE-AAPM Dataset

The researchers Elmar et al. [26] used SVM for 10 training images and 60 testing images, they got 78.08% accuracy. Our proposed approach is also tested with 10 training images and 60 testing images and achieved 91.66% classification accuracy. The researchers Tim Adams et al. [2] used SVM for 10 training images and 73 testing images, they got 58.09% accuracy. The proposed approach is also tested with 10 training images and 73 testing images and achieved 82.43% classification accuracy. These results are placed in Table 2. The experimental results clearly shows that the proposed approach provides better results compared with earlier researcher results [2, 26].

6. Conclusion

Early detection of lung cancer is so important which might increase the chances of patient's survival rates and expectancy. Our proposed work has been implemented by applying the Fuzzy C Means algorithm for segmentation and features were extracted using Local Binary Pattern which was fed to SVM for classification. Experiments were conducted with two benchmark datasets LIDC and SPIE-AAPM. The experimental results shows that the proposed work gives better results than state-of-the-art methods.

References

[1] https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data.

[2] Tim Adams, Jens Dörpinghaus, Marc Jacobs & Volker Steinhage (2018). Automated lung tumor detection and diagnosis in CT Scans using texture feature analysis and SVM. Communication Papers of the Federated Conference on Computer Science and

Information Systems, 17, 13–20. Doi : 10.15439/2018F176.

[3] Asuntha, A., A.Brindha, S. Indirani, A. Srinivasan (2016). Lung cancer detection using SVM algorithm and optimization techniques. Journal of Chemical and Pharmaceutical Sciences 9(4), 3198-3203.

[4] Suren M, P.W.C. Prasad, AbeerAlsadoon, A.K.Singh, A.Elchouemi (2018). Lung cancer detection using CT scan images. Procedia Computer Science 125, 107-114. https://doi.org/10.1016/j.procs.2017.12.016.

[5] Kaucha Deep Prakash, P.W.C. Prasad, AbeerAlsadoon, A.Elchouemi, SasikumaranSreedharan (2017). Early detection of lung cancer using SVM classifier in biomedical image processing. 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI).IEEE. doi: 10.1109/ICPCSI.2017.8392305.

[6] Tidke, Swati P., and Vrishali A. Chakkarwar (2012). Classification of lung tumor using SVM. International Journal Of Computational Engineering Research 2(5): 1254-1257.

[7] Peña, Diego, ShouhuaLuo, and Abdeldime Abdelgader (2016). Auto diagnostics of lung nodules using minimal characteristics extraction technique. Diagnostics 6(1),1- 13. doi: 10.3390/diagnostics6010013.

[8] Sasidhar, B., G. Geetha, B.I.Khodanpur, Ramesh Babu (2017). Automatic Classification of Lung Nodules into Benign or Malignant Using SVM Classifier. Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications. Springer, Singapore, 2017. https://doi.org/10.1007/978-981-10-3156-4_58.

[9] Devarapalli, R. M., Kalluri, H. K., & Dondeti, V. (2019). Lung Cancer Detection of CT Lung Images. International Journal of Recent Technology and Engineering (IJRTE), 7(5S4), pp. 413-416, ISSN: 2277-3878 (Online). https://www.ijrte.org/wp-content/uploads/papers/v7i5s4/E10870275S419.pdf

[10] Xiaobing Peng, Yuquan Zhu. (2018). An Improved Support Vector Machine Algorithm Based On Minimum 2-Norm. Revue D'intelligence Artificielle; Cachan . 32(5-6), 719-728. doi:10.3166/Ria.32.719-728

[11] Kumar, N., & Nachamai, M. (2012). Noise removal and filtering techniques used in medical images. Indian Journal of Computer Science and Engineering, 3(1), 146-153. doi: http://dx.doi.org/10.13005/ojcst/10.01.14.

[12] https://in.mathworks.com/help/images/ref/medfilt2.html.

[13] https://in.mathworks.com/discovery/image-segmentation.html.

[14] Yang, Yong, and Shuying Huang (2012). Image segmentation by fuzzy c-means clustering algorithm with a novel penalty term. Computing and Informatics 26(1), 17-31.

[15] Miyamoto, Sadaaki, et al. Algorithms for fuzzy clustering. Heidelberg: Springer, 2008.

[16] Wasim, M., Aziz, A., & Ali, S. F. (2017). Object's shape recognition using Local Binary Patterns. International Journal of Advanced Computer Science and Application, 8(8), 258-262. doi: 10.14569/IJACSA.2017.080833.

[17] Sengupta, N., Sahidullah, M., & Saha, G. (2017). Lung sound classification using local binary pattern. 1-21. arXiv preprint arXiv:1710.01703.

[18] Chan, Y. H., Zeng, Y. Z., Wu, H. C., Wu, M. C., & Sun, H. M. (2018). Effective pneumothorax detection for chest X-ray images using local binary pattern and support vector machine. Journal of healthcare engineering, 2018.https://doi.org/10.1155/2018/2908517.

[19] Kailasam, S. P., & Sathik, M. M. (2019). A Novel Hybrid Feature Extraction Model for Classification on Pulmonary Nodules. Asian Pacific journal of cancer prevention: APJCP, 20(2), 457-468. doi:10.31557/APJCP.2019.20.2.457

[20] Alsallal, M., Sharif, M. S., Hadi, B., & Albadry, R. (2019). Decision support detection system for lung nodule abnormalities based on machine learning algorithms. Journal of Contemporary Medical Sciences, 5(3). http://www.jocms.org/index.php/jcms/article/view/615

[21] Polat, H., & Danaei Mehr, H. (2019). Classification of Pulmonary CT Images by Using Hybrid 3D-Deep Convolutional Neural

Network Architecture. Applied Sciences, 9(5), 940. https://doi.org/10.3390/app9050940.

[22] https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

[23] https://wiki.cancerimagingarchive.net/display/Public/SPIE-AAPM+Lung+CT+Challenge

[24] Krewer, Henry, Benjamin Geiger, Lawrence O.Hall, Dmitry B. Goldgof, YuhuaGu, Melvyn Tockman, Robert J. Gillies (2013). Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography. 2013 IEEE International Conference on Systems, Man, and Cybernetics.IEEE. doi: 10.1109/SMC.2013.663.

[25] Antonio Oseas de CarvalhoFilho, Aristofanes Correa Silva, Anselmo Cardoso de Paiva, Rodolfo AcatauassuNunes, Marcelo Gattass (2017). Computer-aided diagnosis system for lung nodules based on computed tomography using shape analysis, a genetic algorithm, and SVM. Medical & biological engineering & computing 55(8) , 1129-1146. https://doi.org/10.1007/s11517-016-1577-7.

[26] Rendon-Gonzalez, Elmar, & VolodymyrPonomaryov (2016). Automatic Lung nodule segmentation and classification in CT images based on SVM. 2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW).IEEE. doi: 10.1109/MSMW.2016.7537995.