CLASSIFICATION OF LUNGS IMAGES FOR DETECTING NODULES USING MACHINE LEARNING

Hussein Hamdan and Umar Alqasemi

Dept. of Electrical and Computer Engineering,
King Abdulaziz University, P.O.Box 80200, Jeddah 21589, Saudi Arabia

ABSTRACT

Lung nodules are tiny lumps of tissue that are common in the lungs. The nodule may be benign or malignant; malignant nodules are cancerous and can grow rapidly. For a long time, X-ray images of the chest have been utilized to diagnose lung cancer. We developed in this paper a computer aid diagnosis system (CAD) to atomically classify a set of lung x-ray images into normal and abnormal (with nodule and no-nodule) cases.

We used 180 images in this work, the images are in full size no filtering or segmenting process were applied, 75 of them are for normal cases and the other 105 are for abnormal cases, at the same time 120 of the images have been used to train the classifier and 60 for testing.

Our classifiers were fed with a variety of features, including LBP (local binary pattern) and statistical features. And a classifier was able to identify cases with nodule from cases without nodule with an accuracy (ACC) of 86.7%

KEYWORDS

Image classification, Lung cancer, support vector machine, Lung nodule, machine learning, MATLAB, CAD.

1. INTRODUCTION

Lung nodules are tiny lumps of tissue seen in the lungs. They appear on the computerized tomography (CT) and X-ray of the chest like white shadows. The size of these masses is variated from 0.5 to 3 cm [1].

The nodule can be benign or malignant, Benign nodules are noncancerous while the Malignant nodules are cancerous and can grow so quickly. If the size of the nodules are more than 3cm it can be called masses[2]. Usually, those nodules are not cancer They may be a result of old infections or scar tissue or any other reason but test are always needed to get sure that they are not a cancer[3].

Lung cancer is the leading cause of cancer deaths worldwide and it's one of the common cancers. The biggest problem is the lung cancer in earlier stage can be asymptomatic, most of patients discovered it in the third or the fourth stage and few of them discovered it at the first stage[4]. In the united states the lung cancer came in the third place between all cancers type in the estimated numbers in 2021 for the people with cancer with 235,760 cases and in the first place of the cancer deaths with 131,880 deaths[5].

DOI: 10.5121/sipij.2022.13601

Using traditional methods in diagnosing, the radiologist will try to detect the nodules in many images. Sometimes one case may need more than one hundred CTs images check to get an accurate diagnosis[6]. In addition to the low efficiency in this method, it's make the doctors tired and affects the diagnosing, so they may misdetection or even not detecting nodules. Trying to avoid those challenges as much as possible, a lot of researchers tried developing CAD systems that can help doctors in diagnosing. it can do works of hours and hours from doctor's time in moments, saving time and energy of the doctors.

2. LITERATURE REVIEW

A lot of computer aided detection systems are following two stages model: first the frames extraction and second try to improving the system by reducing false positive rate[7]. however conventional systems have two clearly weaknesses, first the entire efficiency of the system is low. second is the difference between assumption and the reality, that will lead to deteriorated the overall detection result [8].

Both classification and detection were successfully performed by deep learning because the large data that can be used and the amount of features that can be extracted [9]. Using deep learning in lung nodules detection is now a a hotspot in the area of research[10]. There are many deep learning models and One of the most common deep learning models is CNN the convolutional neural network and it's well suited for classification of images [11].

Using CT scans Zhang et al.[12] performed a CAD system for the detection of lung nodules in 2018. The system divided into 5 stages, first the lung segmentation and then the nodule candidate extraction and third the feature extraction then came the preliminary screening and in the end the reduction of the fails positive. The Average sensitivity and accuracy for the system are 89.3 and 93.6%, respectively.

Also, Li et al.[13] in 2018 they proposed nodule detection system based on CNN . they designed 3 deafferents CNNs, each one has a different input size (12x12,32x32 and 60x60). 154 cases from the JSRT database were used. The highest sensitivity scored was 94% with an average of false positives rete of (5.0).

Raunak Dey et al. in 2018 investigated 3D network to classify lung nodules in the CT image if it's benign or malignant[14]. They said working on 3D volumes yields gives an accurate result than working with 2D slices and a multi-views approximated 3D images. They used 2 datasets in their work, first is called LIDC-IDRI dataset[15] and the other is their own collected dataset. They tested 4 types of 3D networks (Basic CNN, Multi-Output, DenseNet and MoDenseNet). The higher results they got was using the MoDenseNet in both datasets, for the LIDC-IDRI ACC is 90.4% and for their dataset it is 86.84%.

Another automatic detection system for lungs nodules Naqi et al. used a multistage segmentation[16], using the LIDC-IDRI dataset[15]. The identification of lung areas was the first step in segmentation. Then with using morphological processes, edge detection, and a bounding box combiation vessels and other undesired items were removed. Then, for feature extraction, a geometric texture features descriptor (GTFD) has been applied, then an SVM-ensemble classifier is used. Their work get accuracy of 99%.

Earlier system was performed by Akram et al. for detecting and classifying lung nodules[17]. They segmented lung volume using thresholding and hole-filling morphological operators and then using multiple thresholding and pruning to extract candidate nodules. after that they extract features (2D and 3D) and use SVM classifier to classify and evaluate 47 scans taken from LIDC dataset. their work achieved an accuracy of 84.90%, 87.65% sensitivity and 82.15% specificity

Table 1. previous works comparison

#	Authors	Year	Database	Methods	Results
1	Zhang et al.[12]	2018	LIDCIDRI	- segmentation -Nodule locating -Feature extraction - prefatory screening -Fails positive reduction	Sensitivity 89.3%
2	Li et al.[13]	2018	JSRT	-3 CNNs with 3 input size -False positives reduction	Sensitivity 94% - 84%
3	Raunak Dey et al. [14]	2018	LIDCIDRI	-Basic CNN -Multi-Output-DenseNet -MoDenseNet	Accuracy 90.4% - 86.84%
4	Naqi et al. [16]	2018	LIDCIDRI	-Lung region identification -Filtering-geometric texture features descriptor -SVM-ensemble classifier	Sensitivity 98.6% Specificity 98.2%
5	Akram et al. [17]	2015	LIDC	-Thresholding & holefilling morphological operators -extract candidate nodules -2D&3D feature extraction -SVM classifier	accurecy 84.9%

3. METHODOLOGY

The images were collected from a database called “The standard digital image database with and without chest lung nodules” [18]. The database contains 247 x-rays, 154 of which have nodules and 93 of which do not. The images came in high resolution High resolution, they are 2048 x 2048 in matrix size and 0.175mm in pixel size and Useful for the CAD training purposes. In addition, they gave all information that may be needed in the process like the patient age and gender – mass diagnosis is it malignant or benign and its coordinates -degree of subtlety (how easy is to visually detect the nodule). We used the images as they came from the source directly, no segmentation or filtration were applied in the process. Then 120 images were chosen for the training and 60 for testing (105 with nodule & 75 without) we start adjusting the number of features to be extracted many times until we got the best accuracy. We used statistical features and they are: mode, matrix’s mean, variance, standard deviation and median. And same features for the derivatives. Other features were extracted are the local binary pattern features (LBP), they can detect in gray-scale images the uniform local binary pattern of textures [19]. After feature extraction we came to the classifiers, we used 2 classifiers in our work SVM and kNN. We also started adjusting the classifiers to get a better.

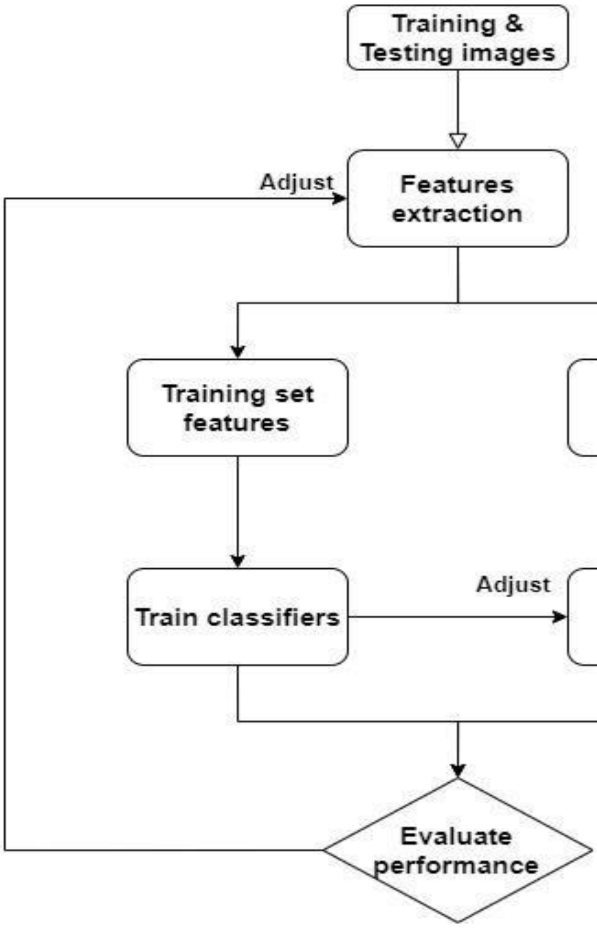

Figure 1. The system's Flowchart

3.1. Pre-Processing

The images obtained from the database have a grayscale appearance. The black and white of the images were inverted when they were imported into MATLAB. In order to retrieve the colors into their original form we used (imcomplement) function in MATLAB.

3.2. Used features

3.2.1. Statistical feature

3.2.1.1. Mean

The average of a group of numbers is the mean $\boldsymbol{\mu}$. It may be calculated by summing the numbers and dividing by the total number of observations.

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$

3.2.1.2. Standard Deviation

The standard deviation $\boldsymbol{\sigma}$ is a statistical metric that quantifies the spread of values.

$$\boldsymbol{\sigma} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{m}-\mathbf{1}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}(\boldsymbol{x_i}-\boldsymbol{\mu})^{\mathbf{2}}}$$

3.2.1.3. Variance

The variance S is a statistical tool used to evaluate variation. It computes the extent to which the dataset is dispersed

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}$$

3.2.1.4. Median

The median is the value that divides a dataset into its upper and lower half. It was utilised to determine the highest frequency value.

3.2.1.5. Mode

The mode is the value with the highest frequency in a data set. It was utilized to determine the most probable value.

In addition, three statistical features were extracted, they were the mean of derivatives, the standard deviation of derivatives, the mode and median of derivatives. All of these features were calculated using MATLAB's tools.

3.2.2. Local Binary Patterns

The local binary patterns (LBP) features recognise the homogeneous local binary pattern of textures in grayscale photos [38]. Utilizing rotationally invariant feature information, the histogram vector has a size of 10 because the size of the neighbours equals 8. Eight neighbouring pixels per pixel were utilised in the calculation.

3.3. The Classifiers

The CADe system concludes with classification. The collected features from the training set will be utilised as classifier inputs. The outcome of each image will then be provided, indicating whether the image contains or not nodules, in order to train the classifiers. Each classifier's output will be examined individually. When each classifier works successfully, the test dataset will be utilised to evaluate the performance of the CADe system. In this work, the following classifiers were employed: Support Vector Machine (SVM), k-Nearest Neighbour (KNN).

SVM was created by V. Vapnik's team (AT&T Bell Labs), a commonly used supervised classifier. SVM can be used to train classifiers that are based on different functions such as linear, polynomial, radial basis and neural network function[20]. The data is represented in space by

SVM, which assigns binary classification to the classified data. In this investigation, our system gives no-nodule images a value of 0 and images with nodule a value of 1. To improve the performance of the classifier, several kernels might be utilized.

kNN is a pattern recognition approach that is commonly used to classify data. It is primarily determined by adjacent sample observation. In the training set, The class of the class given by the majority of its k nearest neighbors will be used to classify the unclassified sample[21].

4. RESULTS

Before presenting our results, you have to know the following parameters:

- TP is abbreviation of True Positive.
- FP is abbreviation of False positive
- TN is abbreviation of True negative
- FN is abbreviation of False negative

The Accuracy (ACC) an expression for the correct predictions

$$ACC = (TP+TN)/(TP+TN+FP+FN)$$

The Sensitivity (Sens) is an expression of the correctly positive cases were diagnosed

$$Sens = TP/(TP+FN)$$

Specificity (Spec) s an expression of the correctly negative cases were diagnosed

$$Spec = TN/(TN+FP)$$

AUC is an expression to give a general look about the system performance how good is it

Table 2. Top three performances in percentage % of the classifiers

Classifier	Training\testing	Sens	Spec	Acc	AUC
KNN - k=3	Training	77	82	80	79
	Test	65	73	70	68.5
SVM - RBF	Training	87	86.5	86.7	85.7
	Test	68.4	70	70	67.4
KNN - k=1	Training	100	100	100	100
	Test	100	81.4	86.7	84

5. DISCUSSION

This paper proposed an automated system for detection of lungs nodules using x-ray images of the chest. As we said before, using the traditional methods for diagnosing the radiologist may need to check a lot of images for one case which mean more time and efforts. Using the CAD systems in diagnosis will aid in the process, it takes no time to check a set of images and may detect nodules that's didn't been detected by radiologists.

Most of the studies in this field were using CT images of the lungs, in same time filters and segmentation were used in the process and these studies achieved a high accuracy system and a lot of them score more than 90%. The proposed system focused on using raw x-ray images with their whole size 2048x2048 directly and achieved an accuracy of 86.7%. we believe that the system can score more accuracy with some processing on the images like cropping small size tissues images or segmenting the lungs area or filtering unwanted objects like thoracic cage Bones but that will be out of our goal. To be honest we can't consider this system as a final diagnosis decider, it always will be the doctors and radiologists especially that's from a patient side I'll not be receptive when i know my health situation was decided by a machine. So we considering this system as a doctors and radiologists assistance.

6. CONCLUSION

Lung cancer is one of the most frequent malignancies, and it is in the first-place cause of deaths by cancers in the world. The major issue is that lung cancer can be asymptomatic in its early stages, most patients find out about it in the third or fourth stages, and just a few find out about it in the first stage.

In this paper, we have proposed a CAD system that may help in diagnosing lung nodule. The system proposed using MATLAB R2022a. As shown in the results, we reached a system with 86.7% accuracy. We believe that with more work we and you can improve the performance of the system. We also recommend using other dataset in addition to or without the dataset we have used (JSRT), since the dataset is limited to 154 of images with nodule.

LIST OF ABBREVIATIONS

Table 3. List of abbreviations

Abbreviation	Stand for	Abbreviation	Stand for
LBP	local binary pattern	CAD	computer aided detection
CT	computerized tomography	kNN	k-Nearest Neighbours
ROC	receiver operating characteristic	JSRT	Japanese Society of Radiological Technology
SVM	support vector machine	AUC	Area Under Curve

REFERENCES

[1] M. D. Eric J. Olson. (2020). Can lung nodules be cancerous? Available: https://www.mayoclinic.org/diseases-conditions/lung-cancer/expert-answers/lung-nodules/faq20058445

[2] C. R. T. MaryAnn De Pietro, "What to know about lung nodules," 2019 .

[3] The American Cancer Society, "Lung Nodules" .

[4] K. D. Miller et al., "Cancer treatment and survivorship statistics, 2019," CA: a cancer journal for clinicians, vol. 69, no. 5, pp. 363-385, 2019 .

[5] A. C. society. (2021). American Cancer Society | Cancer Facts & Statistics. Available: http://cancerstatisticscenter.cancer.org/

[6] D. Gu, G. Liu, and Z. Xue, "On the performance of lung nodule detection, segmentation and classification," Computerized Medical Imaging and Graphics, vol. 89, p. 101886, 2021/04/01 .2021/

[7] D. W. De Boo, M. Prokop, M. Uffmann, B. van Ginneken, and C. M. Schaefer-Prokop, "Computeraided detection (CAD) of lung nodules and small tumours on chest radiographs," European journal of radiology, vol. 72, no. 2, pp. 218-225, 2009 .

[8] J .Ding, A. Li, Z. Hu, and L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2017, pp. 559-567: Springer .

[9] X. Chen, J. Li, Y. Zhang, Y. Lu, and S. Liu, "Automatic feature extraction in X-ray image based on deep learning approach for determination of bone age," Future Generation Computer Systems, vol. 110, pp. 795-801, 2020 .

[10] H. Jiang, H. Ma, W .Qian, M. Gao, and Y. Li, "An automatic detection system of lung nodule based on multigroup patch-based deep learning network," IEEE journal of biomedical and health informatics, vol. 22, no. 4, pp. 1227-1237, 2017 .

[11] A. Krizhevsky, I. Sutskever, and G .E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, pp. 1097-1105, 2012

[12] W. Zhang, X. Wang, X. Li, and J. Chen, "3D skeletonization feature based computer-aided detection system for pulmonary nodules in CT datasets," Computers in biology and medicine, vol. 92, pp. 64-72, 2018 .

[13] C. Li, G. Zhu, X. Wu, and Y. Wang, "False-positive reduction on lung nodules detection in chest radiographs by ensemble of convolutional neural networks," IEEE Access, vol. 6, pp. 16060-16067, 2018 .

[14] R. Dey, Z. Lu, and Y. Hong, "Diagnostic classification of lung nodules using 3D neural networks," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp.778-774

[15] S. G. Armato III et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," Medical physics, vol. 38, no. 2, pp. 915-931, 2011 .

[16] S. M. Naqi, M. Sharif, and M. Yasmin, "Multistage segmentation model and SVM-ensemble for precise lung nodule detection," International Journal of Computer Assisted Radiology and Surgery, vol. 13, no. 7, pp. 1083-1095, 2018/07/01 2018 .

[17] S. Akram, M. Y. Javed ,A. Hussain, F. Riaz, and M. Usman Akram, "Intensity-based statistical features for classification of lungs CT scan nodules using artificial intelligence techniques," Journal of Experimental & Theoretical Artificial Intelligence, vol. 27, no. 6, pp. 737-7.2015 02/ 11/2015 ,51

[18] J. Shiraishi et al., "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," American Journal ofRoentgenology, vol. 174, no. 1, pp. 71-74, 2000 .

[19] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 7, pp. 971-987, 2002 .

[20] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in Proceedings of IEEE computer society conference on computer vision and pattern recognition, 1997 ,pp. 130-136: IEEE .

[21] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," in Classic works of the Dempster-Shafer theory of belief functions: Springer, 2008, pp. 737-760 .

AUTHORS

Eng. Hussein Hamdan M.Sc. Graduate Student of Biomedical Engineering at the Dept. of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, Saudi Arabia. Bachelor’s degree in Electrical Engineering from Al Jouf University, Sakaka, Saudi Arabia.

Dr. Umar S. Alqasemi, Associate Professor of Biomedical Engineering at the Dept. of Electrical and Computer Engineering, King, Abdulaziz University, Jeddah 21589, Saudi Arabia. PhD and MSc degree in Biomedical Engineering from UConn, Storrs, USA. Research work in ultrasound, optical, and photoacoustic imaging, medical imaging recognition, bioelectronics, and digital and analog signal and image processing