Original
ISSN 2413-0516

Decision support detection system for lung nodule abnormalities based on machine learning algorithms

Muna Alsallal,[a*] Mhd Saeed Sharif,[b] Bydaa Hadi,[a] and Ruwaida Albadry[a]

[a]Samawa Technical Institute, ALfurat Alawsat Technical University, Iraq.
[b]School of Architecture, Computing and Engineering, University of East London (UEL) London, UK
[*]Correspondence to Muna Alsallal (E-mail: mona27793@gmail.com)

(Submitted: 14 October 2018 – Revised version received: 28 November 2018 – Accepted: 19 January 2019 – Published online: 26 June 2019)

Objective Investigates the possibility of the early detection in the case of lung infection. Most cases of lung cancer are detected in advance stages as this type is hard to be detected in premature phases. The Zero-Change dataset was chosen to measure the systems' performance on nodule growth. The chosen dataset is assumed as a proven clinical dataset and was used by several researchers in their proposed systems. The designed detection technique has been considered to be used as a decision support tool. This technique is based on using two machine learning algorithms for classification purposes.
Methods Machine learning techniques was applied to detect interesting patterns and manipulate the dataset images in order to enhance the classification task. Pre-processing procedures also have been applied using different MATLAB functions. In addition two well-known techniques that related to the support vector machines (SVMs); the radial basis function kernel-based SVMs and the polynomial kernel-based SVMs have also applied using MATLAB© package named PRTools.
Results The performance of this paper proposed technique was evaluated based on several values of both chosen techniques. The procedure was implemented on the basis of leave-one-out-cross-validation procedure in order to generate unbiased outcomes. The results of cross-validation procedure is averaged and presented as a classifier outcome. The misclassification error, sensitivity, specificity and accuracy are calculated to show a clear image about the two classifiers.
Conclusion The experimental results have shown that the proposed system has scored high accuracy by polynomial kernel SVM. A set of distinguishable representative features are correlated together by a statistics association. Also, this designed system can be considered as a benchmark for developing of other tissues abnormalities signs detection systems.
Keywords machine learning, lung cancer, lung nodule, detection system, classification, image processing, SVM

Introduction

Lung nodules growth which is so-called a "abnormal pulmonary nodules" represents an important factor for lung cancer.[1,2] It is considered to be the foremost reason of human cancerous deaths worldwide among the two genders.[1] The pulmonary nodules has taken an oval shaped growth in the infected lung which sometimes called lung's spots. Patients with cancerous lung are mostly were detected at advanced stage due to difficulties of detection at an early phases.[1] The most important reason for late detection of this disease is inefficiency of X-rays technique in detecting such cases.[1] The only efficient method is the using of the CT-scan device. Another reason for detection difficulties is the need for extensive expertise from the radiologists to be able in categorizing the lung nodules as normal or abnormal. From the radiologists point-of-view the size of normal lung nodules naturally ranges from 2 to 5 mm in thickness and from 1.5 to 3 cm in diameter. However, if the growth in thickness or in area is larger, then the case needs to be considered as critical. It is also highlighted that the radiologists time is limited when compared with the number of patients that they have to see on daily basis. So designing a technique that can recognize the abnormal nodules at an early stage is important and can be assumed as a proactive step to prevent the case aggravation.

Nowadays, machine learning is being a well-known method for improving robust automatic techniques to analyze wide-range of biomedical data.[3] Sajda in 2006 in his paper has reviewed several state-of-the-art of machine learning techniques that have demonstrated their effectiveness in many tasks such as disease detection, diagnosis and treatment monitoring. The review also defined the expansions in machine learning techniques, concentrating on the main two types of learning techniques; supervised and unsupervised. It also went through linear and non-linear approaches and show advantages and disadvantages of each. Such systems were called computer aided diagnosis (CAD) systems, which are widely used in healthcare research area and mostly based on machine-learning algorithms. The standard detection techniques relied on a reference datasets that can work as a foundation for developing robust systems. The Zero-Change dataset was chosen to measure the systems' performance on nodule growth. The chosen dataset was used by Krishnan et al.[2] in their open source system for detecting the progress in lesion sizing. They have evaluated their proposed system on a proven clinical dataset (Zero-Change images dataset). This dataset comprises of 12 pairs of images, each pair contains one for the whole lung and the other is for just the region around the nodules.

One more thing that needs to be noticed is the factors that emphasizes the efficiency of classification tasks in such designed systems. This efficiency can be influenced by two factors; chosen techniques and the set of features. Current scholars highlighted that machine-learning techniques demonstrated their ability to enhance the performance of detection systems.[3] Bengio et al.[4] stated that learning procedures can significantly enhance by combining several types of algorithms such as linear and non-linear approaches in order to gain featured deceptions of data. Another paper that was written by Xu et al.[5] was highlighted the usefulness of using non-linear machine-learning approaches for such systems and high-weight the value of the feature extraction phase.

This research has proposed an automatic detection technique relied on non-linear machine-learning algorithms to classify if the nodule size is normal or abnormal as a second opinion to support the radiologist decision. The system was based on four main techniques; dataset acquisition, pre-processing technique, image features extraction and finally applying a machine-learning technique for classification process. The proposed system trained the nodules sizes after extracting the required features. The next section explores the related work that was done in detecting nodules abnormalities using machine-learning approaches. The following section clarifies the experimental design which includes the system phases. Then, the results will be discussed in Section 4. Finally, we conclude this paper content, method and outcomes in conclusion section.

Background

Many studies were conducted using several machine-learning approaches in order to detect lung nodules abnormality. Bellotti et al.[6] has used contour-based model to detect nodules abnormalities. They scored high performance by gaining 88.5% for detection accuracy. Another group of researchers Riccardi et al.[7] has proposed a new system using 3D radial transforms to detect nodules with overall 71% of detection accuracy. Group of researchers used three different algorithms to determine the pulmonary nodules in CT scan as described in Camarlinghi et al.[8] An advanced technique that based on feed forward neural networks used by Abdulla and Shaharum[9] has been implemented by X-ray images. Specific features set proposed for their study included the area, perimeter and shape. A study that was conducted by Kuruvilla and Gunavathi[10] proposed six discrete features, three of them was mentioned in Abdullah and Shaharum.[9] They add skewness in addition to the time of extraction features from segmented slices that contained two lungs. All those separated features were trained by non-linear machine-learning algorithm, and reported good outcomes.

From the other side of the problem, Support Vector Machines algorithms and their non-linear derivatives were widely used in detecting abnormalities in medical images. The support vector machines (SVMs) were used to detect colon abnormality as a classifier[11,12] with encouraging outcomes. The same SVM applications also implemented for ovarian abnormalities which has revealed good results.[13] Furey et al.[13] used the SVMs technique to several kinds of abnormality data which related to different human body parts such as blood, colon and more. The application of SVM scored high accuracy percentage in most classification cases. Papers written by Segal et al.[14,15] applied the SVM for classification purposes to separate two types of cells with different characteristics. They reported in their results high classification accuracy and has given new indicators to be noted by researchers. A study by Statnikov et al.[16] worked in assessing several machine learning algorithms for their classification performance based on wide range of gene expression to detect abnormality, recommended SV, recommended SVMs as a promising approach in this field. So SVMs has two vital characteristics that make them superior to their peers.

Support Vector Machine

Support vector machines are a group of correlated supervised learning algorithms used commonly for classification and regression. They are considered among the most recent, sophisticated, and high-performance algorithms in artificial intelligence. They aim to separate high-dimensional data in "hyperspace" ("space" with a dimensionality equal to the number of features derived from the training set) using a hyperplane.[17] SVM can be defined as a linear model and it always looks for a hyperplane to separate one class from another.[18] SVMs technique is optimized by connecting with what so-called kernels. Kernels works on a notion that change the depiction of the dot-product in the linear formulation to be non-linear. To illustrate: Vapnik[18] stated that SVM based on dot-products for limited dimension which can be defined as Equation (1)

$$f(y) = x^T y \tag{1}$$

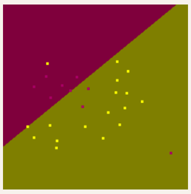
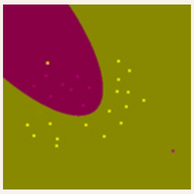
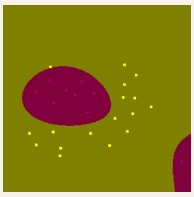
where y represents the outcome of deploying the procedure on non-linear conversion to the proposed data, for example, $yi = \varphi(xi)$, classification is performed by taking the sign of $f(y)$. The notion of the non-linear transmute is mapping the data into a high-dimensional space, where the transmuted data is divided by a hyperplane and linearly separated. To optimize this separable process, the conversion of the data is completed by a kernel trick.[19] The kernels is used to enhance the separation process by replacing the dot-products by non-linear kernel functions. In high level language we can briefly describe the main three types of SVM kernels as shown in Table 1

Experimental Design

This paper proposed an automatic detection system consisted of four main components. Dataset acquisition, features engineering technique which consists of image features extraction and classification component. The four components are shown in Fig. 1.

The same SVM applications also implemented for ovarian abnormalities which has revealed good results.[13] Furey et al.[13]

Table. 1 **Briefly reviewed three types of kernels**

Kernel time	Kernel visualisation	Kernel description
Linear kernel		There is just a "normal" dot-product, thus in 2D decision boundary is always line. Separation of most of points correctly, but due to the "stiffness" of the hypothesis not all points can be captured.[19]
Polynomial Kernel		The polynomial kernel induces space of polynomial combinations of the features, up to certain degree. Consequently we can work with slightly "out of straight line" decision boundaries, such as parabolas.[18]
RBF Kernel		RBF Kernel induced space is Gaussian distributions space ... each point becomes a probability density function (up to scaling) of a normal distribution. In such space, dot-products are integrals and consequently the flexibility is very high.[18]

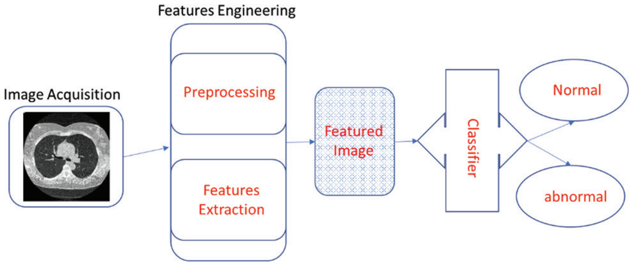


Fig. 1 **This represents a block diagram of the proposed detection system for lungs nodules abnormality classifier[11,12] with encouraging outcomes.**

used the SVMs technique to several kinds of abnormality data which related to different human body parts such as blood, colon and more. The application of SVM scored high accuracy percentage in most classification cases. Papers written by Segal et al.[14,15] applied the SVM for classification purposes to separate two types of cells with different characteristics. They reported in their results high classification accuracy and has given new indicators to be noted by researchers. A study by Statnikov et al.[16] worked in assessing several machine learning algorithms for their classification performance based on wide range of gene expression to detect abnormality, recommended SV, two SVMs algorithms were used to perform classification process; radial basis function (RBF) and polynomial function kernels, respectively. These two techniques are trained using zero-change dataset to find an optimal mode to classify images into their corresponding classes. Then, during the classification phase, the images are classified of whether normal or abnormal. The overall process of dime-sized damage progress detection will be described in details in the following sub-sections.

For the purpose of this paper proposed system, a lung images data sets from a publicly available database The dataset comprises of twelve scan pairs from the CRPF_Database as the first six scan pairs of images take the similar portion thickness which normally equal to 1.25 mm. While the second six pairs scan set has different portion thickness which is equal to 2.5 mm. To describe the pair images; one of the pair scanned images shows the entire lung. The second one concentrates on the nodule region with the same scan resolution of the first. keeping in mind that the person who entitled to perform the scanning process has not change his location during the two scan processes.

For researchers convenient, the access of the 12 scan pairs are publicly available for use in a single compressed file of DICOM images. It can be downloaded conveniently from https://veet.via.cornell.edu/cgi-bin/datac website with all images accessories. They can be obtained from the CRPF_ Database homepage below the title "Repeat Single Session". Furthermore, the scan images can also be downloaded separately in several versions using the direct function for image database download function. To prepare for implementing the proposed system phases a particular pre-processing procedure is applied. We first applied the shade correction technique, as the unbalanced lightning if the scan image needs to be modified if a specific object has to be properly spotted. The second step of pre-processing procedure is applied using morphological-opening task as a proactive step to assess the background. This task was applied using MATLAB© function which is called "imopen". This function typically is applied for morphological-opening task that can be performed scan with 12 pixels structuring element. After morphological-opening is completed by applying "imtophat" MATLAB© function, the task of increasing the image contrast is performed. The task is completed by implementing "imadjust" MATLAB© function as shown in Figure 2. Then, "bwareaopen" another MATLAB© function is applied to eliminate the noise of the background. As a result the concluding output, which only displays the affected scan image region, is gained.

For this paper purpose seven types of features were determined to be extracted from each image to train the classifier. The set of chosen features that needs to be extracted is as follows: definite number of pixels in nodule region (DNP), Marginal length around the region feature (MLAR) (Perimeter of the entity) which can be clearly shown in Fig. 3 as the

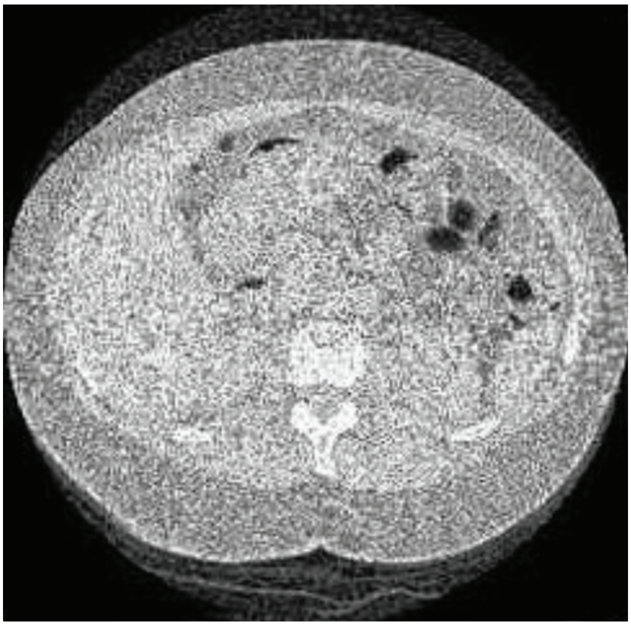


Fig. 2 **This represents the scan lung image after implementing the task of increasing the image contrast.**

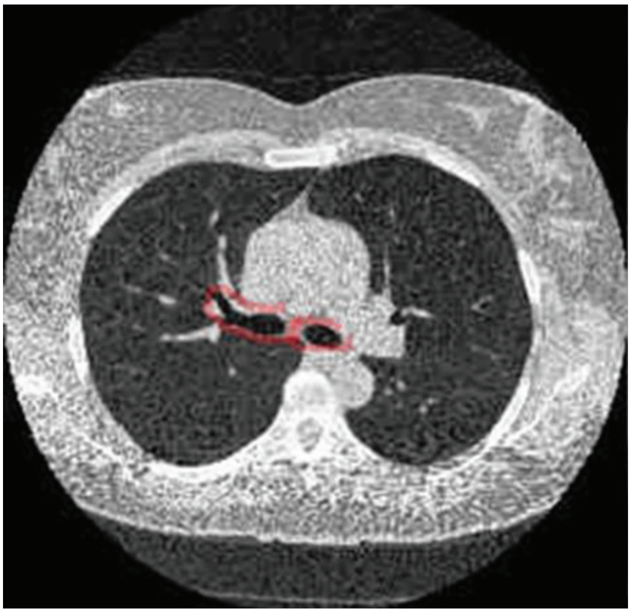


Fig. 3 **This represents the marginal length around the region feature which perimeter of the entity.**

element surrounded by red color, maximum length of the major axis in the region (MaLR), minimum length of the minor axis in the region (MiLR), feature ratio [equal to (MaLR) divided by (MiLR) and finally Region roundness (RR) which is calculated by ((4 * π * DNP)/(MLAR^2))].

The classification procedure is implemented via a MATLAB© package named PRTools.[18] This paper proposed system based on implemented two well-known techniques that related to the SVMs; the RBF kernel-based SVMs and the polynomial kernel-based SVMs.

Results

The performance of this paper proposed technique was evaluated based on several values of both chosen techniques. The procedure was implemented on the basis of leave-one-out-cross-validation procedure to generate unbiased outcomes. The results of cross-validation procedure is averaged and presented as a classifier outcome as shown in Table 2. The misclassification error, sensitivity, specificity and accuracy are calculated to show a clear image about the two classifiers. The experimental results show that the two classifiers are able to identify both classes; however, the polynomial kernel-based SVMs outperformed the RBF kernel-based SVMs.

On the other hand, the system performance were assessed based on radiologists opinions which was compared with system performance as shown in Table 3. The nodules features that characterised by radiologists and the proposed system are premeditated and then compared. The presence of nodules which sized from 1.5 to 3 cm in diameter and thickness that ranges from 2 to 5 mm represented the normal lung nodules.

However, if those measures are characterised in larger forms then they should be considered as an abnormal case which needs more attention and investigation. In order for the system performance to be tested when compared with the radiologist's opinions, a statistical test were applied. *t*-Test was applied to measure the differences in mean between the marked images and the system output. The results are then depicted in Table 3, the value of sigma that was obtained under 95% confidence interval, 0.211. This sigma value demonstrated that the means of the radiologist's opinions and the system output are not pointedly varied.

Discussion

The proposed system was based on the notion of integrating several techniques starting from choosing the dataset, features engineering to transform the original image to a set of proposed features then implementing the classifier. The pre-processing mechanism represents an important proactive step to optimise the classifier performance. The detection of such cases is assumed as a challenging task for most automated detection systems. This challenge relates to the speciality of nodules characteristics with regard to their size, thickness and location. The performance of a lung nodule abnormalities based on machine learning algorithms is measured by calculating sensitivity, specificity and accuracy. After applying the both proposed classifiers on zero-change dataset, the confusion matrix was generated using leave-one-out-cross-validation. Table 2 presented the values of sensitivity, specificity and accuracy. The accuracy value refers to the total number of correct predictions that was made by each classifier. As shown the polynomial kernel SVM outperformed the other algorithm. Sensitivity represents the ability of the system that can detect nodules abnormalities which is matter. While specificity represents the system ability to identify the normality of nodules. In addition, the sigma value that resulted from *t*-test as shown in Table 3 has shown that the differences between the marked images (radiologist opinions) and the system outcomes is not significant.

Table. 2 **Averaged results of two classifiers**

Measure	Polynomial kernel SVM	RBF kernel SVM
Misclassification error	0.0919	0.1826
Accuracy	0.9191	0.7273
Specificity	1	0.3224
Sensitivity	0.881	0.763

Table 3. **The analysis of t-test between the proposed system performance and radiologists opinions**

		Mean	*N*	Std. deviation				
Pair 1	Marked	1.71	12	0.419				
	Testing	1.90	12	0.331				
Correlations of paired samples								
		N	Correlation	Sig.				
Pair 1	Marked and testing	12	−0.019	0.901				
				Paired differences				
		Mean	Std. deviation	Std. error mean	95% Confidenceinterval of the difference		*t*	Sig. (two-tailed)
					Lower	Upper		
Pair 1	Marked and testing	−0.101	0.525	0.081	−0.265	0.069	−1.131	0.211

N = The number of nodules images analysed by the radiologists and the proposed system

Conclusion

This paper produced a detection system for categorising lung nodules as either normal or abnormal. This paper system worked on a notion of integrating several techniques; image acquisition, features engineering to accurately generate a clear depiction of nodules images. The zero-change dataset was used for the purpose of this paper proposed system. The experimental results have shown that the proposed system has scored high accuracy by polynomial kernel SVM. A set of distinguishable representative features which are correlated together by a statistics association. Also, this designed system can be considered as a benchmark for developing of other tissues abnormalities signs detection systems. In terms of developing the proposed system in future, the authors intend to integrate more intelligent techniques for feature extraction to facilitate the detection capabilities. Furthermore deep learning techniques will be involved to compare two generation of machine learning algorithms performance.

Conflicts of interest

None. ■

References

1. Kumar D, Wong A, Clausi DA. Lung nodule classification using deep features in CT images. In: 2015 12th Conference on Computer and Robot Vision (CRV), IEEE, Halifax, NS, Canada, 2015, pp. 133-138.
2. Krishnan K, Ibanez L, Turner WD, Jomier J, Avila RS. An open-source toolkit for the volumetric measurement of CT lung lesions. Opt Express. 2010;18:15256–15266.
3. Sajda P. Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng. 2006;8:537–565.
4. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013; 35:1798–1828.
5. Xu Y, Mo T, Feng Q, Zhong P, Lai M, Chang EI. Deep learning of feature representation with multiple instance learning for medical image analysis. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Florence, Italy, 2014, pp. 1626–1630.
6. Bellotti R, De Carlo F, Gargano G, Tangaro S, Cascio D, Catanzariti E, et al. A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model. Med Phys. 2007; 34:4901–4910.
7. Riccardi A, Petkov TS, Ferri G, Masotti M, Campanini R. Computer-aided detection of lung nodules via 3D fast radial transform, scale space representation, and Zernike MIP classification. Med Phys. 2011;38:1962–1971.
8. Camarlinghi N, Gori I, Retico A, Bellotti R, Bosco P, Cerello P, et al. Combination of computer-aided detection algorithms for automatic lung nodule identification. Int J Comput Assist Radiol Surg. 2012; 7:455–464.
9. Abdulla AA, Shaharum SM. Lung cancer cell classification method using artificial neural network. Inform Eng Lett. 2012;2:49–59.
10. Kuruvilla J, Gunavathi K. Lung cancer classification using neural networks for CT images. Comput Methods Programs Biomed. 2014;113:202–209.
11. Moler EJ, Chow ML, Mian IS. Analysis of molecular profile data using generative and discriminative methods. Physiol Genomics. 2000;4:109–126.
12. Liu Y. Active learning with support vector machine applied to gene expression data for cancer classification. J Chem Inf Comput Sci. 2004;44:1936–1941.
13. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000;16:906–914.
14. Segal NH, Pavlidis P, Antonescu CR, Maki RG, Noble WS, DeSantis D, et al. Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. Am J Pathol. 2003;163:691–700.
15. Segal NH, Pavlidis P, Noble WS, Antonescu CR, Viale A, Wesley UV, et al. Classification of clear-cell sarcoma as a subtype of melanoma by genomic profiling. J Clin Oncol. 2003;21:1775–1781.
16. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005;21:631–643.
17. Alsallal M. A Machine Learning Approach for Plagiarism Detection. PhD diss., Coventry University, 2016.
18. Vapnik V. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, New York, 2013.
19. Aizerman MA, Braverman EM, Rozonoer LI. Theoretical foundations of the potential function method in pattern recognition learning. Automation Remote Contr. 1964;25:821–837.
20. Duin RPW, Juszczak P, Paclik P, Pekalska E, de Ridder D, Tax DMJ, et al. *PRTools4.1. A Matlab Toolbox for Pattern Recognition*, Delft University of Technology.