Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.XX n.XX, p. 69-74. July/August., 2024.
DOI: https://doi.org/10.5935/jetia.v10i48.1202

RESEARCH ARTICLE
ACCESS

OPEN

DETECTION OF LUNG NODULES USING SUPPORT VECTOR MACHINE

***Jhon Anthony Fernández Castro[1], Marlen Pérez Díaz[2] and Rubén Orozco Morales[3]**

[1,2,3] Universidad Central "Marta Abreu de las Villas" – UCLV. Santa Clara-Villa Clara, Cuba.

[1]http://orcid.org/0009-0003-7576-921x, [2]http://orcid.org/0000-0002-3706-9154, [3]http://orcid.org/0000-0002-6240-1569

Email: *jhonanthonyfernandezcastro@gmail.com[1], mperez@uclv.edu.cu[2], rorozco@uclv.edu.cu[3]

ARTICLE INFO

Article History
Received: July 01th, 2024
Revised: July 08th, 2024
Accepted: July 08th, 2024
Published: July 18th, 2024

Keywords:
Chest x-ray,
Pulmonary nodule,
Support vector machines,
Machine learning.

ABSTRACT

Lung cancer is a disease of high mortality worldwide. Therefore, early diagnosis and treatment can save lives. Lung cancer appears as a solitary nodule on chest x-ray, which is sometimes very difficult to detect for the human eye. Therefore, developing a computer-aided diagnosis (CAD) system for the detection of lung nodules, using machine learning (ML) could have a significant impact on patient prognosis. The proposed algorithm begins by pre-processing the images to improve their quality. The lung area is then segmented by thresholding. In the next step, nodule candidates are determined using a sliding band filter and segmented by applying a threshold algorithm, based on adaptive distance (ADT). Next, the suspicious areas are processed by a support vector machine (SVM), based on 15 shape and texture characteristics. Three SVM models were trained and validated with images from a public JSRT database. The best result was obtained with the radial base model (87 % sensitivity). This performance is valued as favorable with respect to human performance.

I. INTRODUCTION

Data from the World Health Organization show that lung cancer is the third most common type of tumor and the leading cause of cancer-related death in the Americas, with more than 324,000 new cases and nearly 262,000 deaths each anus [1]. It encompasses a set of diseases resulting from the malignant growth of cells in the lung tissue [2]. Early detection (when they are a nodule) generally allows for earlier therapeutic intervention and a better prognosis. Pulmonary nodules present radiographically as rounded opacities, no larger than 3 cm [2]. With computed tomography (CT), the detection of the smallest has increased [3]. Although this technique is more sensitive for detecting lung cancer at an early stage, chest x-rays are also used for this task, due to their low cost, simplicity, and low radiation dose. However, the false negative rate is around 30% [4]. Tumor characteristics such as size, visibility and location, radiographic image quality and patient positioning and movement are factors that determine the probability of missing a lung neoplasm. The experience of the radiologist also plays a role in this context [5]. The 20-60% error rate has remained the same for years. To reduce the risk of misdiagnosis, computer-aided diagnosis (CAD) systems are developed [6].

CADs have been presented as a second opinion for the radiologist in the early detection of different diseases [7-9]. The increase in computing capabilities in recent years has allowed the emergence of systems based on machine learning techniques. A revised literature review shows that the sensitivity of the human eye varies between 49% and 65% without the use of CAD, and between 68% and 93% with its assistance for lung nodules [10].

The increase in processing capabilities, data availability and storage of current computers allowed the emergence of ML. These use pre-processing and segmentation stages of the lung region and the use of equations to describe the lesions. The main characteristic is that they use a classifier, which is able to learn to separate the classes: nodule / normal, from a vector of features previously extracted from the images [11].

DL algorithms use convolutional neural networks, which to achieve adequate model training and generalization power, require a very high volume of data, which in medicine is not always available. For this reason, even algorithms that do not use DL continue to be competitive [12].

Support vector machines (SVMs) [13] are among the most successful ML algorithms of the last 20 years. Some reasons that explain this achievement are its good properties for the generalization and convergence of models [14].

SVMs are therefore a classification and regression prediction tool, which uses machine learning (ML) theory to maximize predictive accuracy and automatically avoid overfitting the model to the learning data. They can be defined as systems that use a hypothesis space of linear and nonlinear functions, in a high-dimensional feature space [13].

SVMs use statistical learning theory to find a regularized hypothesis that fits the available data well. The greatest limitation of SVMs lies in the choice of a kernel that suits the problem under analysis. Furthermore, they depend on feeding with a set of characteristics that are appropriate to interpret the problem well.

The objective of the work has been to develop a CAD system based on SVM, which is capable of detecting lung nodules from chest x-rays.

II. MATERIALS AND METHODS

II.1 THEORETICAL DESCRIPTION OF SVM

If we consider that we have a real training data set (R) $(x_1, y_1), \dots, \dots, (x_i, y_i) \in R^n$ where x_i is the ith vector of the sample and y_i its corresponding label, SVM aims at linear discrimination, finding a hyperplane that separates samples of different classes. At the same time, the calculated hyperplane should maximize the distance between the marginal samples of each class. These two conditions are part of an optimization problem. The solution to this problem is based on a small percentage of marginal samples called support vectors. In that case, after computing the optimal solution, the decision function is given by Optimization equation:

$$f(x) = b + \sum \alpha_i y_i K(x_i , y_i) \quad (1)$$

where b denotes a constant bias value, a_i, i =1 … n correspond to multipliers for each sample, having a non-zero value only for the support vectors [15] and $K(x_i, y_i)$ is a kernel.

In the case of nonlinear class separation, the input samples can be mapped to a higher dimensionality space, where they are linearly separated. This is achieved with an appropriate kernel function $K(x_i, y_i)$. A small number of support vectors are selected, which minimizes the computational requirements during testing [15].

II.2 DESCRIPTION OF THE IMAGE SETS USED

An annotated radiograph database, JSRT [16], was used in this investigation. They are images with a spatial resolution of 1024 x 1024 pixels. Images with nodules have the location of the lesion and each one presents a single lesion. 93 normal images and 154 images with nodules were used, for a total of 247 images.

II.3 SOFTWARE AND HARDWARE USED

The software used was MATLAB R2018b (9.5.0.944444). A desktop computer with the following features was used: Processor (CPU) Intel® Core™ i3-6100U (2.3 GHz, 3 MB cache, 2 cores), Graphics Card: Intel® HD Graphics 520 (Integrated), RAM: 8 GB DDR4-2133 SDRAM (2*4GB).

II.4 CAD SYSTEM, 1ST STAGE: IMAGE PREPROCESSING

At this stage, two pre-processing were carried out, one aimed at improving visibility in the image and another necessary for the segmentation of the lung region. For the first one, a smoothing of the image was carried out. Noise was reduced by applying a 5x5 pixel averaging convolution filter. Then, using local normalization (LN) filtering, global contrast equalization was achieved across the image [17], and finally, edges were enhanced with filtering homomorphic. In this case, a logarithmic mapping was carried out in the space domain, to separate the illumination and reflectance components. To make the image illumination more uniform, the high frequency components (reflectance) were increased and the low frequency components (illuminance) were decreased. High-pass filtering was used in this case to suppress low frequencies and amplify high frequencies. Figure 1 shows the result of this stage of the System.

Figure 1: Original image (left). Preprocessed image (right). Source: Authors, (2024).

Lung region segmentation was performed during the second preprocessing to limit nodule detection to the segmented lung fields and prevent false positive (FP) detections. For this, the multi-level threshold method was used, as performed in [17], since the present work is a continuation of this.

An initial segmentation is first obtained using the multi-level threshold method. Then, because in some cases this method leaves out parts of the lung region, an auxiliary segmentation is performed, using another variant of the multilevel threshold method, with the objective of identifying the outer edges of the lungs to correct the initial segmentation. Both images are added and morphological operations of opening and closing the image are applied [18] with which the final segmentation is obtained. In this case, a structuring element in the shape of a disk with a radius of 65 pixels was used to obtain the initial segmentation. Aperture removed all pixels in regions that are smaller than the structuring element, smoothed out external bumps, broke up narrow sections, and removed thin bumps. The closure filled in the smaller holes and concavities. Subsequently, the multi-level threshold method was applied again to the processed image, thus obtaining auxiliary segmentation. Auxiliary segmentation precisely defines the borders of the lungs. This was done with

adaptive morphology, using an edge-linking algorithm, as in [17]. Figure 2 shows the result of this step.

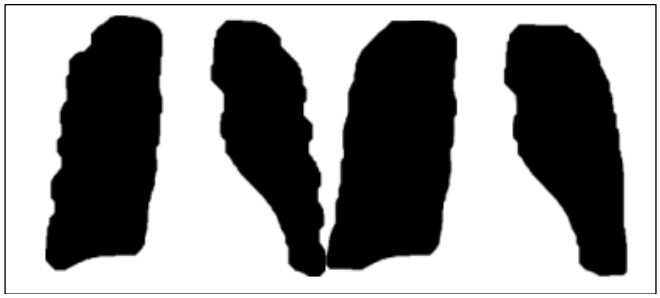

Figure 2: Initial segmentation masks (left) and corrected segmentation (right).
Source: Authors, (2024).

II.5 CAD SYSTEM, STAGE 2: DETECTION AND SEGMENTATION OF CANDIDATE NODULES

After pre-processing, an algorithm was applied to detect candidate nodules. This means locating regions of the image that may be potential lung nodules. For this, a local convergence filter (LCF) was used as in [19],[20], with the difference that the convergence filter used in this case was a sliding band filter [20]. With this filter you can perform detection of nodules at various scales and at the same time introduce the background into the detection approach. Although lung nodules generally have a circular shape, this is not always the case; they can also be spiculated, lobulated, poorly defined, or with irregular edges, so a convex shape and a limited range of sizes for the area were assumed thereof.

The LCF evaluates the degree of convergence of gradient vectors within a local area (support region) towards a pixel of interest (central location of the area) [17]. The degree of convergence is related to the distribution of the directions of the gradient vectors and not to their magnitudes. This makes it easy to define a global threshold that is not affected by image illumination. The local convergence of a gradient vector at a given pixel is defined as: the cosine of its orientation with respect to the line connecting that pixel and the central pixel of the area.

After candidate detection, an adaptive distance-based thresholding (ADT) algorithm was applied to segment each candidate nodule [17-20]. These segmentations were used as a key part of the calculation of most of the characteristics, to identify the real nodule among all the candidates. Figure 3 shows the result of these steps.

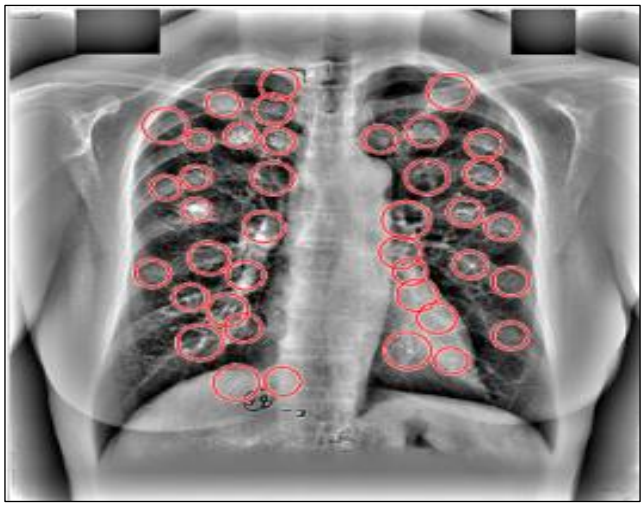

Figure 3: Detection and segmentation of candidate nodules.
Source: Authors, (2024).

II.6 CAD SYSTEM, STAGE 3: FEATURE EXTRACTION

For each nodule candidate, a set of features were calculated and presented in Table 1. Geometric features were calculated using only the shape and position information provided by adaptive segmentation. Intensity features were calculated for the preprocessed images, using the segmentation mask to define the candidate boundary.

Table 1: Characteristics of possible nodules.

Characeristic	**Content**
Lesion size	Maximun diameter after segmentation
Area	Surface
Perimeter	Contour Long
Circularity 1	$A_1=\pi r_m^2$ r_M is the maximun radius (2)
Circularity 2	$A_2=\pi r^2$ r_m is the minimun radius (3)
Excentricity	Deviation respect a circunference.
Internal and external mean	Average of pixel intensity in and out ADT segmentation.
Mean separation	$(Mediainterna - Mediaexterna)$ $/(Mediainterna + Mediaexterna)$ (4)
Internal and external standar deviation	Dispersion respect to the mean value in and out the ADT segmentation
Contrast 1	$(Mediainterna - Mediaexterna)$ $/Mediainterna + Mediaexterna$ (5)
Contrast 2	$Contraste1 / (Des.estandar_{interna} + Des.$ $estandarexterna)$ (6)

Source: Authors, (2024).

II.7 CAD SYSTEM, STAGE 4: CLASSIFICATION OF POSSIBLE LUNG NODULES

The objective of any classifier is to find a boundary that allows the classes to be separated (in this case nodule / non-nodule). Classification was performed by a binary SVM classifier. The classifier receives a set of training features, a result of the previous stage, each labeled for the classes: nodule and non-nodule. From these characteristics, a model was built that was subsequently used to classify the test images. Features from 133 nodule and non-nodule images were used for model training.

Three kernels were implemented to train the model, so that a non-linear decision surface can be transformed into a linear equation and into a greater number of dimension spaces. Mathematical formulation of the kernels used:

Linear kernel

$$K(x,y) = \langle x,y \rangle \quad (7)$$

Gaussian or radial base kernel

$$K(x,y)=e^{-\gamma|x-y|2}, \gamma>0 \quad (8)$$

Polynomial kernel of degree P

$$K(x,y)=[\gamma\langle x,y\rangle+\tau]^p \quad (9)$$

To obtain each SVM model, the MatLab function *fitcsvm* (X, Y) was used, using the feature vectors X and the classification labels in the vector Y. Figure 4 shows the scheme of the nodule classification stage.

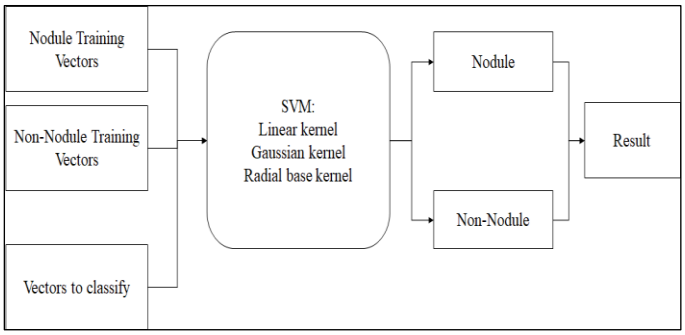

Figure 4: Scheme for the classification of possible nodules.
Source: Authors, (2024).

At the end of the training of each model, with the different nuclei, validation was carried out with 15 images of nodules and non-nodules (not used for training). For this, sensitivity, precision, accuracy, balanced accuracy and the F1-*score* value were used as metrics.

IV. RESULTS AND DISCUSSIONS

IV.1 FEATURE VECTORS

Although no statistical analysis was done to prove the existence of significant differences between the characteristics for each class, based on the centroids per class of each one, it can be seen that the characteristics that differ the most between the two classes are: the area of the nodule, eccentricity, external standard deviation and circularity 2. However, this aspect must be scientifically corroborated in future work. For the purposes of this work, all characteristics were used in the subsequent stages, not just those that a priori appear to be the determining ones. Figure 5 shows the results of features in each class.

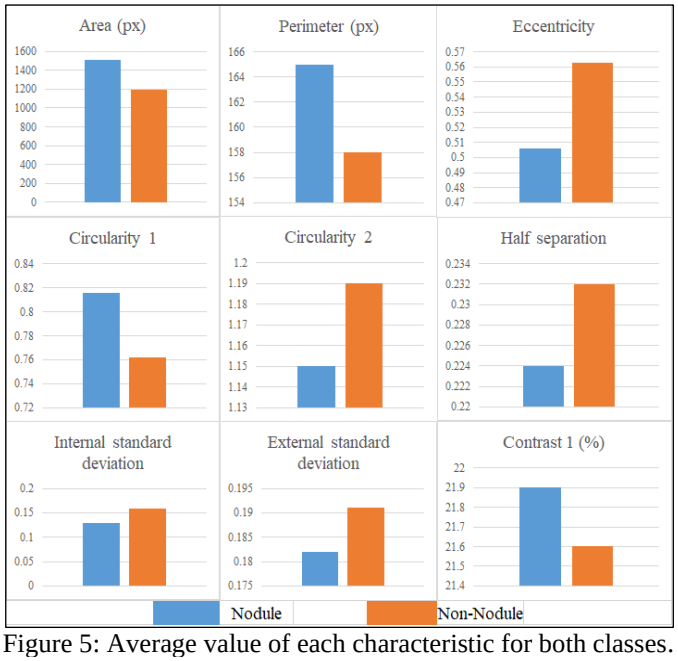

Figure 5: Average value of each characteristic for both classes.
Source: Authors, (2024).

IV.2 MODEL TRAINING AND VALIDATION RESULTS

Figure 6 shows the validation results of the 3 SVM models tested for the vector including all features and Table 2 shows their respective metrics.

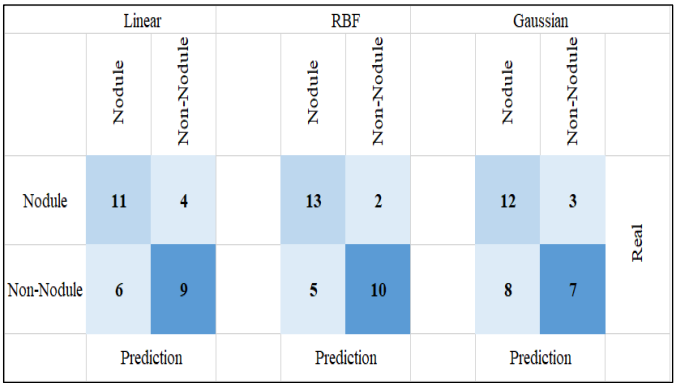

Figure 6: Confusion matrices for validation data.
Source: Authors, (2024).

Table 2: Model performance metrics in validation.

Metrics (%)	Linear	RBF	Gaussian
Accuracy	66,67	76,67	63,33
Sensitivity	73,33	86,67	80,00
Precision	64,71	72,22	60,00
F1 (%)	68,75	78,79	68,57
Balanced accuracy	66,67	76,67	63,33

Source: Authors, (2024).

In the clinical setting, the most valued metric is sensitivity, which provides the percentage of findings of true nodules (VP). The consequences of evaluating a healthy area incorrectly involve simply repeating the test. However, the consequences of incorrectly evaluating a VP are the most unfavorable, since the early diagnosis of the disease can be affected. In the experiment carried out, the best result was that obtained with the radial base model (87% sensitivity). This performance is rated as favorable compared to human performance, which is between 49% and 65% [10].

IV.3 GENERAL DISCUSSION

In the present work, a classification stage has been added to the CAD system designed in [17]. In this previous work, we only reached the phase of detecting candidate nodules. This article takes advantage of the benefits of SVMs to separate the classes into nodule or non-nodule.

From digital radiographic images, radiologists can typically see 45% to 68% of actual nodules. It has been noted that if the nodules are smaller than 10 mm, only 29% are detected. In fact, detection is strongly determined by location [10]. Hence the importance of using automated systems as a second opinion.

For example, Figure 7 shows a nodule that is very difficult to visualize by the human eye, according to the BD JSRT annotation [16], due to its low contrast and its overlap with a rib. Note in the zoom on the right that for this nodule the contrast difference with respect to the surrounding background is practically undetectable by the human eye. The radial-based model, however, was able to detect such a nodule, indicating its potential for the task.

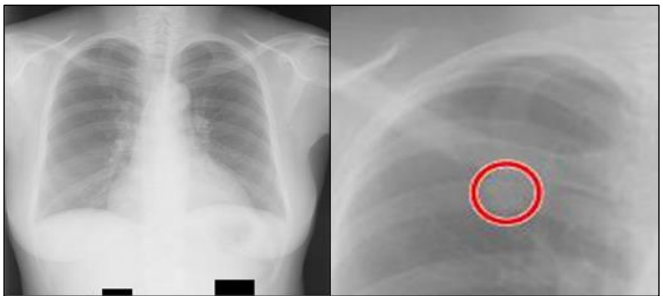

Figure 7: Nodule undetectable to the human eye.
Source: Authors, (2024).

The computational cost of training for the three models was 10 seconds per model, for the computational capabilities used. Once the model is trained, the approximate execution time of all the stages of the CAD system, for an image, is 3-4 min.

When reviewing four studies with lung nodule detection models published in the literature, in which the JSRT database was used, the results that are summarized in Table 3 were found. These reinforce the expressed criterion that the proposed system manages to classify lung nodules with sensitivity values similar to other systems previously published by international authors, at least for the validation training data. Future studies with data from different origins should be carried out to test the generalization power of the system.

Table 3: Results of CAD systems made with the JSRT DB

Studies	Sensitivity (%) (Images with correct detection/Total images with nodules)
Wei [21]	80 (123/154)
Coppini[22]	60 (93/154)
Schilham [23]	67 (103/154)
Hardie [19]	63 (88/140)

Source: Authors, (2024).

As directions for future work, it is proposed to introduce a rigorous method for selecting class-determining characteristics, such as principal component analysis, test other classifiers such as random forest trees or cluster analysis, and carry out a perceptual study, with human observers, under standardized conditions, to evaluate the real detection difference between the proposed system and the performance of human experts on the task.

V. CONCLUSIONS

It has been proven that SVMs have potential for the task of detecting lung nodules from chest x-rays. An SVM with a radial basis model presented potential for the task in terms of computational effectiveness and efficiency. Of the characteristics used as input vector for the SVM, the most sensitive a priori for classifying nodules were: nodule area, eccentricity, external standard deviation and circularity 2.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Marlen Pérez Díaz,
Methodology: Marlen Pérez Díaz, and Rubén Orozco Morales.
Investigation: Jhon Anthony Fernández Castro
Discussion of results: Jhon Anthony Fernández Castro, Marlen Pérez Díaz, and Rubén Orozco Morales,
Original Draft: Jhon Anthony Fernández Castro.
Writing – Review and Editing: Jhon Anthony Fernández Castro, Marlen Pérez Díaz, and Rubén Orozco Morales,
Resources: Marlen Pérez Díaz.
Supervision: Marlen Pérez Díaz, and Rubén Orozco Morales.
Approval of the final text: Jhon Anthony Fernández Castro, Marlen Pérez Díaz, and Rubén Orozco Morales,

VII. ACKNOWLEDGMENTS

This work has been partially funded thanks to Agency for Nuclear Energy and Advanced Technology of Cuba (AENTA), Project Code PS211LH02.

VIII. REFERENCES

[1] W. W. H. Organization. (26/04/2022). *Cáncer*. Available: https://www.who.int/es/news-room/fact-sheets/detail/cancer

[2] H. Mahersia, M. Zaroug, and L. Gabralla, "Lung cancer detection on CT scan images: a review on the analysis techniques," *Lung Cancer*, vol. 4, no. 4, 2015.

[3] J. Jeremy, H. McAdams, and S. Rossi, "Neoplasias pulmonares primarias," Diagnóstico *por Imagen del Tórax*, pp. 933-935, 2011.

[4] Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, "Machine Learning Applications in Lung Cancer Diagnosis, Treatment and Prognosis," *ArXiv*, vol. 2203.02794, 2022.

[5] A. Del Ciello *et al.*, "Missed lung cancer: when, where, and why?," *Diagnostic and Interventional Radiology*, vol. 23, no. 2, p. 118, 2017.

[6] M. Haber, A. Drake, and J. Nightingale, "Is there an advantage to using computer aided detection for the early detection of pulmonary nodules within chest X-Ray imaging?," *Radiography*, vol. 26, no. 3, pp. 170-178, 2020.

[7] M. Souto *et al.*, "Detección automática de nódulos pulmonares en tomografía computarizada. Un estudio preliminar," *Radiología*, vol. 50, no. 5, pp. 387-392, 2008.

[8] J. D. López-Cabrera, L. A. L. Rodríguez, and M. Pérez-Díaz, "Classification of breast cancer from digital mammography using deep learning," *Inteligencia Artificial*, vol. 23, no. 65, pp. 56-66, 2020.

[9] E. D. S. Aday, M. P. Díaz, and R. O. Morales, "Diseño de Sistema Automatizado para Detección de Anomalías en Imágenes Digitales de Mama," *Journal of Health and Medical Sciences*, vol. 5, no. 4, pp. 229-243, 2019.

[10] I. Bush, "Lung nodule detection and classification," *Report, Stanford Computer Science*, 2016.

[11] H. Seo *et al.*, "Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications," *Medical physics*, vol. 47, no. 5, pp. 148-167, 2020.

[12] F. Shan *et al.*, "Lung infection quantification of COVID-19 in CT images with deep learning," *arXiv preprint arXiv:.04655*, 2020.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.

[14] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning* (Springer Texts in Statistics). Springer New York, NY, 2013.

[16] 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 of Roentgenology*, vol. 174, no. 1, pp. 71-74, 2000.

[17] E. Martinez-Machado, M. Perez-Diaz, and R. Orozco-Morales, "Automated System for the Detection of Lung Nodules," *Lecture Notes in Computer Science*, vol. 13055, pp. 337-348, 2021.

[18] Y. Zhang and L. Wu, "Optimal multi-level thresholding based on maximum Tsallis entropy via an artificial bee colony approach," *Entropy*, vol. 13, no. 4, pp. 841-859, 2011.

[19] R. C. Hardie, S. K. Rogers, T. Wilson, and A. Rogers, "Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs," *Medical Image Analysis*, vol. 12, no. 3, pp. 240-258, 2008.

[20] C. Supanta, G. Kemper, and C. del Carpio, "An algorithm for feature extraction and detection of pulmonary nodules in digital radiographic images," in *2018 IEEE International Conference on Automation/XXIII Congress of the Chilean Association of Automatic Control (ICA-ACCA)*, 2018, pp. 1-5: IEEE.

[21] J. Wei, Y. Hagihara, A. Shimizu, and H. Kobatake, "Optimal image feature set for detecting lung nodules on chest X-ray images," Berlin, Heidelberg, 2002, pp. 706-711: Springer Berlin Heidelberg.

[22] G. Coppini, S. Diciotti, M. Falchini, N. Villari, and G. Valli, "Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 4, pp. 344-357, 2003.

[23] A. M. Schilham, B. Van Ginneken, and M. Loog, "A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database," *Medical Image Analysis*, vol. 10, no. 2, pp. 247-258, 2006.