Transactions on Machine Intelligence 1(2) (2018) 99–105

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI

A New Approach to Feature Extraction Based on Lung CT Images Using Machine Learning Algorithms for Lung Disease Classification

M. R. Fazel Najafabadi[1,*], S. Ayat Najafabadi[1], S. Fekri Ershad[1]

[1] Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Isfahan, Iran

ARTICLE INFO

Article History:
Received 10 March 2018
Received in revised form 24 May 2018
Accepted 26 June 2018
Available online 27 June 2018

Keywords:
Human Tissue Density Analysis, Gray Level Co-occurrence Matrix, Lung Disease, Moments, Machine Learning, Feature Extraction, Support Vector Machine, Optimum-Path Forest

ABSTRACT

Accurate diagnosis of lung diseases based on processing and analyzing lung CT images is crucial for aiding medical decision-making. This study presents a new feature extraction method based on human tissue density patterns, called Analysis of Human Tissue Density (AHTD). This method is compared with the Gray Level Co-occurrence Matrix (GLCM), Hu Moments (HM), Statistical Moments (SM), and Zernike Moments (ZM). The dataset of chest tomography images was obtained from the Walter Cantidio University Hospital in Fortaleza, Brazil. Four machine learning classifiers were used in this study: Bayesian Classifier, Optimum-Path Forest (OPF), k-Nearest Neighbors (KNN), and Support Vector Machine (SVM) to classify lung diseases in chest images. Feature extraction from lung images was performed in 5.2 milliseconds, achieving an accuracy of 99.01% for lung disease diagnosis and classification. The results of this study suggest that the proposed method can be used in real-time applications due to its rapid processing time and high accuracy for classifying lung diseases based on lung CT images.

1. INTRODUCTION

Medical diagnosis has always been an art. Throughout history, we remember renowned physicians just as we do famous painters or composers. Who is called an artist? Someone who can do things others cannot. This is precisely what a good doctor does during medical diagnosis. They use their education, experiences, and talent to diagnose a disease [1]. Today, computer-based methods have made significant and effective contributions in various fields, particularly in medical science. Artificial intelligence, by offering various techniques, has also had an effective role in solving different issues. In medical science, these techniques have greatly assisted in the early diagnosis of diseases [2]. For instance, one of the AI techniques is image processing, which, as the name suggests, deals with processing and analyzing images. Image processing consists of two major categories: image enhancement and machine vision. Image enhancement includes methods to improve the visual quality of images and ensure their correct display, while machine vision deals with methods to understand the meaning and content of images [3], using them in tasks such as biomedical engineering.

* Corresponding Author: mohamadreza.fazel@yahoo.com
Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Isfahan, Iran

http://dx.doi.org/10.47176/TMI.2018.99

Today, identifying diseases using medical image processing and analysis techniques is crucial to help doctors make accurate diagnoses. Consequently, various studies have been developed to extract information from medical images to identify diseases [4]. Numerous diseases affecting the global population are related to the lungs. Therefore, research in the field of lung diseases is significant in public health studies, primarily focusing on chronic lung diseases [5]. Pulmonary fibrosis (PF) occurs when lung tissues become damaged and scarred. This stiff and thick tissue makes it difficult for the lungs to function correctly. As PF progresses, shortness of breath also advances. The onset of this disease is usually silent, accompanied by shortness of breath [6]. As the disease progresses, all lung volumes decrease, followed by a reduction in expiratory flow and lung capacity. Symptoms include dry cough, loss of appetite, weight loss, fatigue, and eventually heart failure. Emphysema is a chronic lung disease where the air sacs (alveoli) in the lungs are over-inflated [7]. Consequently, the elastic fibers that open and close the air sacs during breathing lose their elasticity. This disease usually occurs in adults between the ages of 55-75 and is more common in men than women. The primary cause of emphysema is smoking. Emphysema often presents itself alongside other lung diseases, such as chronic obstructive pulmonary disease (COPD). COPD is the most common cause of death and disability due to lung diseases. COPD includes a group of diseases characterized by airway obstruction, which increases the resistance to airflow in and out of the lungs [4] and [8].

Computed tomography (CT) was invented by Godfrey Hounsfield and German physicist Allan Cormack in 1972. CT imaging, or computed tomography (CT), uses X-rays in conjunction with algorithms and computer calculations to create images of the body. In CT, an X-ray tube is placed opposite an X-ray detector, and with the help of a ring that moves around the patient in a rotational manner, a computer-generated cross-sectional image is produced [3]. CT images have a wide range of gray levels based on the attenuation coefficient of each human tissue, following the Hounsfield scale. This scale, named after the inventor of CT, ranges from -1000 to +1000 Hounsfield units (HU), where zero represents the attenuation coefficient of water and -1000 represents the attenuation coefficient of air [4].

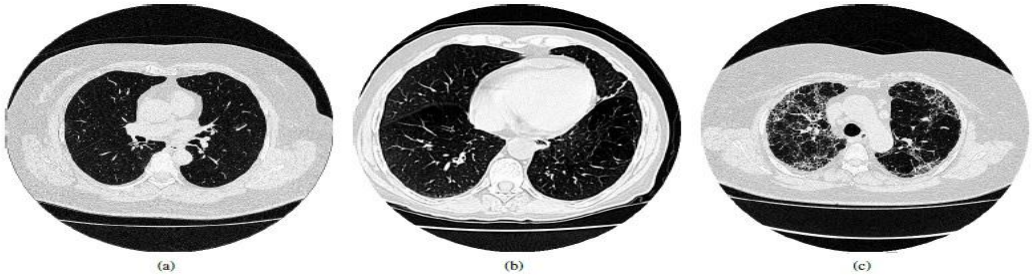

Fig.1. a) Image of healthy lung b) Image of emphysema c) Image of fibrosis [4]

In statistical texture analysis, texture features are calculated through the statistical distributions of observed combinations of gray levels at specific positions relative to each other. Texture statistics are usually categorized into first-order, second-order, and higher-order statistics. The Gray Level Co-occurrence Matrix (GLCM) method is used for extracting second-order texture information. A GLCM is a matrix whose number of rows and columns is equal to the number of gray levels in the image. If an image has G gray levels, the GLCM will be a G×G matrix. The matrix element P(i, j | Δx, Δy) represents the number of times the relationship between two pixels, separated by a distance (Δx, Δy) and having gray levels i and j, occurs based on the defined neighborhood relationship. In other words, the GLCM can extract statistical features such as contrast, mean, variance, entropy, homogeneity, and correlation, considering the distribution of brightness intensities and the position of some pixels simultaneously [6] and [5].

Statistical Moments (SM) are calculated using the gray level distribution of the input image, often derived from the image histogram. These features provide a statistical description of the relationship between different gray levels. Moments of the image are inverse functions to polynomial basis functions, as described by the following equation:

$$M_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} P_{pq} f(x,y) \quad \begin{cases} P_{pq} = x^p y^q & \text{Geometric torque} \\ P_{pq} = (x+iy)^p (x-iy)^q & \text{Complex torque} \end{cases} \tag{1}$$

Geometric and complex moments contain identical information about images and are used to construct rotation-invariant moments [9]. The orthogonal moments of an image, known as Zernike moments, are invariant to rotation and can easily reconstruct the image. This technique can evaluate and measure image features at any moment order by comparing the reconstructed image with the original. These moments are also invariant to scale changes [10]. Teh and Chin, in their study on noise sensitivity and image detail from the first to fifth-order Zernike moments, stated that higher-order moments are more resistant to noise but provide more detailed information about the image [11].

The seven moments invariant to rotation, scale changes, and image translation are known as Hu moments. These moments are recognized as invariant moments, allowing the image to be identified independently of its position, size, and orientation. Despite these properties, they are also invariant to image resolution or high contrast but have very poor discriminative power [4] and [9].

2. NEW FEATURE EXTRACTION APPROACH

The concept of human tissue density analysis using Hounsfield units (HU) obtained from CT images was proposed by Reboucas, Filho, and colleagues [12]. This study presents a new feature extraction method based on human tissue density patterns, named Analysis of Human Tissue Density (AHTD). The proposed method utilizes the radiological densities of human tissues in Hounsfield units to identify suitable features from medical images. The proposed feature extraction method was applied to lung CT images using five density classes (vi). The following classes were defined for the lung CT data:

1. Highly air-inflated lung tissue (1000 to 950 Hu)
2. Normally air-inflated lung tissue (500 to 950 Hu)
3. Low air-inflated lung tissue (500 to 100 Hu)
4. Airless lung tissue (100 to 100 Hu)
5. Bone tissue (600 to 2000 Hu)

In this study, a decision tree based on neighborhood density analysis was proposed to find the best position for initiating an active contour model and to determine whether the objects of interest are inside or outside the lung under analysis [4] and [12]. This study focuses on human tissue density analysis based on CT images for lung disease diagnosis and classification.

The proposed feature extraction method in this study uses the percentage Pi of healthy human tissues based on their density analysis in HU, as per equation (2):

$$P_i = \frac{f(V_i)}{\sum_{j=0}^{N-1} f(V_i)} \tag{2}$$

where N is the number of tissues under analysis.

Based on the previously analyzed region, the function determining the number of points with the given density in each class v_i is defined by equation (3):

$$f(V_i) = \sum_{x=0}^{W} \sum_{y=0}^{H} R(x, y) \tag{3}$$

where W and H are the image dimensions, representing width and height, respectively, and R(x, y) is given by equation (4):

$$R(x,y) = \begin{cases} 1 & \text{and} \quad \lim_{inf}(V_i) < T(x,y) < \lim_{sup}(V_i) \\ 0 & \text{and} \quad Otherwise \end{cases} \tag{4}$$

where $\lim_{inf}(V_i)$ and $\lim_{sup}(V_i)$ are the lower and upper bounds of the density range in HU for class vi, and T(x, y) represents each pixel of the image under analysis.

To introduce the features extracted by the proposed method, Figure (2) shows AHTD maps in RGB for lung CT images based on equation (5), where Pi is defined by equation (2).

$$\mathrm{AHTD_{MAP}} = \begin{cases} \mathrm{R(x,y)} = \mathrm{P_2} \\ \mathrm{G(x,y)} = \frac{\mathrm{P_3}+2\times\mathrm{P_4}}{3} \\ \mathrm{B(x,y)} = \frac{\mathrm{P_0}+2\times\mathrm{P_1}}{3} \end{cases} \quad (5)$$

a b c

Fig. 2. AHTD maps in RGB a) Healthy lung b) Emphysema c) Fibrosis [4]

CT systems were used to obtain image datasets with a resolution of 512 × 512 pixels at 16-bit depth, with tomographic slices defined based on the coordinate plane. The chest CT dataset was acquired from the Walter Cantidio University Hospital in Fortaleza, Brazil [12]. This dataset includes 12 images from healthy volunteers and 24 images from patients, comprising 12 images with lung fibrosis and 12 with emphysema. Since each image shows two lungs, there are 24 samples for each class (healthy, fibrosis, and emphysema). Figure (3) shows an example of grayscale lung images from the original chest CT scans, computed using a window width of 600 HU and a center of 1000 HU.

Feature extraction from the mentioned dataset was discussed and analyzed using a MacBook Pro with a Core i5 2.4 GHz processor, 8 GB RAM, and performed experiments. The proposed feature extraction method was compared with SM, HM, ZM, and GLCM. All these methods are rotation-invariant. The Gray Level Co-occurrence Matrix (GLCM) was computed for correlation, contrast, homogeneity, and energy texture descriptors in four directions [4].

Pattern recognition experiments in this study utilized various machine learning algorithms with different settings for lung disease detection and classification in images. The classifiers used in this study include Bayesian, Optimum-Path Forest (OPF), k-Nearest Neighbor with Euclidean distance (KNN) [13], and Support Vector Machine with Radial Basis Function (SVM). Classification was estimated using 10-fold cross-validation.

Table 1. Accuracy of Methods and Their Ranking

Ranking	Feature	Accuracy
1	AHTM	99.01
2	GLCM	98.67
3	SM	97.56
4	HM	90.22
5	ZM	70.37

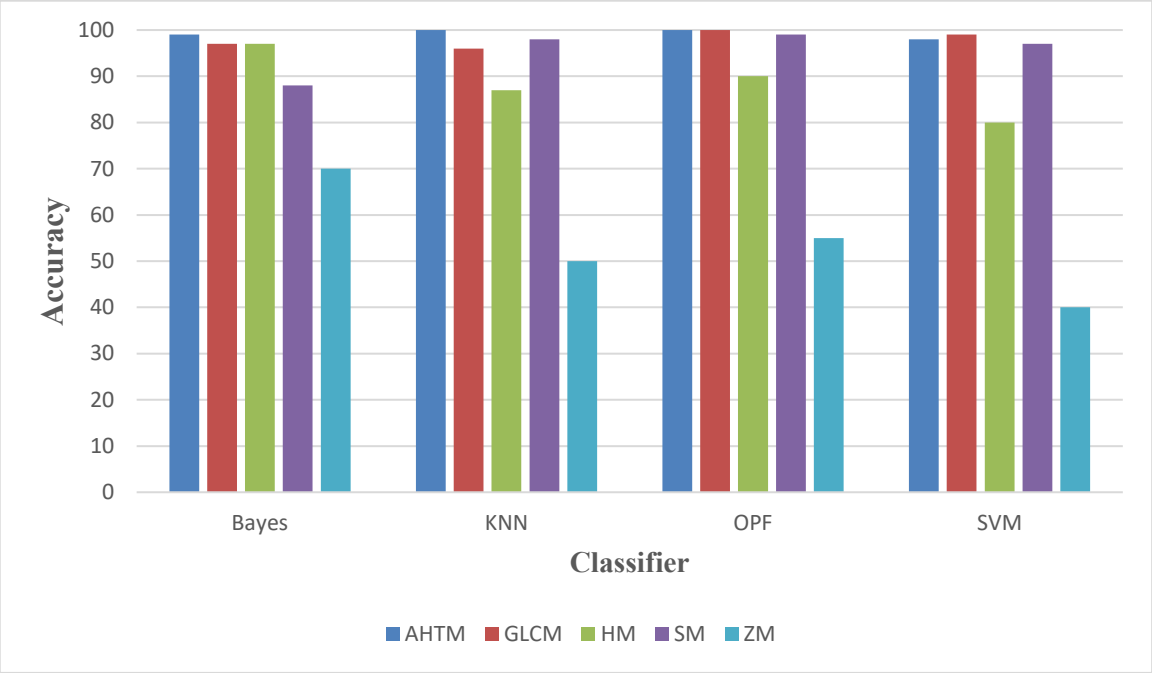

Fig. 3. Accuracy percentage of methods with AHTD using different classifiers [4]

As shown in Figure (3), the AHTD-based feature extraction method has higher classification accuracy than other methods across all classifiers. Additionally, Figure (4) shows that the proposed method has significantly lower computation time compared to other methods.

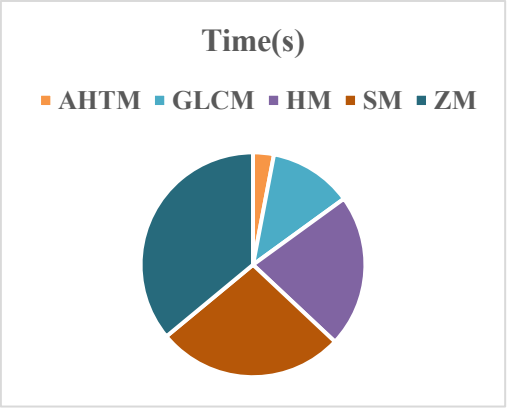

Fig. 4. Comparison of computation time of the proposed method with other methods [4]

3. CONCLUSION

This study presents a new feature extraction method for medical images based on the radiological density of human tissues. The proposed method was compared with the Gray Level Co-occurrence Matrix (GLCM), Hu moments, statistical moments, and Zernike moments. Four machine learning classifiers were used for lung disease classification in chest CT images for these analyses.

The reviewed AHTD method demonstrated faster feature extraction and achieved the highest accuracy for the evaluated dataset, extracting features from lung images in 5.2 milliseconds and achieving 99.01% accuracy for lung disease detection and classification, as shown in Table 1. These results indicate that the proposed method can be

used for disease classification in medical images and can serve as an alternative method for real-time applications due to its rapid processing time and suitable features.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Khan, I. Y., Zope, P. H., & Suralkar, S. R. (2013). Importance of artificial neural network in medical diagnosis of diseases like acute nephritis disease and heart disease. *International Journal of Engineering Science and Innovative Technology*, 2(2), 210-217.

[2] Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications*, 36(4), 8610-8615. https://doi.org/10.1016/j.eswa.2008.10.032

[3] Ozkan, H., Osman, O., & Sahin, S. (2013). Computer aided detection of pulmonary embolism in computed tomography angiography images. In *2013 International Conference on Electronics, Computer and Computation (ICECCO)* (pp. 355-358). https://doi.org/10.1109/ICECCO.2013.6718301

[4] Rebouças Filho, P. P., de S. Rebouças, E., Marinho, L. B., Sarmento, R. M., Tavares, J. M. R. S., & de Albuquerque, V. H. C. (2017). Analysis of human tissue densities: A new approach to extract features from medical images. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2017.02.005

[5] Ramalho, G. L. B., Rebouças Filho, P. P., de Medeiros, F. N. S., & Cortez, P. C. (2014). Lung disease detection using feature extraction and extreme learning machine. *Revista Brasileira de Engenharia Biomédica*, 30(3), 207-214. https://doi.org/10.1590/rbeb.2014.019

[6] Neto, E. C., Cortez, P. C., Cavalcante, T. S., Rodrigues, V. E., Reboucas Filho, P. P., & Holanda, M. A. (2016). 3D lung fissure segmentation in TC images based in textures. *IEEE Latin America Transactions*, 14(1), 254-258. https://doi.org/10.1109/TLA.2016.7430087

[7] Pforte, A. (2004). Epidemiology, diagnosis, and therapy of pulmonary embolism. *European Journal of Medical Research*, 9(4), 171-179.

[8] Eskildsen, S. F., Coupé, P., Fonov, V. S., Pruessner, J. C., & Collins, D. L. (2015). Structural imaging

biomarkers of Alzheimer's disease: Predicting disease progression. *Neurobiology of Aging*, 36(S1), S23-S31. https://doi.org/10.1016/j.neurobiolaging.2014.04.034

[9] Gonzalez, R. C., & Woods, R. E. (1992). *Digital image processing*. Addison-Wesley.

[10] Khotanzad, A., & Hong, Y. (1990). Invariant image recognition by Zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5), 489-497. https://doi.org/10.1109/34.55109

[11] Teh, C.-H., & Chin, R. T. (1988). On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4), 496-513. https://doi.org/10.1109/34.3913

[12] Reboucas, F., Pedro, P., Cortez, P. C., & Holanda, M. A. (2011). Active contour models CRISP: New technique for segmentation of the lungs in CT images. *Revista Brasileira de Engenharia Biomédica*, 27(4), 259-272. https://doi.org/10.4322/rbeb.2011.021

[13] Jurkovic, I.-A., Stathakis, S., Papanikolaou, N., & Mavroidis, P. (2016). Prediction of lung tumor motion extent through artificial neural network (ANN) using tumor size and location data. *Biomedical Physics & Engineering Express*, 2(2), 025012. https://doi.org/10.1088/2057-1976/2/2/025012