ISSSD 2019

XIX INTERNATIONAL SYMPOSIUM ON SOLID STATE DOSIMETRY

Host

CONSEJO ZACATECANO DE CIENCIA TECNOLOGÍA E INNOVACIÓN

Universidad Autónoma de Zacatecas
"Francisco García Salinas"

Proccedings
Volume I

Zacatecas, Zac. México
October 7 to 11th, 2019

Tuberculosis disease: Diagnosis by image processing

Eduardo Perez-Careta[1], Wendy Morales-Castro[1],
Teodoro Córdova-Fraga[2] and Rafael Guzmán-Cabrera[1]

[1]Universidad de Guanajuato, km 3.5 + 1.8 road Salamanca-Valle de Santiago
Salamanca, Guanajuato Mexico 36730.

[2]Universidad de Guanajuato, Campus León
León, Guanajuato México

E-mail: eduardo.perezc@me.com

Abstract

Tuberculosis is one of the first human diseases of which there is evidence, it´s very harmful and easy to spread through the air, this disease is caused by the bacterium "Mycobacterium tuberculosis". One way to detect tuberculosis is by chest x-rays, by analyzing the x-ray you can obtain the detection of any abnormality (parenchymal, ganglionic or pleural). The development of algorithms for image processing has gained greater momentum in the medical field, essentially helping for diagnosis and treatment. This paper presents a method that allows the presence of tuberculosis in medical X-ray images to be identified. Three methods of classification were implemented for the evaluation of the method: Support Vector Machine, Logistic Regression and K-Neighbors Classifier. Two classification scenarios were implemented: cross validation and training and test sets. The results obtained allow us to see the viability of the proposed method.

Keywords: Tuberculosis Detection, Classification of medical images, Medical diagnosis.

1. INTRODUCTION.

Tuberculosis (TB), also known as Tisis, is a chronic infectious disease, caused by a germ called “Mycobacterium tuberculosis”. The bacteria usually attack the lungs mainly, but can also damage other organs of the human body. TB spreads through the air, when a person with pulmonary TB coughs, sneezes or speaks [Jaeger 2012], thus causing the spread of the disease. It can be preventable and curable if it´s detected on time otherwise it could cause the death of the patient. To find out if a person has tuberculosis disease, tests such as a chest x-ray or a culture of a sputum sample (phlegm that is expelled from the lungs when coughing) [Jaeger 2014] can be performed. In Mexico, the main cause of mortality is tuberculosis with a rate of 9.24% per 100 thousand inhabitants, data provided by the Statistical and Death System of the General Directorate of Epidemiology. According to the World Health Organization (WHO), tuberculosis is one of the 10 leading causes of mortality in the world. In 2015, about 10.4 million people became ill with tuberculosis and 1.8 million died from this disease. More than 95% of deaths from tuberculosis occur in third world countries [Hogeweg 2010]. In Figure 1 we can see 4 X-ray images, the two images on the left show healthy patients and the two on the right show patients with the detected disease.

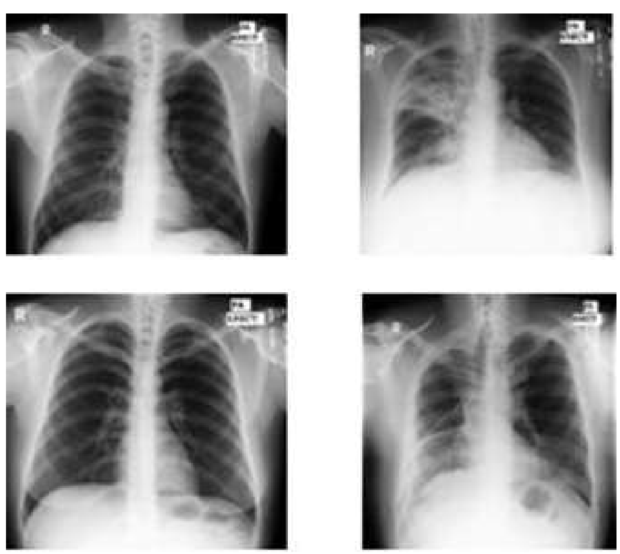

FIGURE.1.- X Ray Images.

Globally, the incidence of tuberculosis is decreasing at a rate of approximately 2% annually. The methods of diagnosis and treatment of the disease have improved, however, it will not be possible to achieve the goal of the End of TB Strategy by 2020 [Xu 2013]. Undoubtedly, this public health problem remains a great challenge for the health system of the countries, mainly in the process of development, it´s for this reason that in this work a methodology based on the automatic classification of medical images that allow to the expert to improve his medical diagnosis through the use of algorithms based on the application of artificial intelligence. The rest of the article is organized as follows: a brief description of the related work is presented in section 2. Section 3 presents the proposed method, as well as a description of the different stages that compose it. Finally, the results of the evaluation of the proposed method and the conclusions of this work are presented.

2.- RELATED WORK

In this paper, artificial intelligence is applied in the automatic classification of chest radiograph images of patients with tuberculosis and without tuberculosis. The term Medical Image covers different lines of work that complement each other, since it includes the set of specific techniques for image acquisition and reconstruction by the multiple available technologies; the display of said images; compression and storage thereof; the improvement of its quality and contrast; the tasks of registering images together with their processing for the segmentation and extraction of information from said images, such as the quantification of anatomical and physiological parameters. The integration of all these disciplines allows addressing multiple applications, such as aid in surgical planning and image-guided interventions, the description of anatomical regions based on atlases, the visualization and analysis of organs and tissues, among others. A very interesting review of the evolution of medical image analysis and processing techniques since the 1980s can be found in [Duncan *et al*., 2000]. Image classification refers to the task of extracting information classes from an image, there are two types of classification: supervised and unsupervised, Supervised classification part of a set of known classes, these classes must be characterized according to the set of variables by measuring them in individuals whose membership in one of the classes does not present doubts, while the unsupervised classification does not establish any class, although it´s necessary to determine the number of classes we want to establish, and let them be defined by a statistical procedure [Karargyris *et al*., 2016]. In our case we intend to classify images into two possible categories: sick or healthy. If we use the random classification method, we would have a result of 50% probability of hitting the classification, while the automatic classification methods throw us 80%, 90% or more, which gives us greater security in the assertiveness of the method, some advantages of the automatic classification method are:

- No expert is needed to classify images.
- Reduces the tedious work of experts.
- It greatly reduces human errors that may arise.
- It´s more effective than the method of chance.
- You can classify large amount of images in a short time.

Image processing is widely used, for example, in [Guzmán-Cabrera *et al*., 2017] image processing is used to extract regions of interest with properties that may be potentially related to Parkinson's medical diagnosis, use computer-assisted diagnostic technology to process the images, extract the textures, make a segmentation of the image and find the area of interest. Within [Quintanilla *et al*., 2016] we find the use of image processing, pattern recognition and artificial intelligence to help detect a cluster of micro calcifications in digitalized mammography images. There are other techniques to perform image processing such as [Martinez *et al*., 2019] using a Raman spectroscopy technique to obtain spectral maps with specific spatial resolution (1 to 5 micrometers) over a selected region of the sample to access and visualize relevant information about the spatial distribution in any sample about its biochemical composition. Some works of interest where a compendium of techniques for improving medical images can be found are [Sonka *et al*., 1993; Clarke *et al*., 1995; Vovk *et al*., 007], and for eliminating image noise using techniques ranging from erosion, extraction and others commonly used in the state of the art, they can be consulted in [Bao and Zhang 2003; Buades *et al*., 2012; Manjón *et al*., 1980]. Other research found on image processing is found in [Guzmán- Cabrera *et al*., 2012] which, like [Quintanilla *et al*., 2016] helps the early diagnosis of breast cancer using image processing, within this investigation we can find that it uses the technique of texture segmentation, the images used as evidence are from a database, which has images of carcinogenic masses and micro calcifications manually labeled by experts. In [Guzman-Cabrera *et al*., 2016] they carry out the identification of breast cancer using thermal images, perform a digital processing of the images, using a texture analysis of the images to identify and extract all regions of interest.

3.- PROPOSED METHOD.

The proposed method was developed in Python. The characteristics of the images that are used as classification attributes are extracted with KERAS. KERAS is a library of Open Source Neural Networks created in Python that contains the architecture of RESNET50, this architecture will help to extract the characteristics of the images through arrangements. In this work, three classification methods were used: The first method used is SVM (Support Vector Machine), which is a supervised learning model with associated algorithms that analyze data and recognize patterns, is used for classification and analysis regression The second method used is Logistic Regression, which is a machine learning classification algorithm that is used to predict the probability of a categorical dependent variable that is dichotomous, that is, that contains data that can be classified into one of two possible categories. (alive or dead, sick or healthy, yes or no, etc.). A logistic regression, therefore, requires that the dependent variable be binary. In addition, the level 1 factor should represent the "desired" value. Only significant variables should be included as independent variables that, in turn, should be independent of each other. The third method used is K-Neighbors Classifier is an algorithm based on instances of supervised type of Machine Learning. It can be used to classify new samples (discrete values) or to predict (regression, continuous values). It essentially serves to classify values by looking for the most similar data points (by proximity) learned in the training stage and by guessing new points based on that classification [Turney 2002]. For this work the Montgomery database was used, the X-ray images of this database were acquired from the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA. UU. This set contains 138 posterior-anterior radiographs, of which 80 radiographs are normal and 58 radiographs are abnormal with manifestations of tuberculosis. All images are identified and available in DICOM format. The set covers a wide range of anomalies, including spills and military patterns. The data set includes radiology readings available as a text file. Each of the images has a label that helps identify the images. The labels can be: with TB (label with the number "1", success), and normal or without TB (Label with the number "0", failure). Preprocessing was performed on the images used, the diagram of the preprocessing process is shown in Figure 2.

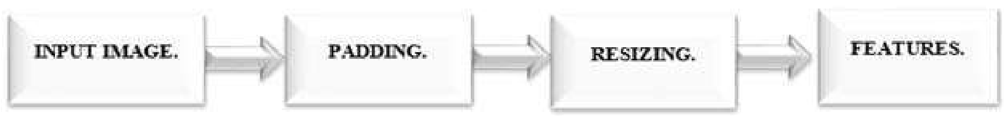

FIGURE.2.- PRE PROCESSING.

As can be seen in this figure, there are two main parts in the preprocessing: (1) Padding (2) Resizing. Both stages go after the input of the images, resulting in a matrix for each input image with dimensions of 224x224 and with numbers from 0 to 255, this corresponds to an image of 224x224 in 3 channels (RGB). With this matrix, the arrangements with dimensions of 2048 will be obtained, which will contain the characteristics of the images. In the Padding option, X-ray images within the data set are of variable size and most images have only one color channel and some have 3 color channels. And because the model cannot process the images in this way, first all the images are converted to uniform dimensions by applying an additional fill in the image that creates a uniform dimension, that is 450x450 pixels in Portable Network Graphics format (PNG). Otherwise, for resizing (Resize Image), memory is the biggest challenge for image classification models, for this reason it´s necessary to make the resizing of images Resizing is the most used technique to overcome errors of memory, once the padding was applied to the resizing images change the size to 128x128 pixels in size. All images used for processing were previously preprocessed with these tools. The last stage of preprocessing consists in obtaining the characteristics that will be used as classification attributes. 2048 features are obtained for each image.

3.1. Processing for cross validation.

Once the arrangements with the characteristics of the images for TB and Normal are obtained, the labels are created in a text document to name each of the images that will be used for the training of the program. Within the processing program, labels and features are called and relationships between tags and features are created and then converted into arrangements. When the program finishes ordering the data for better interpretation, it converts the relationships to 0 and 1, in order to interpret them, then the part of the classification methods mentioned above begins (SVM, Logistic Regression and K-Neighbor Classifier). Figure 3 shows the diagram of the processing performed for the cross-validation scenario, with each of the methods.

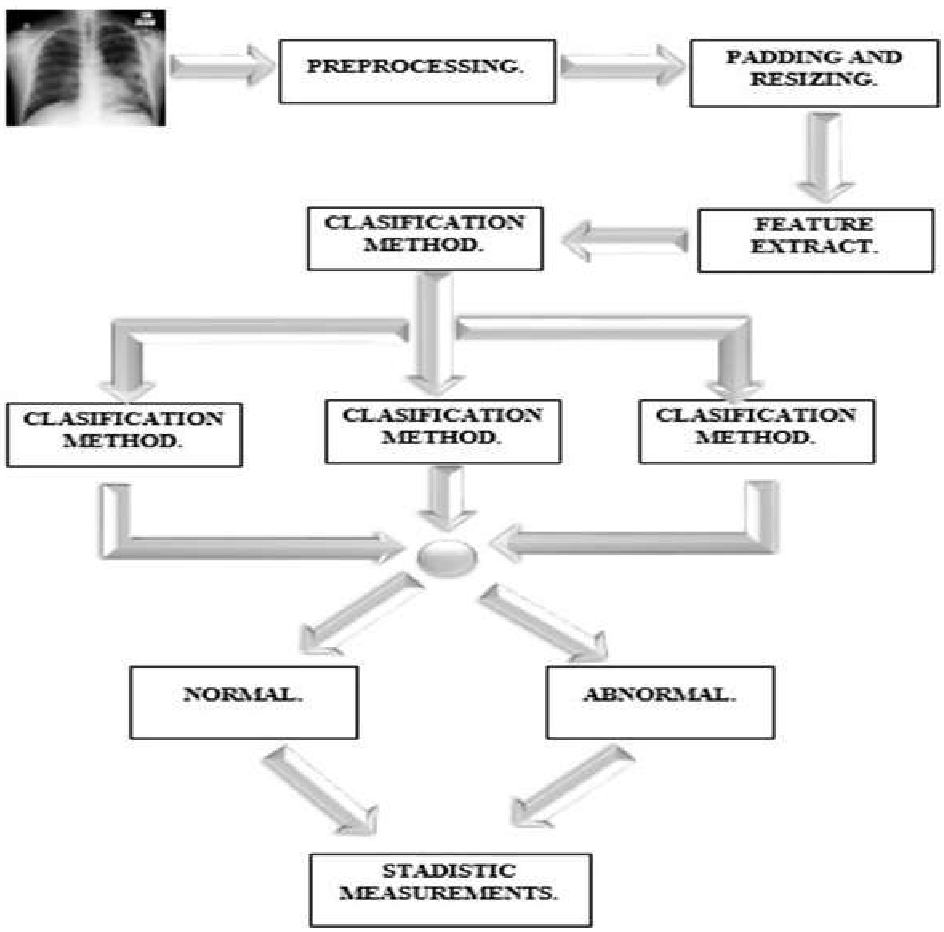

Figure. 3.- Cross Validation Diagram

3.2. Processing for training and test set.

Training and test sets were created for the second scenario. 80% of the images were used for training and the remaining 20% was used for testing, this so that the test images are not seen again by the training set, just like the first scenario the arrangements were created With the characteristics of the images and labels for training and testing in a text file lastly, preprocessing provides four text files that contain the labels and characteristics of each set created. These files enter the classifier where the relationships between characteristics and labels of each set are made, and after this way, carry out the classification with the aforementioned methods and also perform tests with the different values of "C" as in the scenario of cross validation. The diagram of the Methodology used for the training sets formed is shown in Figure 4.

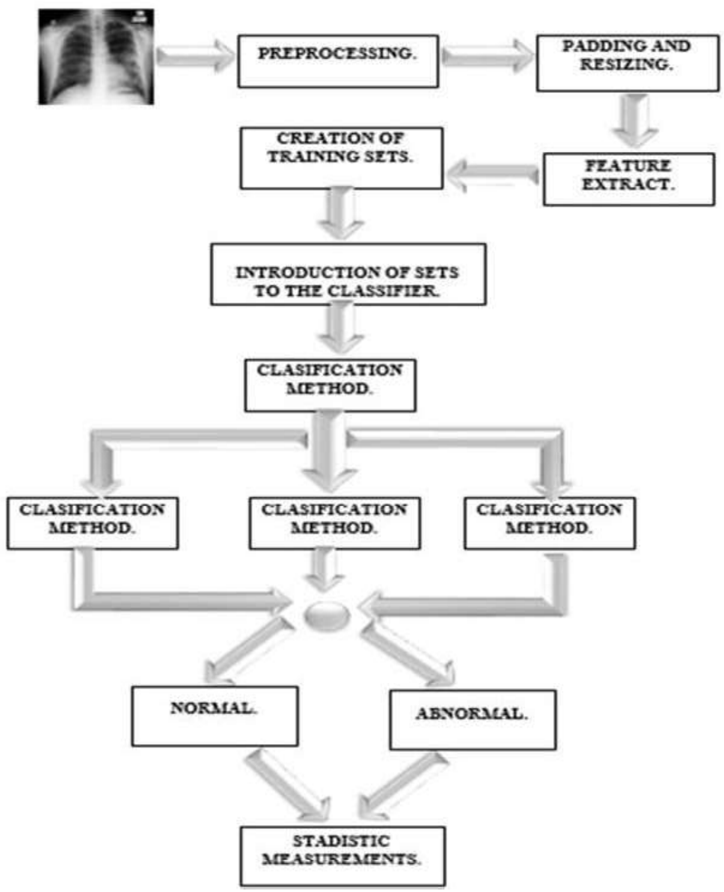

Figure. 4.- Training Set Diagram.

For each classification made, we recorded the following evaluation metrics: Accuracy, Precision, Recall, F1.

4. RESULTS.

The evaluation metrics used allow us to observe which of the scenarios used obtains better results with the classifiers used.

In table 1 we can see the data obtained in the cross-validation scenario. And in table 2 the data obtained with the training and test sets formed.

Table 1. Cross Validation Results.

Classification	Accuracy	Precision	Recall	F1
SVM	0.76	0.80	0.74	0.73
Logistic Regression	0.75	0.79	0.73	0.72
K- Neighbors Classifier	0.70	0.68	0.67	0.66

Table2. Training Sets Results.

Classification	Accuracy	Precision	Recall	F1
SVM	0.86	0.86	0.85	0.85
Logistic Regression	0.81	0.82	0.79	0.80
K- Neighbors Classifier	0.63	0.61	0.60	0.60

As can be seen both in the tables and in Figure 5, the precision and accuracy values obtained are greater than 85%, definitely much higher than the accuracy provided by chance. The scenario that shows the best performance is the training and test sets, which is definitely the best classification scenario (most desirable to perform) because the training set never sees the test set, thus avoiding the to be able to have some type of influence when carrying out the assignment of the category to the image under study. It can also be seen that in both scenarios the classifier that shows the best performance is SVM.

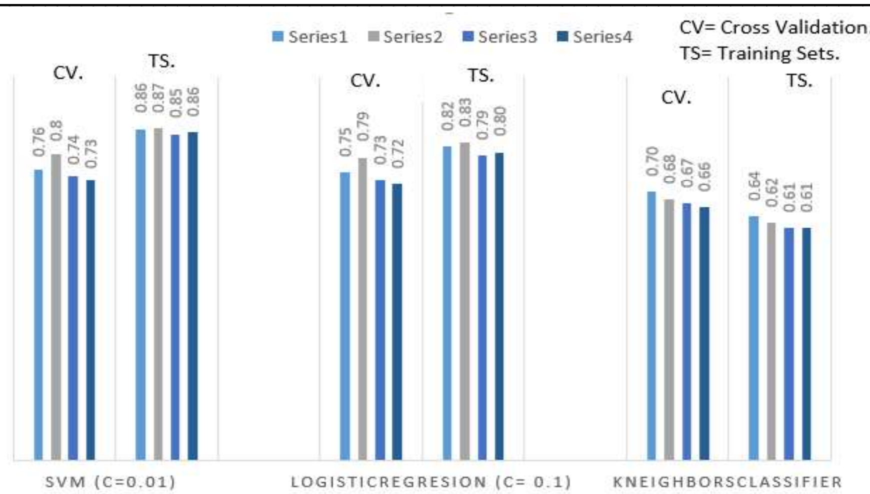

Figure 5. Result Graph.

5. CONCLUSIONS.

This paper presents results of automatic classification of medical images into two categories: with and without TB. Two classification scenarios were implemented: cross validation and training and test sets. The scenario with the best results was in which training and test set were formed with an accuracy greater than 85%. The classification method that shows the best performance in the two scenarios implemented in this work is SVM. As can be seen in the results obtained in the present work, they overcome by far random, and allow to carry out the classification of images in an efficient way avoiding in this way having a specialist perform this work, which in a Large volume of images would become a tedious job. The results obtained allow us to see the feasibility of the methodology used. It also allows us to identify the best classification scenario and machine learning method to carry out the classification of radiographs with and without tuberculosis.

REFERENCES.

Bao, P., Zhang, L. (2003). *Noise reduction for magnetic resonance images via adaptive multiscale products thresholding*. IEEE Trans. Med. Imaging **22**: 1089–99.

Buades, A., Louis Collins, D., Robles, M. (2012). *New methods for MRI denoising based on sparseness and self-similarity. Med. Image Anal*. **16**:18–27.

Clarke, L., Velthuizen, R., Cmacho, M., Heine, J., Vaidyanathan, M., Hall, L., Thatcher, RW, Silbiger, M. (1995). *MRI segmentation: methods and applications. Magn. Reson. Imaging* **13**: 343–368.

Duncan, J., Ayache, N. (2000). *Medical image analysis: Progress over two decades and the challenges ahead.* IEEE Trans. Pattern Anal. Mach. Intell. **22:** 85–106.

Guzmán-Cabrera R., Gómez-Sarabia M., Torres-Cisneros M., Escobar-Acevedo M., Guzman-Sepulveda J.R. (2017). *Parkinson´s Disease: Improved Diagnosis using image processing*. Proceedings Photonics North: 22–24.

Guzmán-Cabrera R., Guzman-Sepulveda J.R., González A., Rosales J., Torres M. (2016). *Digital Processing of Thermographic Images for Medical Applications.* Revista de Chemie, **67:** 53-56.

Guzmán-Cabrera R., Guzmán-Sepúlveda J.R.,· Torres-Cisneros M., May-Arrioja D., Ruiz-Pinales J., Ibarra-Manzano O. (2012). *Digital Image Processing Tecnique for Breast Cancer Detection.* International Journal of Termophysics, **33:** 150-163.

Hogeweg L. (2010). *Fusion of local and global detection systems to detect tuberculosis in chest radiographs*, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer: 650–657.

Jaeger S. (2012). *Detecting tuberculosis in radiographs using combined lung masks*, Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE: 4978–4981.

Jaeger S. (2014). *"Automatic tuberculosis screening using chest radiographs"*, IEEE transactions on medical imaging, **33**: 233–245.

Karargyris A. et al. (2016). *"Combination of texture and shape features to detect pulmonary abnormalities in digital chest x-rays"*, International journal of computer assisted radiology and surgery, **11:** 99–106.

Manjón, J. V, Coupé, P., Marr, D., Hildreth, E., (1980*). Theory of edge detection. Proc. R. Soc. London. Ser. B, Biol.* Sci. **2**: 187–217

Martínez J. C., Guzmán-Cabrera R., Córdova T., (2019). *Generation of Raman images through spectral.* Journal of Intelligent & Fuzzy Systems **36**: 4991-4999.

Quintanilla J., Gordillo-Sosa J.A., Estrada-García H., Barrón-Adame J., Lozano-GarcíaJ., Guzmán-Cabrera R., (2016). *Analysis of Mammograms Using Texture Segmentation*. Computación y Sistemas, 20: 279-288.

Sonka, M., Hlavac, V., Boyle, R., (1993). *Image Processing, Analysis, and Machine Vision Second Edition, Chapman an. ed.*

Turney, P. D. (2002). *Thumbs up or thumbs down: semantic orientation applied to unsupervised classification of reviews.* In Proceedings of the 40th annual meeting on association for computational linguistics (ACL): 417–424.

Vovk, U., Pernus, F., Likar, B., (2007). *A Review of Methods for Correction of Intensity Inhomogeneity in MRI.* Med. Imaging, IEEE Trans: 34-42.

Xu T. (2013), *Novel coarse-to-fine dual scale technique for tuberculosis cavity detection in chest radiographs*, EURASIP Journal on Image and Video Processing, **2013**: 1-18.