SOM Neural Networks as a Tool in Pleural Tuberculosis Diagnostic

Alvaro D. Orjuela-Cañón
GIBIO Electronics and Biomedical Faculty
Universidad Antonio Nariño
Bogotá D.C., Colombia
Signal Processing Laboratory
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
dorjuela@ieee.org

José Manoel de Seixas
Signal Processing Laboratory
COPPE/Poli Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
seixas@lps.ufrj.br

Anete Trajman
Post-graduation in Health Education,
Gama Filho University
Rio de Janeiro, Brazil
Medical School,
McGill University
Montreal, Canada
atrajman@gmail.com

***Abstract* — Diagnostic tasks in a contagious disease as pleural tuberculosis is essential to send the patients to a correct treatment to cure the ailment. Tools for supporting this task are still a challenge because different variables are necessary, including invasive standard tests. In the present work Self Organizing Maps were used for clustering of patients of pleural tuberculosis in three risk groups. The first approach employs just anamnesis variables and a second methodology uses additional information about some classical result tests. The last approach exhibits best results compared to the approach using anamneses variables**

Keywords—Neural Networks; Diagnosis; Pleural Tuberculosis; Clustering

I. INTRODUCTION

Tuberculosis (TB) is an infectious disease considered global emergency by the World Health Organization (WHO) because it is the second leading cause of death from infectious diseases each year [1]. This disease is caused by *Mycobacterium tuberculosis*, and usually is divided into pulmonar (affecting the lungs) and extrapulmonar (affects other body parts). TB is transmitted through the air when a person with pulmonar TB coughs and expels the bacteria. Each year, TB causes millions of deaths, as mention the latest statistics reported in 2011, where nine million of new cases and 1.4 million of deaths were reached. This reason leads the WHO to declare TB as emergency global public health [1].

In Colombia were reported a total of 11,699 cases and the country was placed as the fifth highest number of reported cases in 2011, making a national public health problem. By 2012, 10,194 cases were reported nationwide, with Valle and Antioquia regions as the most affected [2].The situation is even more critical in Brazil, where in 2010, 71,000 new cases were found and an incidence of 37.2 for each thousand patients was appointed[4][4].

Additionally, in Colombia, in the last 13 years there have been 12,781 cases of TB BK is negative and highly probable cases have been lost due to the absence of complementary diagnosis [2][3]. In the Rio de Janeiro state exists the most incidence number with 74,06 from 100.000 habitants, and in the city of Rio de Janeiro, the incidence reaches 66,94 new cases for each 100.000 habitants [3].

A strategy for disease control is to meet and treat patients for smear (BK) positive. The BK is a quick and inexpensive diagnostic methodology but has low sensitivity ranging between 20% and 42% due to paucibacillary of HIV positive patients. A negative BK can avoid picking up a patient if not used complementary methodologies such as sputum culture, which is not accessible in all regions of developing countries [2].

Pleural tuberculosis (pTB), differently to pulmonary TB, happens when the pleura is affected by the TB bacillus. This kind of extrapulmonary TB is one of the most frequently found. The diagnostic in pTB is a current problem because information from a test based on culture of pleural fluid is necessary. These cultures are obtained by removing tissue through thoracentesis procedure, which is invasive, expensive and required well trained human resources [5][6].

Neural networks have proven to be effective in supporting the tuberculosis diagnostic, using models that have been developed based on clinical and histopathological variables, where tools for the diagnostic with characteristics as noninvasively, online applications or *triage* approaches are necessary. An example of this can be sought in the pulmonary tuberculosis problem, where Multi-Layer Perceptron networks (MLP) were used[7][8]. In pleural tuberculosis (pTB) also these techniques have shown relevant results using MLP neural networks [9][10].

Clustering tasks have been implemented for classifying the patients in risk groups. Adaptive Resonance Theory (ART) neural networks were used to this kind of classification, obtaining interesting results [11]. Similar approach was implemented in pulmonary TB but using Self Organizing Maps (SOM) and a *k-means* algorithm for classification of patients into different risk groups [12]. Also, for pTB

clustering is an alternative to classify the patients in low, medium and high risk of having the disease. Fuzzy-ART neural networks were implemented to do this task, reaching good classification results [13].

The present work presents results showing how SOM neural networks perform a clustering in pleural tuberculosis patients. This article is organized as follows: Section II describes the database used, information about SOM neural network design and its training. Section III shows the results obtained for classification, and a discussion of these results is provided in Section IV. Finally, conclusions are presented in Section V.

II. MATERIALS AND METHODS

This section describes the data used and the methodology applied to train the neural network. Characteristics about SOM design and how the clustering post-training was employed and also provided.

A. Database

The database used in this work was also used in a previous work [4]. All patients with pleural effusion admitted for diagnostic at the Hospital Geral da Santa Casa da Misericórdia (Rio de Janeiro, Brazil) were included in the study, after acquiring to sign a participation agreement.

Two approaches are assessed in this work; first one, called pre-test analysis, uses just four anamnesis variables: age, gender, smoking status and HIV status. This approach is implemented to support *triage* tasks, where additional information is not provided. A second approach has nine variable, which includes anamnesis variables and some classical test results, as adenosine deaminase (ADA), acid-alcohol resistant bacillus (BAAR), serology test (ELISA), polymerase chain reaction (PCR), and pleural fluid culture. This approach is called post-test analysis, and yields a support for the medical diagnosis. Table I summarizes both approaches.

Each patient's record was consulted to gather information on 11 variables. The two added variables correspond to a pleural tissue culture, and a histopathological test using a biopsy of the pleura [9].The results of these tests were not used because it sought a diagnosis that is little invasive and painful to the patient. Additional to each 11 variables, the final diagnostic is stored in the database.

The variables were coded as +1 for a positive result, -1 for a negative result, or zero in case the specified test was not available or the patient failed to provide information, except forage, whose numerical value (in years) was kept normalized, and gender, which was coded as +1 for male and -1 for female patients.

The database contains records of 135 patients, 96 have been diagnosed with pTB and 36 of them were diagnosed without the disease.

TABLE I. TWO DESIGN APPROACHES

Type of test	
Pre-test ***(Anamnesis)***	***Post-test*** ***(Experimental Tests Results)***
Age Gender Smoking Status HIV Status	Age Gender Smoking Status HIV Status Adenosine deaminase (ADA) Acid-alcohol resistant bacillus (BAAR) Serology (ELISA) Polymerase chain reaction (PCR) Pleural fluid culture

B. SOM Training Process

SOM neural networks are capable of arranging the input data into a discretized two-dimensional space (a map), which attempts to preserve the topological properties of the original input space. This is motivated by the behavior of visual, aural and sensory areas of human cerebral cortex [14].

As difference from MLP models, the main advantage of SOM architectures is training, as in the most of cases is made in an unsupervised way. This is useful in clustering tasks due to similarities in the data can be found by the map [15], and in k groups may be assigned efficiently to patients.

SOM uses the information from the input to do a representation across a nonlinear mapping in an output space with reduced dimensionality. This new space is taken to analyze the original dataset in a graphical way, where different areas of the map preserve characteristics of the classes employed in the training process [14].

Learning process is conformed by three stages: competitive, cooperative and adaptive. In competitive learning, Euclidian distance (weights) from each input to all units or neurons is computed. Unit with weight most similar to the input is defined as the best matching unit (BMU). Then a cooperative process is given around BMU, and units close to it are updated based on a neighborhood function. Finally, adaptive process changes BMU weights according to the input [12][16]. This is reached through the expression:

$$w_i(t+1) = w_i(t) + \eta(t)h_{ij}(t)(x(t) - w_i(t)) \qquad (1)$$

where $w_i(t)$ are weights of the map, $\eta(t)$ is a learning coefficient, $h_{ij}(t)$ is a neighborhood function and $x(t)$ is the input vector.

For training SOM network is necessary to proportionate: number of units, size, type of lattice map and neighborhood function parameters. Number of units and size define the map resolution, type of lattice defines unit's arrangement from regular or irregular forms, and the base size of the neighborhood function controls cooperative process.

There are heuristical rules to compute number of units and dimension map, one of them is based on principal component analysis (PCA). Ratio of first and second principal components from the training dataset can be an initial value for obtaining the length and width relation of the map [14]. As

the variables are not continues, PCA analysis was not the appropriated to calculate the dimensions ratio. In this case, a Multiple Correspondence Analysis (MCA) was used because the data was presented in a categorical way [17]. Computation of the map dimensions were obtained from MCA inertial information, in a similar way to PCA analysis, just that for categorical data. In addition to that, it is attempted what all units have been activated by the data. These rules were followed to determine the number of units and size. Hexagonal topology for lattice was implemented due to distance between adjacent units in the beginning of the training is same.

Finally, neighborhood function establishes how strong the link between units is. In the present work, it is based on Gaussian distribution, given by:

$$h_{ij}(t) = \exp(-d_{ij}^2 / 2\sigma^2(t)) \quad (2)$$

where d_{ij} is the Euclidian distance between the j unit and BMU, and $\sigma(t)$ is the basis of the function in the iteration t. This parameter changes during the training, beginning with a basis of four units and ending with just a one unit.

The map size and area where the neighborhood function has significant values determine accuracy and generalization of classification [16].

As three risk groups are proposed: high, medium and low, the training of the SOM is developed searching this number of clusters. This assists the training and the choice of final model for the classification. It is important to note that all units of the map have been activated by any input pattern because units without activation can confuse the classification. In this way, models with units without activations are not considered.

C. Post-training Analysis

As number of output units of map can be different to three groups, as is required by the main objective of this work, it is necessary a post-training analysis after the map training. This procedure takes advantage of nonlinear processing that makes the map and can be developed by a simpler clustering process.

In this stage, the units of SOM network are clustered in three main groups. This work is developed by the *k-means* algorithm [18], who builds the groups using the measure of distances between the clusters in the map. Three groups were pre-defined to do an analysis similar to *triage* diagnosis of diseases.

According with the number of activations of the units in each group, labels were given using the number of patients with and without TB. Labels as: high, medium and low were used depends that number. The group with the highest number of pTB cases was named with high risk group, and the group with the highest number of cases without pTB was called as low risk group. When the number of cases was similar for pTB and non pTB the group was labeled as medium risk group.

III. RESULTS

Results after SOM training are shown in two parts: pre-test and post-test analyses. For pre-test analysis, it was found that the appropriate dimension ratio was 1.1. This value was obtained by the inertial information from MCA analysis. The selected map has a dimension of 3 x 3 units. Nine outputs of this map were used in the *k-means* algorithm to classify the units in the three risk groups. Figure 1.1 shows the results for the map and the output clustering in the pre-test analysis. In each unit the number of activations given for patients with pTB are marked by positive (+) pTB, and activations given for patients without pTB are marked by negative (-) pTB.

According with the number of activations in each group, the selected area of the map was labeled. Table II shows the results for pre-test analysis, where a percentage of patients is calculated. In this case, the medium risk group was labeled in this way due to the balance in both classes. High risk group was labeled in this way because the 84.48% of its activations are due to patients with pTB.

In post-test analysis were developed the same stages as pre-test case, but the SOM had as input nine variables. Dimensions ratio was defined by the MCA analysis, obtaining a ratio of 1.2676, which was extracted using the first and second inertia. With this information and maintaining all units activated, a map with dimensions of 5 x 4 was used.

After training, the *k*-means algorithm was used to obtain the three risk groups. The trained map with three sections after the *k-means* is shown in the Figure 2. In the same way as pre-test the figure shows the number of activations in each case.

Table III shows the results for this clustering, where the high risk group is obtained because the units of this section of map were activated by patients with pTB. The low risk group contains the highest number of activations given by the patients without the disease, and the medium risk group is obtained because activations in this section of map are from pTB and just one without the disease.

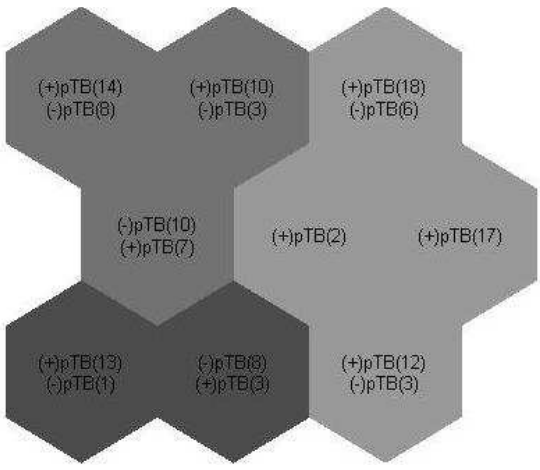

Fig. 1. Map with the three risk groups by the pre-test analysis.

TABLE II. RESULTS FOR PRE-TEST ANALYSIS

RISK GROUPS	Patients with TB		Patients without TB		Total
	Pacientes	(%) TB	Pacientes	(%) NTB	
High	49	84.48%	9	15.52%	58
Medium	31	59.62%	21	40.38%	52
Low	16	64%	9	36%	25
TOTAL	96		39		135

TABLE III. RESULTS FOR POST-TEST ANALYSIS

RISK GROUPS	Patients with TB		Patients without TB		Total
	Pacientes	(%) TB	Pacientes	(%) NTB	
High	66	100%	0	0%	66
Medium	15	93.75%	1	6.25%	16
Low	15	28.30%	38	71.70%	53
TOTAL	96		39		135

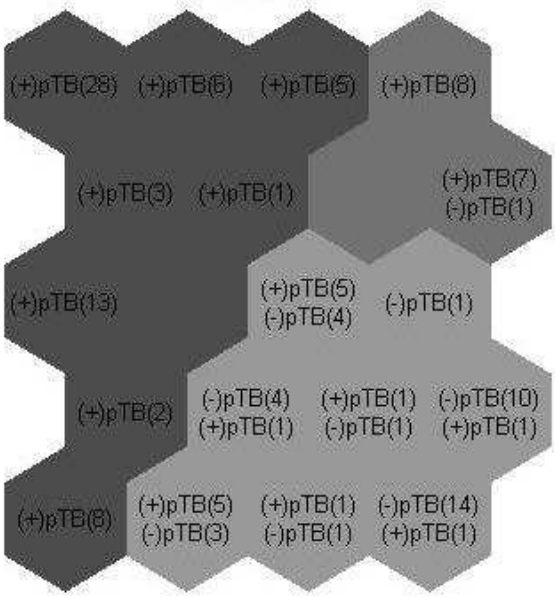

Fig. 2. Map with the three risk groups by the post-test analysis.

IV. DISCUSSION

It is possible to observe that the pre-test analysis yield a poor result because a low risk group cannot be well defined. This clustering can be used in two important groups for disease detection. In that case, a combination of groups of high and medium risk is implemented, obtaining the 83.33% of pTB cases. This means that can be detect the pTB using the trained map in addition with the *k-means* clustering.

Post-test analysis presented better results, which can be explained by the information that is higher than pre-test case. For this analysis, the risk groups were labeled in an easy way. The high risk group just has activations from pTB patients as shows the figure 2, and the low risk group has a high population of patients without the disease. In the same way as pre-test analysis, it is possible to calculate the pTB cases proportion, combining the number of activations of groups of high and medium risk, obtaining 84.38%. This computation is made because is important to medical staff in pre-diagnostic tasks, due to patients in this combined group can be sent to treatment or specific tests. Proportion to non pTB cases can be calculated using the number of activations of patients without pTB in the low risk group, obtaining a 97.44%.

It is possible to see that the results in pre-diagnostic applications offer a difference of 1.05 % for the two approaches, showing that just pre-test information can be useful. This could be explaining by the use of variables most important in the detection task, characteristics that have been studied in previously works [9][13],where the anamnesis variables are enough to detect pTB with results close to 90% in sensitivity.

Other difference is observable for the non pTB cases because the post-test analysis improved impressively the results. This was obtained using information from classical test, where these patients are located in the low risk group. At same time for this analysis, the clustering is clearer as can be shown in the map (Fig. 2).

V. CONCLUSIONS

SOM neural networks have shown good characteristics of clustering in the pTB diagnostic problem. In this paper is proposed a methodology using a SOM neural network and a *k-means* algorithm, where patients are placed into three risk groups that can be used in a pre-diagnosis. Thus patients classified in the high risk group can be referred for further analysis for confirmation of the disease and its subsequent treatment start.

Two approaches with different number of variables were implemented. In a first analysis the use of anamnesis variables can obtain a proportion of 83.33% for patients with pTB. In this case, the system remains a useful tool for medical staff in a pre-diagnosis.

In a post-test analysis, where variables including more information about classical results test data present a clear clustering for three risk groups. In this case, the pTB patients proportion reached was 84.38%, providing similar advantages that pre-test analysis. For the non pTB cases the results are better when information from post-test analysis is inserted in the clustering system.

ACKNOWLEDGMENT

This work was supported under grant: PI/UAN-2013-562GB from Universidad Antonio Nariño, Colombia and CNPq, FAPERJ and CAPES, Brazil. Authors want to thank theUniversidade Federal do Rio de Janeiro and Hospital Geral de Santa Clara da Misericórdia and COPPE/UFRJ by their support to this work.

REFERENCES

[1] World Health Organization WHO, "Global tuberculosis report 2012", ISBN 978 92 4 156450 2

[2] Weekly Epidemiologic Report, Subdirección de Vigilancia y Control en Salud Pública, Instituto Nacional de Salud. No. 33, Agosto 2012. – *In Spanish*

[3] Hijjar, M., procopio, M., freitas, L., et al. "Epidemiologia da tuberculose: importância no mundo, no Brasil e no Rio de Janeiro". In: Pulmao, RJ, pp. 310–314, 2005. - *In Portuguese*

[4] Piller RVB, "Epidemiologia da Tuberculose", Pulmao RJ, 2012; 21(1), pp. 4 – 9. - *In Portuguese*

[5] A. Trajman, C. Kaisermann, R. R. Luiz, R. D. Sperhacke, M. L. Rossetti, M. H. F. Saad, I. G. Sardella, N. Spector and A. L. Kritski. "Pleural fluid ADA, IgA-ELISA and PCR sensitivities for the diagnosis of pleural tuberculosis".

[6] Llaca Diaz J. M., Flores Aréchiga A., Martínez Guerra M. G., Cnatú Martínez P. C., 'The Smear and Culture in Diagnosis of Extrapulmonary Tuberculosis', Revista Salud Pública y Nutrición, Vol 4 No. 3 July-September 2003, Nuevo León, México.*In Spanish.*

[7] Er O., Termutas T., Tanrikulu a. C., "Tuberculosis Disease Diagnosis Using Artificial Neural Networks". In: Journal of Medical Systems, Vol. 34, pp. 299-302, Junho 2010.

[8] Elveren E., Yumusak N., "Tuberculosis Disease Diagnosis Using Artificial Neural Network Trained with Genetic Algorithm", In: Journal of Medical Systems, Vol 35, pp. 329-332, 2011.

[9] J. Faria, J. Seixas, J. Souza Filho, A. Orjuela, A. Vieira, A. Kritski, I. Silva, A. Trajman. "Pleural Tuberculosis Diagnosis Based on Artificial Neural Networks Models", X Congreso Brasileiro de Inteligencia Computacional CBIC 2011, November 2011, Fortaleza, Ceará – Brazil.

[10] J. M. Seixas, J. Faria, J. B. O. Souza Filho, A. F. M. Vieira, A. Kritski, A. Trajman, "Artificial neural network models to support the diagnosis of pleural tuberculosis in adult patients", *In*: International Journal of Tuberculosis and Lung Diseases, Vol. 17, No. 5, pp. 682-686, 2013.

[11] Baptista de Oiveira e Souza Filho J., Silva Antunes P. H., Seixas J., Maidantchik C., "Neural Networks Applied to Paucibacilar Pulmonary Tuberculosis Diagnosis", *AutomaticBrazilianCongress,* 2010. - *In Portuguese*

[12] Cascao L. V. "Models of Computational Intelligence for Aid in the Clinical Pulmonary Tuberculosis Diagnosis", Master dissertation, Universidade Federal do Rio de Janeiro, October 2011 – *In Portuguese*

[13] Orjuela-Cañón A. D., Seixas J.M., "Fuzzy-ART Neural Networks for Triage in Pleural Tuberculosis", In: Proceedings of V Biomedical Engineering and Bioengineering Colombian Congress, 2013.

[14] Kohonen T., Self Organizing Maps, Springer, 2000.

[15] Faussete L., Fundamentals of Neural networks: architectures, algorithms, and applications.Third Edition, Pearsons Education Publishers.

[16] Haykin S., Neural Networks: A Comprehensive Foundation, Prentice Hall, Second Edition, 1999.

[17] Agresti, A. An Introduction to Categorical Data Analysis. Wiley, 2007.

[18] Kanungo T., Mount D. M., Netanyahu N., Piatko C., Silverman R., Wu A., "An Efficient *k*-Means Clustering Algortihm: Analysis and Implementation", In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002.