X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

PLEURAL TUBERCULOSIS DIAGNOSIS BASED ON ARTIFICIAL NEURAL NETWORKS MODELS

João Felipe Faria[1], José Manoel de Seixas[1], João Baptista de Oliveira e Souza Filho[2], Alvaro D. Orjuela[1,3], André Filipe Marcondes Vieira[4], Afrânio Lineu Kritski[5], Isabel Michelly F. Silva[4] and Anete Trajman[4]

[1]Signal Processing Laboratory, Universidade Federal do Rio de Janeiro

[2]Dept. of Electrical Engineering, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca/CEFET-RJ

[3]Universidad Antonio Nariño

[4]Faculty of Medicine, Universidade Gama Filho

[5]Clementino Fraga Filho University Hospital, Universidade Federal do Rio de Janeiro

{joao, seixas}@lps.ufrj.br, jsouza@cefet-rj.br, dorjuela@ieee.org, {atrajman, andrefmarcondes, kritskia}@gmail.com, isabelmichelly@ig

Abstract – The diagnosis of pleural tuberculosis (pTB) is still a challenge, because the *de facto* standard diagnostic test (histopathological examination) and all classical diagnostic tests require thoracentesis, which is an invasive procedure and incurs on increased costs and risks. Innovative approaches which take in account only clinical features are thus highly relevant to provide less invasive diagnostic tools for pTB. This paper proposes the investigation of non-linear models using artificial neural networks to develop a diagnosis support tool based solely on the clinical variables available without the need to perform thoracentesis exam. Models based on multilayer perceptrons achieved 84.7% (95%CI, 82.4% to 87.1%) of diagnostic accuracy versus 82.5% (95%CI, 80.2% to 84.8%) obtained through histopathological examination. Moreover, the sensitivity of the models was again tested, but only on patients that were diagnosed with pTB based only on clinical grounds, as classical testing did not yield positive results; the resulting estimated sensitivity indicates that the models can exploit the underlying clinical features even in pTB cases undetected by classical testing, but the low number of patients in this situation poses a limitation to this analysis.

Keywords – Artificial neural networks, pleural tuberculosis, diagnosis support, nonlinear models.

Resumo – O diagnóstico da tuberculose pleural (pTB) ainda se apresenta como um desafio, pois o teste diagnóstico tido como padrão *de facto* (exame histopatológico) e todos os testes diagnósticos clássicos requerem a realização da toracocentese, procedimento invasivo que incorre em custos e riscos elevados. Abordagens inovadoras que levem em conta somente aspectos clínicos são portanto muito relevantes para prover ferramentas para um diagnóstico menos invasivo da pTB. Este artigo propõe a investigação de modelos não-lineares utilizando redes neurais artificiais para desenvolver uma ferramenta de suporte ao diagnóstico baseada tão somente no processamento de variáveis clínicas disponíveis sem a realização da toracocentese. Perceptrons multicamadas alcançaram 84.7% (95%CI, 82.4% to 87.1%) de acurácia no diagnóstico, versus 82.5% (95%CI, 80.2% to 84.8%) obtidos pelo exame histopatológico. Adicionalmente, a sensibilidade dos modelos foi novamente avaliada, mas somente em pacientes cujo diagnóstico de pTB se deu apenas com base no acompanhamento clínico, não tendo obtido resultado positivo em nenhum dos testes clássicos; a sensibilidade estimada desta maneira indica que os modelos são capazes de extrair características relevantes dos dados clínicos, mesmo em casos em que a pTB não é detectada pelos testes usuais, mas o pequeno número de pacientes nesta situação se apresenta como uma limitação para esta análise.

Palavras-chave – Redes neurais artificiais, tuberculose pleural, suporte ao diagnóstico, modelos não-lineares.

1. INTRODUCTION

The tuberculosis (TB) is an infectious disease considered a world emergency by the World Health Organization, as it has high infection rates worldwide and is the main cause of death in people aged between 15–49 years [1]. The most frequent extrapulmonary site of infection is the pleura, in which case it is denominated pleural tuberculosis (pTB) and has been found to be a major cause of pleural effusion in many countries [2–4]. However, accurate pTB diagnosis is still a challenge, because acid-fast bacillus (AFB) smear and *M. tuberculosis* culture in pleural fluid removed through thoracentesis procedure exhibit low sensitivity [5–8] and pleural biopsy is thus frequently required, usually through histopathological examination, but lack of specificity, increased risk of complications and costs added by the biopsy procedure pose limitations to this approach [8]. Newer tests have been proposed to achieve higher accuracy in diagnosing pTB using pleural fluid specimens only [9], but innovative approaches which take into account only clinical features and provide less invasive yet accurate diagnostic tools for pTB are highly relevant, specially in areas with high TB prevalence and poor resources available.

Decision support systems based on computational intelligence may provide a useful tool in pTB diagnosis. In this context, artificial neural networks (ANNs) are a non-linear signal processing tool, frequently used in multidisciplinary problems, due to their universal function approximation properties [10]. They are capable of extracting non-evident but meaningful linear and non-linear relationships buried deep even in high-dimensional data sets with complex interactions between its variables and implement useful mapping functions [10]. In addition, they are valuable tools due to their ability to generalize and yield reasonably good models with relatively few data points [10]. Extensive use of ANNs in medical applications can be found in the existing literature, ranging from pattern recognition in medical images applied to decision support systems, to prediction of treatment outcomes and physiological measurements [11]; earlier works have applied binary classification trees and artificial neural networks in predicting smear negative pulmonary TB [12]. Given the challenges faced in the development of simple, efficient, non-invasive and low-cost tools for pTB diagnosis, this paper proposes ANN prediction models as tools to tackle those challenges. These models may provide an aid tool for non-invasive diagnosis of pTB, avoiding the need for thoracentesis and pleural biopsy procedures.

2. NEURAL NETWORK DESIGN

Neural models used two-layered feedforward multilayer perceptron neural network (MLP) with one output node for a binary pTB diagnosis decision. The optimal size of the hidden layer was defined through a cross-validation approach. Hyperbolic tangent was set as the transfer function for all nodes. The cost function adopted was the mean square error (MSE) and early stop was used as the learning mechanism control to avoid overtraining [10]. Target outputs were set as +1 for a positive pTB diagnosis and -1, otherwise. The database was split into training and test sets, the latter used as the validation set for early stop. Variables with values outside of the range [-1, +1] were normalized so that the biggest and smallest values found in the training set were linearly mapped to +1 and -1, respectively. The parameters obtained to normalize the training set were used to normalize the test set as well; values eventually mapped outside of the range [-1, +1] were squashed to +1 or -1.

In order to estimate the performance of the ANN models, a stratified random subsampling (stratified hold-out) [13] method was used, which consists on repeatedly and randomly split the data set into training and test sets, with both sets exhibiting the same proportion of positively and negatively diagnosed pTB patients as found in the original data set. Each split considered a proportion of 70% of the available events dedicated to the training set and the 30% remaining ones for the test.

Models produced had their performance measured regarding sensitivity, specificity, accuracy and the sum-product (SP) index value. Performance metrics based on the SP index are suitable to select classifiers which show balanced values of sensitivity and specificity, as a significant drop in either of them forces a dramatic drop in the index [14]. The expressions for calculation of the different performance indexes are summarized in Table 1.

The measure of the proposed model's classification performance considered the mean performance attained over all runs (100 in this work). In order to lower the model uncertainty associated with weight and bias initialization, in each split the ANN was randomly initialized and trained several times (100 in this work) and the best performing model (with regard to SP index) was taken as the sample performance of that split. The complete procedure is outlined in Algorithm 1.

Training was made in batch mode [15] and the performances using four different training algorithms were compared: Resilient Backpropagation [16] (RPROP), Levenberg-Manquardt [17] (LM), BFGS quasi-Newton [18] and One Step Secant [19] (OSS). The classifier was completed by feeding the output of the ANN to a hard decision device, which maps non-negative output values into positive pTB diagnosis and negative pTB diagnosis otherwise.

Table 1: Calculation of the different performance indexes used.

Sensitivity (S)	Specificity (E)	Accuracy (A)	Sum-product (SP)
$S = \frac{TP}{TP+FN}$	$E = \frac{TN}{TN+FP}$	$A = \frac{TP+TN}{TP+FN+TN+FP}$	$SP = \sqrt{\left(\frac{S+E}{2}\right) . \sqrt{S.E}}$

Where TP, TN, FP and FN stand respectively for the number of true-positive, true-negative, false-positive and false-negative classification outcomes.

Algorithm 1 Estimating network performance via a stratified random subsampling method.

```
for n := 1 ... 20:
        repeat 100 times:
                draw randomly 30% of positively diagnosed pTB patients to
                        to form the test set;
                draw randomly 30% of negatively diagnosed pTB patients to
                        to form the test set;
                form training set with remaining patients;
                normalize variables;
                repeat 100 times:
                        randomly initialize network with n nodes in the
                                hidden layer;
                        train network with the training set;
                        simulate network on the test set;
                end;
                select the best performing network;
        end;
        store all performances, networks and train/test partitions;
end;
```

3. CONFIDENCE INTERVAL ESTIMATION

Performances found in study results in medical research are often reported along with *p*-values to assess the statistical significance of the results, but in the past two decades the reporting of confidence intervals in addition to or instead of *p*-values is widely used and has been increasingly adopted as a guideline by the main medical journals [20]. Briefly, confidence intervals are estimates of the precision of the population parameters' estimates inferred from a limited sample [20]. A 95% confidence interval is thus defined as the interval which will contain the *true* population parameter for 95% of the samples randomly drawn from the population [21]. The actual percentage value used is arbitrarily set according to how strict one wants to be when considering the statistical significance of findings; common values usually found in literature are 99%, 95% and 90%. [20]

As models in this work are evaluated in terms of the mean value of performance indexes obtained for different data splits, confidence intervals can provide information about the precision of the performances' estimates.

In order to construct the confidence interval, a conservative approach based on Chebyshev's inequality [22], i.e.:

$$P\left(|x - \bar{x}| \geq \delta\right) \leq \frac{\sigma_x^2}{\delta^2} \tag{1}$$

One should note that no assumption over the samples' distributions is considered.

The 95% confidence interval CI is thus calculated as follows:

$$CI = (\hat{\mu} - \delta, \hat{\mu} + \delta) \tag{2}$$

$$\delta = \sqrt{\frac{\hat{\sigma}_M^2}{0.05}}, \hat{\sigma}_M = \frac{\hat{\sigma}}{\sqrt{N}} \tag{3}$$

where $\hat{\mu}$ is the mean value of the performance index values attained across all runs, $\hat{\sigma}$ is the standard deviation of the sample's performance index values and N is the number of runs.

Table 2: Best performance obtained with each training algorithm.

Algorithm	Mean SP-index	95% CI	Mean Accuracy	95% CI	No. of hidden layer neurons
Levenberg-Manquardt	83.8%	81.8% to 85.9%	84.7%	82.4% to 87.1%	6
BFGS quasi-Newton	82.9%	80.8% to 84.9%	84.2%	82.3% to 86.2%	20
One Step Secant	81.9%	79.7% to 84.0%	83.9%	81.9% to 85.9%	20
Resilient Backpropagation	81.4%	79.3% to 83.5%	84.2%	82.3% to 86.1%	20

4. DATABASE

The database used in this work was also used in previous works by Trajman et al [9], which assessed novel pleural fluid tests in the diagnosis of pTB. 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. Each patient's record was consulted to gather information on 12 variables derived from different sources: anamnesis (age, gender, smoking status, HIV status), classical test results (histopathologic, pleural tissue culture, pleural fluid culture, AFB smear), experimental test results (adenosine deaminase -ADA, IgA-ELISA and nucleic acid sequences by amplification tests -NAAT- using in house polymerase chain reaction -PCR) and the final diagnosis of the patient [9]. 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 for age, whose numerical value (in years) was kept, and gender, which was coded as +1 for male and -1 for female patients.

The database contains records of 137 patients, two of which were excluded from this study because of their records' lack of final diagnosis information. Of the remaining 135 patients, 96 have been diagnosed with pTB and 22 of them were so based solely on clinical grounds, i.e. none of the tests yielded a positive result. Only the anamnesis variables and HIV testing results were used as inputs of the network, as they are the only ones that would be available in case thoracentesis had not been performed.

5. RESULTS

The effect of the number of neurons in the hidden layer on the mean value of the SP-index can be seen in Figure 1. The confidence intervals estimated as detailed in Section 3 are represented by the error bars. A slight increasing tendency can be observed in the performance number of hidden neurons grows. The performance of the ANN models derived through different training algorithms was similar as well, with the best overall performance attained by LM algorithm. The best mean SP-index obtained through each training algorithm are summarized in Table 2.

The performance of the classical tests was estimated in a similar fashion as those of the models, i.e. their accuracy, sensibility, specificity and SP-index were evaluated 100 test sets randomly generated containing 30% of the patients. The mean performances attained by the classical tests, accompanied by the corresponding confidence intervals, can be found in Figure 2, along with the results for the best performances reached with the networks using different training algorithms (see Table 2).

Model uncertainty associated with weight and bias initialization was also estimated. For this analysis, only models with 6 hidden neurons and trained with LM algorithm were considered. Performance estimation in the same fashion as before and, for each of the 100 splits, two models (out of 100 generated) were picked as the sample performance of that split. One was picked at random, so the performance estimation will take in account the uncertainty introduced by initialization, and the other was picked by choosing the best performing model according to the SP-index. The mean performance and confidence intervals were estimated for accuracy, sensibility, specificity and SP-index. These results are plotted in Figure 3.

To further investigate the performance of the models, we analyzed the output of each of the networks trained (except for the ones discarded to lessen the performance fluctuation due to initialization – see Section 2) when presented to each one of the 22 patients whose diagnostics was based solely on clinical grounds, i.e. those for which none of the classical diagnostic tests yielded a positive result but a continuous medical accompaniment led to the conclusion that the patient indeed suffered from pleural tuberculosis. For each model derived through a training training algorithm, the 22 patients' data was fed to the 100 networks generated and the overall diagnosis accuracy obtained by the models for this group of patients was determined, expressed in terms of mean value and its associated confidence interval. To check for generalization, the diagnosis accuracy obtained by the models for the patients in this group that were not used to train that network, i.e. those that were drawn to form the test set of that particular iteration. Across all splits, of those 22 patients, at most 11 and at least 2 were drawn to the test set, while the median number was 7. The results can be seen in Figure 4. Note that, in this scenario, the accuracy is equal to the sensitivity, as only patients with positive pTB diagnosis are taken in account.

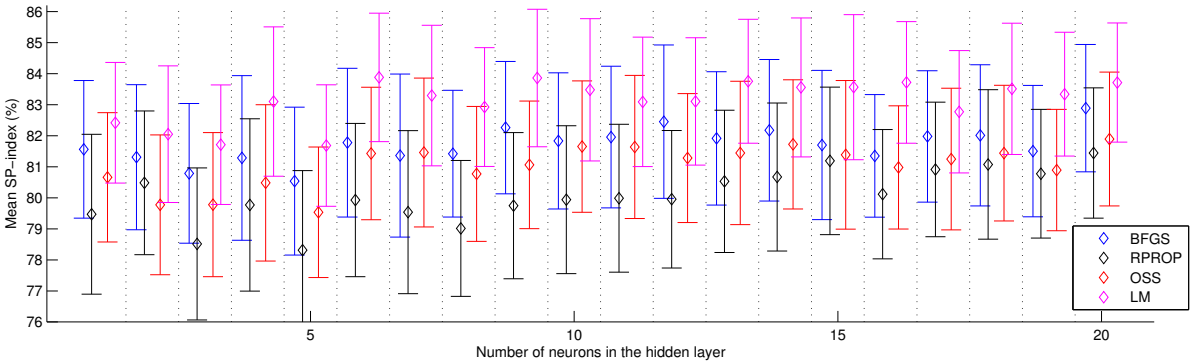

Figure 1: The effect of the number of neurons in the network's hidden layer in the mean SP-index value. Error bars reflect the estimated confidence intervals.

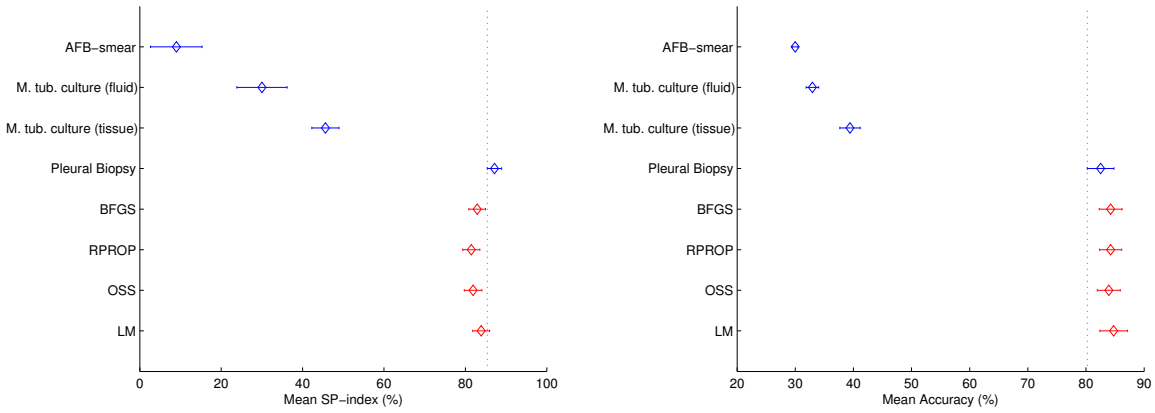

Figure 2: Comparison of two of the classical tests and the best performing network topologies obtained with different training algorithms.

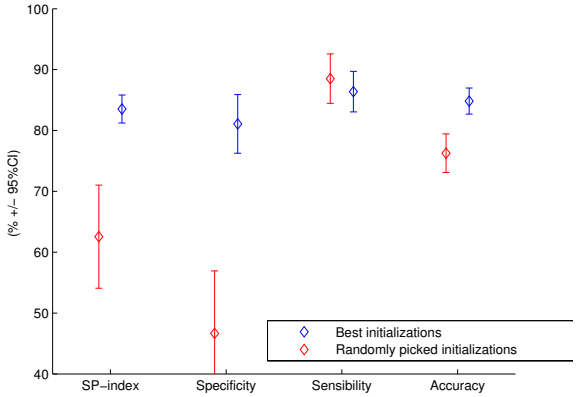

Figure 3: Effects of random initialization on estimated performance. Comparison of estimation results with and without compensation for fluctuations associated with the initialization of weights and bias.

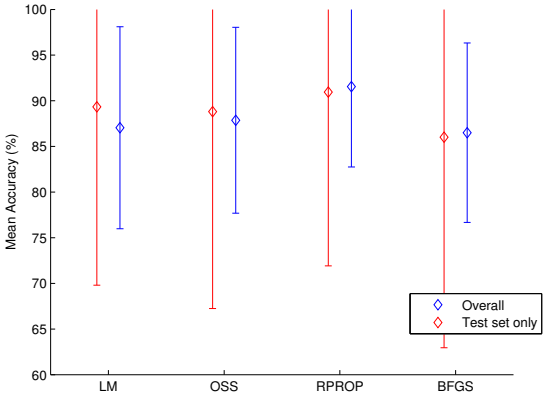

Figure 4: Accuracy (sensitivity) of the networks for diagnosing pleural tuberculosis undetected by classical diagnostic tests.

6. Conclusions

Pleural biopsy through histopathological examination is currently the best available diagnostic tool for pleural tuberculosis [9], but it is a highly invasive procedure which incurs in increased risks and costs. Diagnostic tests using only the pleural fluid collected via the somewhat less invasive procedure of thoracentesis are available but exhibit several shortcomings [5], and thus the development of novel diagnostic tools is highly relevant, specially to assist doctors practicing in low resource settings in areas with high rates of tuberculosis infection. This paper evaluated the potentials and challenges involved in the development of artificial neural network models for diagnosis support of pleural tuberculosis with the main aim of greatly diminishing the need for invasive procedures such as the aforementioned thoracentesis and pleural biopsy.

The results obtained show that the ANN models derived using only simple and readily available clinical information (age, gender, smoking status, and HIV status) were able to correctly identify patients in accordance with their final pTB diagnosis in most of the cases, attaining performances similar to that of histopathological examination, the *de facto* standard for pTB diagnosis [9], and definitely superior to other classical tests in which a positive result grants a positive pTB diagnosis (AFB-smear and *M. tuberculosis* culture in pleural tissue or fluid). It should be noted that the results, albeit promising, were obtained using a database with a small number of control patients not diagnosed with pTB (around 30%). This fact poses one of the major limitations of this study, as it can affect the assessment of specificity [5] and, in turn, of the performance indexes used (SP-index and accuracy). These results do however serve the purpose of evaluating the potentials and challenges of such an approach for the support of pTB diagnosis. Analysis of Figure 3 confirms that an accurate assessment of specificity is more difficult to be made and that the model learning suffers considerably from the lack of control patients. A proper selection of bias and weight initialization may, however, compensate this fact, leading to a model which correctly identifies control patients almost as well as pTB patients, as confirmed by the small difference between accuracy and SP-index attained by the models.

An interesting finding is that good mean accuracies were obtained even for patients whose tuberculosis went undetected through classical diagnostic testing. The small number of patients in this category (22) is however a major limitation in this analysis, and specially when testing for generalization, as the number of such patients in the test set is even lower for any given split. Nevertheless, the estimated accuracies in both cases are close, indicating good generalization, and they are close to the accuracies obtained earlier for the entire test sets as well, which indicates that the underlying clinical characteristics are sufficient to predict pleural tuberculosis diagnosis, a very useful result considering the reported lack of sensitivity in most classical diagnostic tests [5–8].

Future research may include refinement of the presented models, the use of a larger database currently under construction featuring more patients and more clinical variables, and the use of other techniques, e.g. clustering tools and component analysis, as the development of risk profile analysis tools.

References

[1] P. Dolin, M. Raviglione and A. Kochi. "Global tuberculosis incidence and mortality during 1990–2000". *Bulletin of the World Health Organization*, vol. 72, pp. 213–220, 1994.

[2] C. Liam, K. Lim and C. Wong. "Causes of pleural exudates in a region with a high incidence of tuberculosis". *Respirology*, , no. 5, pp. 33–38, 2000.

[3] L. Selig, A. Cunha, E. Teixeira, M. Belo, M. Branco and A. Trajman. "Anti-HIV testing for tuberculosis patients in Rio de Janeiro State". *Pulmão RJ*, , no. 10, pp. 8–13, 2001.

[4] L. Valdes, A. Pose, E. San Jose and J. Martinez Vazquez. "Tuberculous pleural effusions". *Eur J Intern Med*, , no. 14, pp. 77–88, 2003.

[5] A. Trajman, M. Pai, K. Dheda, R. van Zyl Smit, A. Zwerling, R. Joshi, S. Kalantri, P. Daley and D. Menzies. "Novel tests for diagnosing tuberculous pleural effusion: what works and what does not?" *Eur Respir J*, , no. 31, pp. 1098–1106, 2008.

[6] A. Seibert, J. Haynes, R. Middleton and J. Bass, J.B. "Tuberculous pleural effusion: Twenty-year experience". *Chest*, , no. 99, pp. 883–886, 1991.

[7] E. Bueno, G. Clemente and C. e. a. Castro. "Cytologic and bacteriologic analysis of fluid and pleural biopsy specimens with Cope's needle: Study of 414 patients". *Arch Intern Med*, , no. 150, pp. 1190–1194, 1990.

[8] R. Light. "Useful tests on the pleural fluid in the management of patients with pleural effusions". *Curr Opin Pulm Med*, , no. 5, pp. 245–249, 1999.

[9] 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". *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 8, no. 67, pp. 877–884, 2007.

[10] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, second edition, 1998.

[11] P. Lisboa. "A review of evidence of health benefit from artificial neural networks in medical intervention". *Neural Networks*, vol. 15, no. 1, pp. 11 – 39, 2002.

[12] A. Santos, B. Pereira, J. Seixas, F. Mello and A. Kritski. *Neural Networks: An Application for Predicting Smear Negative Pulmonary Tuberculosis*, pp. 279–289. A Birkhauser book, 2007.

[13] R. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1137–1143. C.S. Mellish, 1995.

[14] A. R. dos Anjos, R. C. Torres, J. M. Seixas, B. C. Ferreira and T. Xavier. "Neural triggering system operating on high resolution calorimetry information". *Nuclear Instruments and Methods in Physics Research*, pp. 134–138, 2006.

[15] R. Duda, P. Hart and D. Stork. *Pattern Classification*. Wiley, 2000.

[16] M. Riedmiller. "Rprop – Description and Implementation Details". Technical report, University of Karlsruhe, 1994.

[17] M. Hagan and M. Menhaj. "Training feed-forward networks with the Marquardt algorithm". *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[18] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[19] R. Battiti. "First and second order methods for learning: Between steepest descent and Newton's method". *Neural Computation*, vol. 4, no. 2, pp. 141–166, 1992.

[20] D. Altman, T. Bryant, M. Gardner and D. Machin. *Statistics with ConfidenceÑConfidence Intervals and Statistical Guidelines*. BMJ Books, London, 2000.

[21] R. Newcombe. "Statistical Applications in Orthodontics. Part I. Confidence Intervals: an Introduction". *J. Orthod.*, vol. 27, no. 3, pp. 270–272, 2000.

[22] P. Z. Peebles. *Probability, Random Variables, And Random Signal Principles*. McGraw-Hill, New York, 2001.