OPEN ACCESS Freely available online

Journal of Infectious Diseases & Preventive Medicine

Research Article

Application of Data Mining Techniques in Tuberculosis (TB) Diagnosis: A Comparison of Multilayer Perceptron Neural Network (MLP) and Adaptive Neuro-Fuzzy Inference System (ANFIS) Efficiency

Azamossadat Hosseini*, **Hamid Moghaddasi, Reza Rabiei, Sara Mohebi Mushaei**

Department of Health Information Technology and Management, Shahid Beheshti University of Medical Sciences, Tehran, Iran

ABSTRACT

Background: Data mining techniques for disease diagnosis help the prediction and control of various diseases, including Tuberculosis (TB). This study aimed to compare the efficiency of two main models of TB diagnosis: MLP (Multilayer Perceptron Neural Network) and ANFIS (Adaptive Neuro-Fuzzy Inference System) to find out which data-mining-based model is more efficient in detecting tuberculosis.

Materials and methods: In this analytical study, database used was for inpatients in a specialized hospital for lung and respiratory diseases. The database included 1159 records, of which 599 records belonged to TB infected patients and 560 records to non-infected patients. With help of 13 factors effective on diagnosis of the disease and using the set of TB records, the two models of MLP and ANFIS were tested and evaluated. Finally, using the ratio test, two models were compared based on their AUC values to see which one is more efficient. The sensitivity, specificity, accuracy, and RMSE of the two models were also compared.

Results: The efficiency of MLP was 0.9921 and the efficiency of ANFIS was 0.8572. MLP's sensitivity, specificity, accuracy, and RMSE were recorded as 93.50%, 94.80%, 94.30%, and 0.1788, respectively. These values for ANFIS equaled 79.60%, 92.60%, 85.63%, and 0.3345, respectively. According to these results, there was a significant difference between efficiency levels of MLP and ANFIS models (p-value<0.0001).

Conclusion: The MLP indicated a higher AUC value compared with ANFIS. The results also showed higher sensitivity, specificity, and accuracy but lower RMSE for MLP. Overall, MLP proved superior to ANFIS for TB diagnosis.

Keywords: Tuberculosis; Modeling of TB diagnosis; Multilayer perceptron; Neural network; Adaptive neuro-fuzzy inference system; Efficiency level

INTRODUCTION

Tuberculosis (TB) is a reemerging infectious disease worldwide [1,2], caused by a bacterium called *Mycobacterium tuberculosis*. Claiming some 1.7 million lives in 2016, the disease was introduced as the world's most infectious killer [3,4]. Due to the lack of adequate diagnostic and treatment facilities, TB has turned into a serious challenge in developing countries [5,6]. Various methods have been yet introduced for TB diagnosis, of which the culture method-culturing *Mycobacterium tuberculosis* organisms-is a major one. Although this approach is considered the standard method of TB diagnosis, it needs high-tech lab instruments and takes a long time, 8-12 weeks, to receive the test results [3,7-11] . The tuberculin skin test, chest X-ray, and sputum smear microscopy are other common methods for TB diagnosis [1,12], with the latter being the most common [9], although a third of TB-infected people, mostly children, are unable to produce sputum [13]. Therefore, a key strategy set by World Health Organization (WHO) to put an end to TB is the use of rapid diagnostic methods rather than those dependent on sputum [3]. Over the last decade, quick molecular diagnostic methods have also been developed, although they are challenged by the absence of required medical equipment in medical centers [1].

Despite the progress in the treatment of TB, the early diagnosis of this infectious disease is still of great importance in increasing the survival rate of patients [9]. TB diagnosis is still challenging since all the existing methods are aggressive approaches associated with high costs and risks [14].

The combination of biology, artificial intelligence, and mathematical

Correspondence to: Azamossadat Hosseini, Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran, E-mail: hosseinia@gmail.com

Received: December 07, 2021; **Accepted:** December 21, 2021; **Published:** December 28, 2021

Citation: Hosseini A, Moghaddasi H, Rabiei R, Mushaei SM (2021) Application of Data Mining Techniques in Tuberculosis (TB) Diagnosis: A Comparison of Multilayer Perceptron Neural Network (MLP) and Adaptive Neuro-Fuzzy Inference System (ANFIS) Efficiency. J Infect Dis Prev Med. 9: 247.

models has led to rapid progress in medicine, helping physicians to better identify diseases [15]. There have been a number of studies conducted on TB diagnosis based on data mining techniques and artificial intelligence [12,16]. Data mining refers to finding and collecting interesting, unexpected, and valuable patterns within large data sets [17]. The techniques are applied in numerous fields, including medicine, marketing, and banking [18]. As for medicine, the data mining techniques have proved highly effective in the early identification of diseases, providing proper treatments at reasonable costs and designing healthcare decision support systems [19,20]. The techniques can also be very helpful in medical discoveries [7]. Classification is a part of the data mining process usually utilized for analyzing medical data [21] as well as making medical decisions [22]. To obtain the highest possible accuracy in TB diagnosis, several researchers have applied various data mining techniques, including artificial neural networks (ANNs), rough sets, adaptive neuro-fuzzy inference system (ANFIS), decision tree, naïve Bayes, support vector machine (SVM), and association rules [5,6,9,11,12,23-27]. As the multilayer perceptron neural network (MLP) and ANFIS are among the most efficient methods of TB diagnosis [12,14,19,24-27], the present study aimed to compare the efficiency of MLP and ANFIS in TB diagnosis.

Benfu et al. conducted a study on the use of ANN in the diagnosis of smear-negative pulmonary TB. This study focused on 560 records belonging to 272 infected people and 288 healthy individuals. Each record consisted of 29 fields containing various data, e.g. demographic information, clinical symptoms, chest X-ray, and laboratory test results. The ANN in this study consisted of a hidden layer of nine neurons. The area under the receiver operating characteristic (ROC) curve, accuracy, sensitivity, and specificity of the created model were 0.989, 93.10%, 88.89%, and 100%, respectively [24].

Ucar and Karahoca adopted a data mining technique in their study for predicting the existence of *mycobacterium tuberculosis* in patients to detect TB by ANFIS, MLP, and PART techniques. The study contained 667 patients along with 30 associated parameters reduced to 20 ones. Upon a comparison of the three models, ANFIS proved superior, whereas the root mean square error (RMSE) for the three algorithms was 18%, 19%, and 20% for ANFIS, MLP and PART, respectively. This study concluded that ANFIS was a more reliable and efficient technique compared to MLP and PART [27].

Faria et al. studied pleural tuberculosis diagnosis applying ANN models on a sample of 135 patients including 12 relevant parameters. The outcomes of an MLP-based TB diagnosis model demonstrated an accuracy rate of 84.7% [14]. Moreover, Filho et al. examined a screening system for smear-negative pulmonary TB via ANNs involving 136 adult patients with smear-negative pulmonary TB between January 2010 to December 2011. The patient records contained 12 parameters of age, cough, bloody sputum, night sweats, fever, weight loss over 10%, breathlessness, lack of appetite, smoking, non-pulmonary TB, the history of hospitalization, and HIV infection. The study compared the CART, MLR, MLP and support vector machine (SVM) models. The results of analyzing the sensitivity, specificity, accuracy, and area under the curve (AUC) values were in favor of MLP with 100%, 80%, 88%, and 0.918, respectively [26].

In previous studies, MLP and ANFIS demonstrated acceptable results in detecting TB. The present study aimed to compare the efficiency of these techniques using a database bigger than those applied in other studies and with some different input parameters.

MATERIALS AND METHODS

Initially, in order to identify the risk factors and main symptoms of TB, the literature was reviewed examined and experts' opinions were then sought about the factors identified. Accordingly, features such as fever, weight loss, night sweats, coughing, bloody sputum, fatigue, lack of appetite, smoking, diabetes, HIV infection, contact with TB-infected patients, alcohol consumption, age, taking immunosuppressant drugs, sex, white blood cell (WBC) count and hemoglobin (Hb) rate were extracted as the most significant risk factors and symptoms (Table 1). In addition, the erythrocyte sedimentation rate (ESR) was also considered as the most important blood test for detecting TB infection.

Table 1: Examining input variables.

No.	Variable	Type of variable	Correlation
1	Fever	Qualitative	0.4416
2	Weight loss	Qualitative	0.3814
3	Night sweats	Qualitative	0.361
4	ESR	Qualitative	0.359
5	Coughing	Qualitative	0.3456
6	Smoking	Qualitative	0.2514
7	Age	Qualitative	0.146
8	Blood sputum	Qualitative	-0.0221
9	HIV	Quantitative	-0.0294
10	WBC	Qualitative	-0.0575
11	Contact with TB-infected patients	Quantitative	-0.0729
12	Sex	Quantitative	-0.1252
13	Hemoglobin	Quantitative	-0.1947

In this study, a database containing 1217 records was used. The database belonged to a clinical setting in MaseehDaneshvari Hospital, Tehran, and included 630 records of TB-infected patients and 587 records of non-infected patients. Each record consisted of 17 fields including risk factors, symptoms, and test results. These 17 features were: fever, weight loss, night sweat, coughing, bloody sputum, smoking, AIDS, contact with a TB-infected patient, age, sex, occupation, number of years of smoking, presence of underlying diseases, white blood cell (WBC) count, fasting blood sugar (FBS), hemoglobin (Hb), and ESR.

In the preprocessing stage, four features, including job, number of years of smoking, underlying diseases, and fasting blood sugar level were deleted due to the large number of null data, i.e. >50% of the information about these characteristics in the existing database lacked information. In order to determine the effectiveness of the remaining 13 features in the diagnosis of the patient group and non-infected group, a correlation function was applied to the data, the results of which are shown in Table 1. In the second step, the records with 50% of the fields blank were deleted. Therefore, 1159 records remained, of which 599 belonged to TB-infected patients and 560 to non-TB-infected individuals.

In the next step, using the preprocessed data, the perceptron neural network was created by MATLAB R2015b with one and two hidden layers. Based on the database considered for the test (30% of the total data), the accuracy, sensitivity, and specificity metrics as well as the RMSC and AUC (ROC) values were calculated for modeling. The ANFIS model was also created by MATLAB R2015b. Based on the specified test database (30% of the total data), the same metrics and values were calculated for the model.

Finally, the accuracy, sensitivity, and specificity metrics, as well as RMSE were determined. In addition, to find out whether there was a significant difference between the efficiency of the two models, p-value was examined at α =0.05. In addition, the AUC for both models was calculated using MedCalc.

RESULTS

To find the number of hidden layers and neurons in each layer, various structures of the MLP were examined. The neural network with one hidden layer was examined for 95 modes, with the number of neurons varying from 6 to 100, and the neural network with two hidden layers was examined for 205 modes. Twenty of modes with the best results are presented in Table 1.

As seen in Table 2, the best results are assigned to MLP with one hidden layer and 42 neurons in the layer. Figure 1 displays the ROC for the best mode of the MLP.

Table 2: Examining various structures of the MLP.

No. of hidden layer(s)	No. of neurons in hidden layer(s)	Sensitivity	Specificity	Accuracy	RMSE	AUC
1	29	90.29%	91.28%	90.80%	0.2536	0.9723
	37	90.30%	85.70%	88.10%	0.2799	0.9624
	42	93.50%	94.80%	94.30%	0.1788	0.9921
	49	91.50%	88.40%	90.00%	0.2752	0.9624
	54	92.00%	92.00%	92.00%	0.2405	0.9769
	62	92.50%	91.89%	91.80%	0.25	0.9743
	68	90.20%	92.30%	91.20%	0.2505	0.9738
	71	91.30%	89.80%	90.60%	0.2565	0.9714
	86	91.50%	90.70%	91.10%	0.2513	0.9726
	91	89.00%	90.20%	89.60%	0.2638	0.9692
2	11-9	89.00%	90.90%	89.90%	0.2785	0.9605
	10-10	91.70%	90.50%	91.10%	0.2596	0.9702
	13-15	89.60%	92.10%	90.90%	0.2497	0.9749
	14-16	91.70%	92.50%	91.10%	0.2585	0.9713
	11-17	91.00%	91.40%	91.20%	0.2507	0.9733
	10-19	92.70%	90.40%	91.50%	0.2513	0.972
	25-35	91.00%	88.80%	89.90%	0.2623	0.9702
	30-45	92.20%	90.70%	91.50%	0.2458	0.976
	50-45	91.20%	88.60%	89.90%	0.2556	0.9718
	40-50	93.70%	93.80%	93.70%	0.2306	0.9811

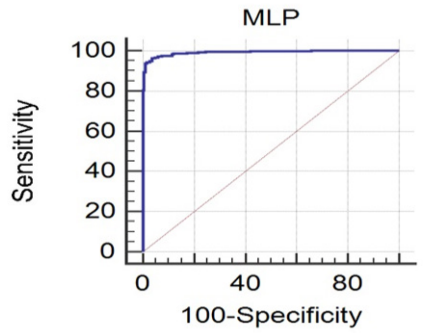

Figure 1: The ROC for the MLP.

We also employed MATLAB R2015b to create an ANFIS trained model and determine the sensitivity, specificity, accuracy, and RSME metrics. To find the best results of the model, the ANFIS was run for different iterations some of them with the best results are recorded in Table 3.

Table 3: Results related to the implementation of the ANFIS Model.

AUC	RMSE	Accuracy	Specificity	Sensitivity	Iteration
0.76	0.3744	77.48%	81.00%	72.26%	200
0.8106	0.371	81.03%	80.65%	81.22%	300
0.816	0.3647	81.61%	81.72%	80.12%	500
0.8201	0.3612	81.93%	81.72%	81.48%	800
0.8253	0.3443	82.20%	86.02%	89.41%	1000
0.8572	0.3345	85.63%	92.60%	79.60%	1500
0.8511	0.3373	85.08%	90.98%	81.03%	2000
0.8448	0.3416	83.86%	90.13%	79.88%	2500

According to Table 3, the best result of the ANFIS model belonged to 1500 iterations. The values of each indicator in this case are given in Table 4.

Table 4: The best mode of ANFIS model.

Sensitivity	Specificity	Accuracy	RMSE	AUC
79.60%	92.60%	85.63%	0.3345	0.8572

To compare MLP and ANFIS models, the accuracy, sensitivity, specificity, and RMSE values were examined at the significance level of α=0.05, with the results presented in Table 5.

According to Table 5, there are significant differences between the sensitivity, accuracy, and RMSE of the two models, but there is no significant difference between their specificity values.

Table 5: Accuracy, sensitivity, specificity, and RMSE Values of the two models.

Metrics	P-value	Notes
Sensitivity	0.0002	Significant difference
Specificity	0.5287	No significant difference
Accuracy	0.0114	Significant difference
RMSE	0.0013	Significant difference

To compare the efficiency of the two models, the area under the ROC curve of the MLP and that of the ANFIS model were calculated by MedCalc at α=0.05 and the results are presented in Figures 2 and 3 and Table 6.

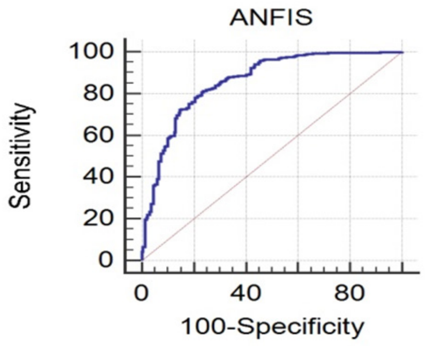

Figure 2: Flow ROC for the ANFIS.

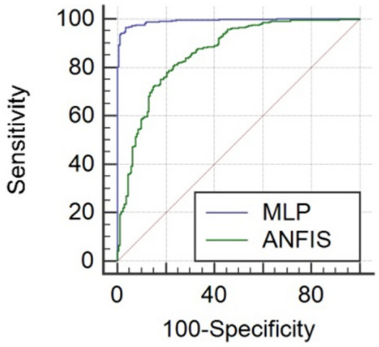

Figure 3: Flow The AUCs of MLP and ANFIS models.

As Table 5 illustrates, the p-value is <0.0001, denoting a significant difference in the efficiency of the two models; denoting the rejection of the null hypothesis.

Table 6: AUCs of MLP and ANFIS Models.

AUC			
Variable 1	ANFIS		
Variable 2	MLP		
Classification variable	diagnosis		
Sample size	1159		
Positive group:diagnosis= 1	599		
Negative group:diagnosis=0	560		
	AUC	Sea	95% Clb
ANFIS	0.86	0.0109	0.839 to 0.880
MLP	0.992	0.00192	0.985 to 0.996
Pairwise comparison of ROC curves			
ANFIS˜MLP			
Difference between areas	0.132		
Standard Error"	0.0106		
95% Confidence Interval	0.111 to 0.153		
z statistic	12.477		
Significance level	P<0.0001		

[a]DeLong et al.. 1988; [b]Binomial exact

DISCUSSION

Although there are several methods for TB diagnosis, each method is associated with some pitfalls. Therefore, there are concerns about controlling of this disease and delays in its diagnosis and treatment. Beside the common methods, data mining techniques could help on-time diagnosis of TB. Classification is a part of the data mining approach usually meant to analyze the medical data [21]. The studies addressing the application of data mining techniques considered the MLP and ANFIS among the frequently used techniques for TB diagnosis.

In this study, to determine the number of hidden layers and the neurons in each layer for achieving the best possible MLP model, all the possible modes for one layer with 6 to 100 neurons, and two layers with 6 to 50 neurons, were examined. The mode with one hidden layer and 42 neurons in the layer resulted in 93.50%, 94.80%, 94.30% for a sensitivity, a specificity, and an accuracy, respectively. For RMSE and AUC the findings were 0.1788, and 0.9921, in order. With respect to the ANFIS, the mode with 1500 recurrences, a sensitivity of 79.60%, a specificity of 92.60%, an accuracy of 85.63%, RMSE=0.3345, and AUC=0.8572 proved the best possible model.

The result of a study conducted by Benfu et al. regarding the application of artificial neural network in the diagnosis of smear negative pulmonary tuberculosis indicated the accuracy, sensitivity, and specificity of 93.10%, 88.89%, and 100%, respectively. The AUC in this study was 0.989 [24]. The AUC value and the accuracy, as well as sensitivity of the MLP model in the present study were higher than those reported by Benfu et al. study [24]. In another research performed by Ucar and Karahoca the ANFIS model indicated a higher efficiency compare to the MLP and PART models [27]. The difference between the findings of the aforementioned study with those of present study could be due to the different input data and parameters used. The finding of a study by Faria et al. showed that the accuracy of the MLP model in TB diagnosis was 84.7% [14] which falls below the MLP accuracy in the present study. Moreover, in research conducted by Joao-Filho et al. the MLP model with a sensitivity of 100%, a specificity of 80%, an accuracy of 88%, and AUC=0.918 proved having a higher efficiency than the other examined models, namely CART, MLR, and SVM [26]. The MLP model in the present study demonstrated higher specificity, accuracy, and AUC, and as a result, a higher efficiency compared to findings reported by JoaoFilho et al.'s study.

CONCLUSION

Herein, after a comprehensive review of the risk factors and the main symptoms of tuberculosis, the following risk factors were identified as the most significant ones: contact with TB-infected patients, having AIDS, smoking, sex, age, and the symptoms of fever, night sweats, bloody sputum, weight loss, coughing, and abnormal WBC and Hb. Furthermore, ESR proved to be the most important test for TB diagnosis.

In comparing the specificity values of MLP and ANFIS by the ratio test, it was revealed that although MLP enjoyed a higher specificity as compared to ANFIS, there was no significant difference between their specificity values. With respect to other metrics, namely, sensitivity, accuracy, and RMSE, the MLP model proved having a higher efficiency compare to ANFIS as there were significant differences between these metrics. Finally, a significant difference was found between the two models in terms of efficiency, as the MLP model indicated a better performance in TB diagnosis.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

Conceptualization: Hosseini A, Moghaddasi H. Data curation: Hosseini A, MohebiMushaei S. Creat train and test TB diagnosis model: Hosseini A, Moghaddasi H. Formal analysis: Hosseini

A, Moghaddasi H, MohebiMushaei S. Writing-original draft preparation: Rabiei R, Hosseini A. Writing-review and editing: Rabiei R. Approval of final manuscript: all authors.

REFERENCES

1. Omisore MO, Samuel OW, Atajeromavwo EJ. A genetic-neuro-fuzzy inferential model for diagnosis of tuberculosis. Appl Comput Inform. 2017;13:27-37.
2. Roberts CA, Buikstra JE. The bioarchaeology of tuberculosis: A global perspective on are-emerging disease: University Press of Florida; 2003.
3. Bobak CA, Titus AJ, Hill JE. Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets. Appl Soft Comput. 2019;264;73-74.
4. Organization TWH. Global Tuberculosis Report 2017.
5. Asha T, Natarajan S, Murthy KB. Association-rule-based tuberculosis disease diagnosis. Second International Conference on Digital Image Processing: International Society for Optics and Photonics. 2010;75462Y.
6. Elveren E, Yumuşak N. Tuberculosis disease diagnosis using artificial neural network trained with genetic algorithm. J Med Syst. 2011;35:329-332.
7. Lakshmi K, Krishna MV, Kumar SP. Utilization of data mining techniques for prediction and diagnosis of tuberculosis disease survivability. Int J Mod Educ Comput Sci. 2013;5:8.
8. Organization TWH. Global Tuberculosis Report 2016. 2016.
9. Saybani MR, Shamshirband S, Golzari S, Wah TY, Saeed A, Kiah MLM, et al. RAIRS2 a new expert system for diagnosing tuberculosis with real-world tournament selection mechanism inside artificial immune recognition system. Med Biol Eng Comput. 2016;54:385-399.
10. Saybani MR, Shamshirband S, Hormozi SG, Wah TY, Aghabozorgi S, Pourhoseingholi MA, et al. Diagnosing tuberculosis with a novel support vector machine-based artificial immune recognition system. Iran Red Crescent Med J. 2015;17.
11. Uçar T, Karahoca A, Karahoca D. Tuberculosis disease diagnosis by using adaptive neuro fuzzy inference system and rough sets. Neural Comput Appl. 2013;23:471-483.
12. Rusdah EW, Winarko E. Review on data mining methods for tuberculosis diagnosis. Information Systems International Conference. 2013:2-4.
13. PeterJG, Theron G, Muchinga TE, Govender U, Dheda K. The diagnostic accuracy of urine-based Xpert MTB/RIF in HIV-infected hospitalized patients who are smear-negative or sputum scarce. PloS one 2012;7:e39966.
14. Faria J, Seixas J, Souza Filho J, Orjuela A, Vieira A, Kritski A, et al. Pleural tuberculosis diagnosis based on artificial neural networks models. X Congreso Brasileiro de Inteligencia Computacional CBIC2011.
15. Polat K, Şahan S, Kodaz H, Güneş S. Breast cancer and liver disorders classificationusing artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism. Expert Systems with Applications 2007;32:172-183.
16. Asha T, Natarajan S, Murthy K. Effective classification algorithms to predict the accuracy of tuberculosis-A machine learning approach. Int J Inf Secur. 2011;9:89.
17. Hand DJ. Principles of data mining. Drug Saf. 2007;30:621-622.
18. Durairaj M, Kalaiselvi G. Prediction of diabetes using soft computing techniques-A survey. Int J Sci Res. 2015;4:190-192.
19. Ahmad P, Qamar S, Rizvi SQA. Techniques of data mining in healthcare: A review. Int J Comput Appl. 2015;120.
20. Koh HC, Tan G. Data mining applications in healthcare. J Health Inf Manag. 2011;19:65.
21. Bakar A, Febriyani F. Rough neural network model for tuberculosis patient categorization. Proceedings of the international conferenceon electrical engineering and informatics. 2007.
22. Pendharkar P, Rodger J, Yaverbaum G, Herman N, Benner M. Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 1999;1223:32-37.
23. Asha T, Natarajan S, Murthy K. Optimization of association rules for tuberculosis using genetic algorithm. Комп'ютинг. 2013;12:151.
24. Benfu Y, Hongmei S, Ye S, Xiuhui L, Bin Z. Study on the artificial neural network in the diagnosis of smear negative pulmonary tuberculosis. 2009 WRI World Congress on Computer Science and Information Engineering: IEEE; 2009.
25. Er O, Temurtas F, Tanrıkulu AÇ. Tuberculosis disease diagnosis using artificial neural networks. J Med Syst. 2010;34:299-302.
26. João Filho BdO, de Seixas JM, Galliez R, de Bragança Pereira B, de Q Mello FC, Dos Santos AM, et al. A screening system for smear-negative pulmonary tuberculosis using artificial neural networks. Int J Infect Dis. 2016;49:33-39.
27. Uçar T, Karahoca A. Predicting existence of *Mycobacterium tuberculosis* on patients using data mining approaches. Procedia Comput Sci. 2011;3:1404-1411.