©2017 Published in 5th International Symposium on Innovative Technologies in
Engineering and Science 29-30 September 2017 (ISITES2017 Baku - Azerbaijan)

Designing Multi-layered Artificial Neural Networks for Risk Analysis of Lung Cancer Disease

[1]Umut Kaya, [2]Atınç Yılmaz and *[2]Ediz Şaykol
[1]Computer Technologies Department, Kavram Vocational High School, Istanbul, Turkey
[2]Faculty of Engineering, Department of Computer Engineering Beykent University, Turkey

Abstract

This paper is intended to propose a multi-layered artificial neural network model for lung cancer risk analysis. Hundreds of thousands of people die every year due to cancer, use of mathematical models especially in decision making may help reduce the loss of patients by augmenting effective predictions in early stages of diagnosis. Feed forward back propagation and cascade forward back propagation network structures have been applied for creating multi-layered ANN model. Levenberg-Marquardt algorithm and Bayes regulation algorithm have been used for training step. Hyperbolic tangent sigmoid function has been chosen as the transfer function and the learning algorithms have been integrated. Four multi-layered ANN models are evaluated with regression and validation analysis on a dataset of 616 people. A comparative analysis is also performed to show that our bayes1ffb model gets higher accuracy value than two existing models in the literature

Key words: Artificial Neural Networks, Lung Cancer, Risk Analysis

1. Introduction

According to the World Health Organization 2004 report, entitled as changing history, lung cancer is one of the most morbid diseases. Usage of mathematical models in medicine may help reduce the loss of patients by augmenting effective predictions in early diagnosis [1]. The main factors that cause lung cancer are stated as exposure to tobacco, genetics, and air pollution, radiation, breathing chemical materials, life quality and nutritional habits. Existing studies about the early diagnosis of the lung cancer show that pre-diagnosis increases the success of treatment [2]. Hence, pre-diagnosis has an important role in treatment of lung cancer disease, so as artificial intelligence to augment the pre-diagnosis stage, as mentioned in the literature [3].

There exist techniques based on fuzzy logic and artificial neural networks (ANN) for predictive modelling and risk analysis of lung cancer with using a set of personal data including smoking habits, age and genetics for diagnosis [4]. In the analysis of lung cancer risk model, sex, age, skin tone, smoking, age of starting smoking, passive smoking environment, occupational status, living environment, genetic status, economic status, and nutritional habits were determined as factors for cancer risk [5]. In [6], ANNs are used to classify the cancer types according to gene recognition profile by using pattern recognition method with using the distance of Euclidean Square for each cancer type was computed and 95% of validation result was attained.

Ganesan et al. applied artificial neural network for cancer diagnosis by using demographic data [4] and attained over 87% validation results in conclusion. Ashwin et al. proposed to detect lung cancer nodule in images with using ANN for cad diagnosis system [7]. Contrast limited adaptive

*Corresponding author: Address: Faculty of Engineering, Department of Computer Engineering Beykent University, 34396, Istanbul TURKEY. E-mail address: edizsaykol@beykent.edu.tr, Phone: 444 19 97 52 99

histogram equalization work was experted trained medical diagnosis with modified BFGS accuracy 96.7%, sensitivity 92.1% and specificity 94.3%.

Kathalkar et al. applied artificial neural network to brain cancer analysis and classification of cancer types at MRI images of the patients [8]. Ahmed et al. [9] researched early detection of lung cancer risk using data mining. Rajan and Chelvan proposed a model to measure data mining techniques to lung cancer dataset could provide reliable performance in the detection of lung cancer at beginning stage applying Kohonen Self Organizing Map [10]. Mottaleb adopted artificial neural network for viroteraphy cancer treatment analysis before the clinic symptoms of breast cancer disease [11]. Utomo et al. utilized artificial neural network to breast cancer diagnosis by using extreme learning techniques [12]. Gorynski et al. studied about early lung cancer detection by using artificial neural network [13].

The main contribution of this study is to provide effective setups for this crucial prediction for human health employing ANNs. To the best of our knowledge, this is the first effort in adapting multi-layered ANN with these different learning algorithms to lung cancer risk analysis. Also multi-layered method has effective ability about to solve non-linear problems as like XOR problems. The remainder of this paper is organized as follows: Our proposed multi-layered artificial neural network model is given Section 2 along with the details of the 4 techniques differing by learning algorithms. These different algorithms are utilized to obtain better learning and decision making performance, and the details of this performance analysis is explained in Section 3. Finally, Section 4 concludes the paper.

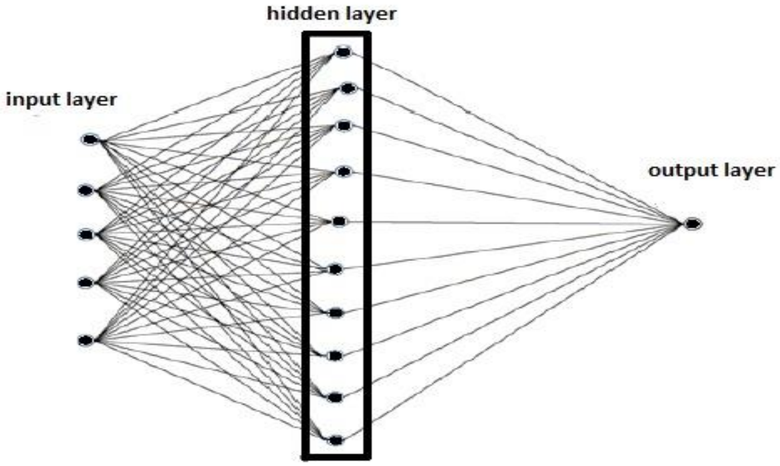


Figure 1: Multi-layered artificial neural network models.

2. Multi-layered Artificial Neural Network Model

Our multi-layered artificial neural network model is shown in Figure 1. Here, Levenberg-Marquardt algorithm and Bayes regulation algorithm have been used for training step. Levenberg Marquardt is fast learning algorithm which has simple structure with best ways of Gauss-Newton and Steepest-Decent algorithm. Output evaluation using with Levenberg-Marquardt starts with to assign an initial value to sequences of initial weights and proceeds to compute the sum of error squares of target value and real value until computing the all sum of error squares. Bayesian network performs a multivariate probability distribution of random variables. This distribution provides the variations of calculations to be done. The partition which is named as node represents the random variables and the connections of among the nodes show the dependency of variables conditions over the probability between these variables in the structure. Bayesian networks are more favorable than the other methods as to show the probability distributions of all variables in the network, to allow the finite values of the variables can be renewed and thus to provide the probability values of the process and the variables can be reflected more realistically with new observations.

Hyperbolic tangent sigmoid function has been chosen as the transfer function and the learning algorithms have been integrated to multi-layered neural networks. ANN models that we have utilized here can be listed as follows:

- bayes1ffb (feed forward back propagation model with Bayes regulation) is shown in Figure 2 and regression graph of this model is given in Figure 3.

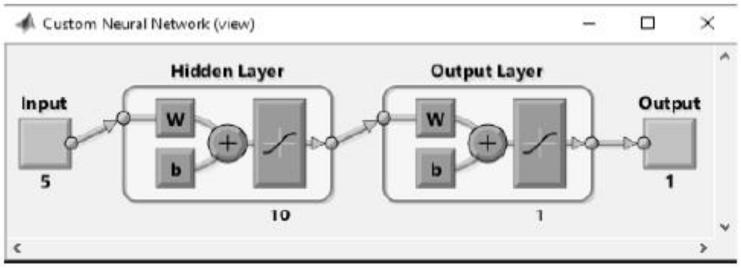


Figure 2: Feed forward back propagation ANN model for bayes1ffb and lm1ffb.

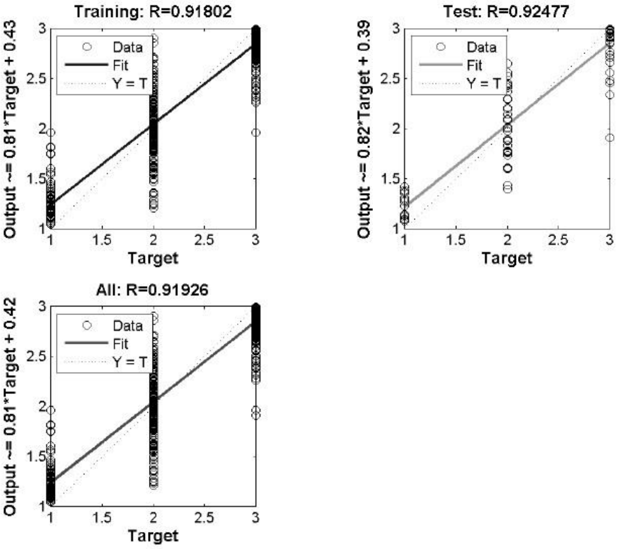


Figure 3: bayes1ffb regression graph.

- bayes2cfb (cascade forward back propagation model with Bayes regulation) is shown in Figure 4 and regression graph of this model is given in Figure 5.

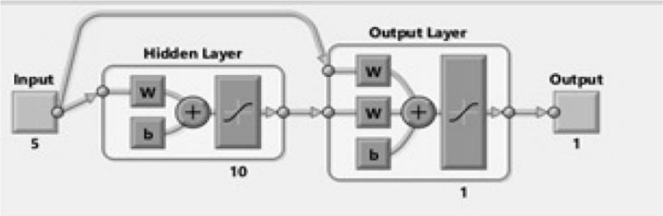


Figure 4: Cascade forward back propagation ANN model for bayes2cfb and lm2cfb.

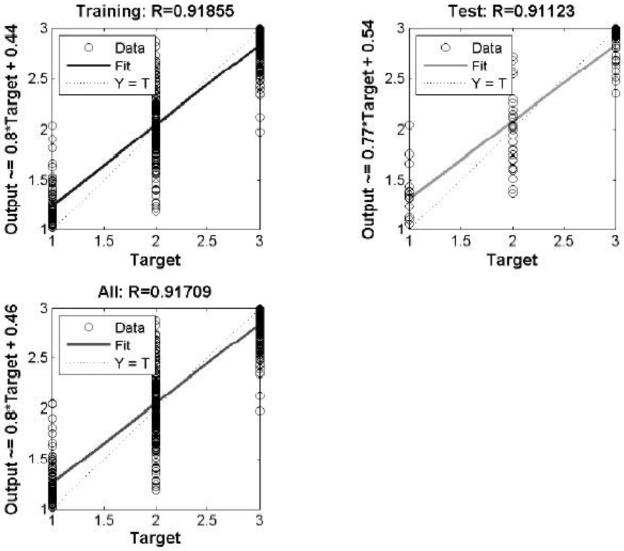


Figure 5: Regression graph of bayes2cfb.

- lm1ffb (feed forward back propagation model with Levenberg-Marquardt) is shown in Figure 2 and regression graph of this model is given in Figure 6.
- m2cfb (cascade forward back propagation model with Levenberg-Marquardt) is shown in Figure 4 and regression graph of this model is given in Figure 7.

The comparison of output values of each of these 4 methods with the target output values are given in Figure 8.

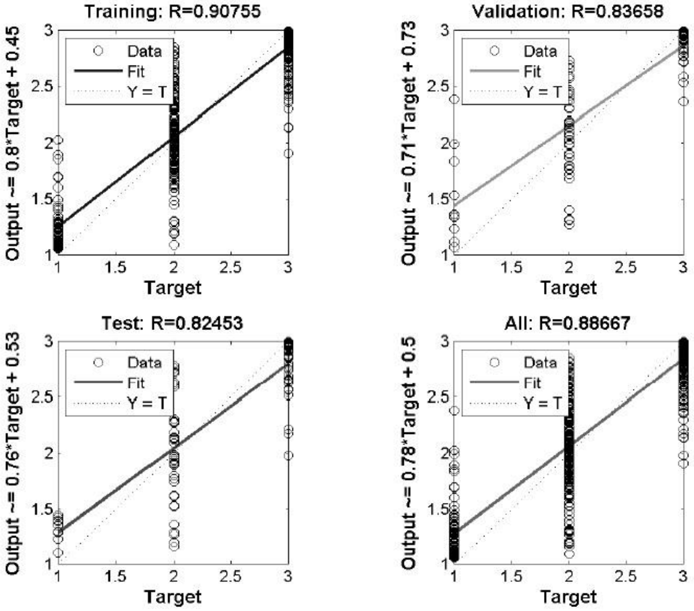


Figure 6: Regression graph of lm1ffb.

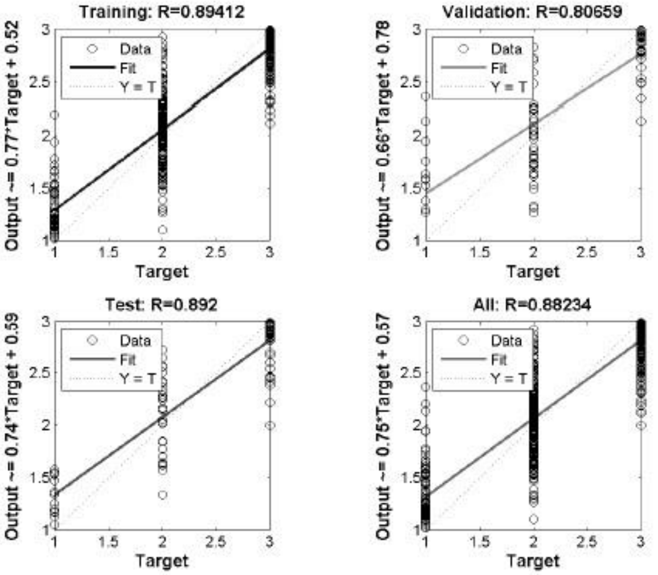
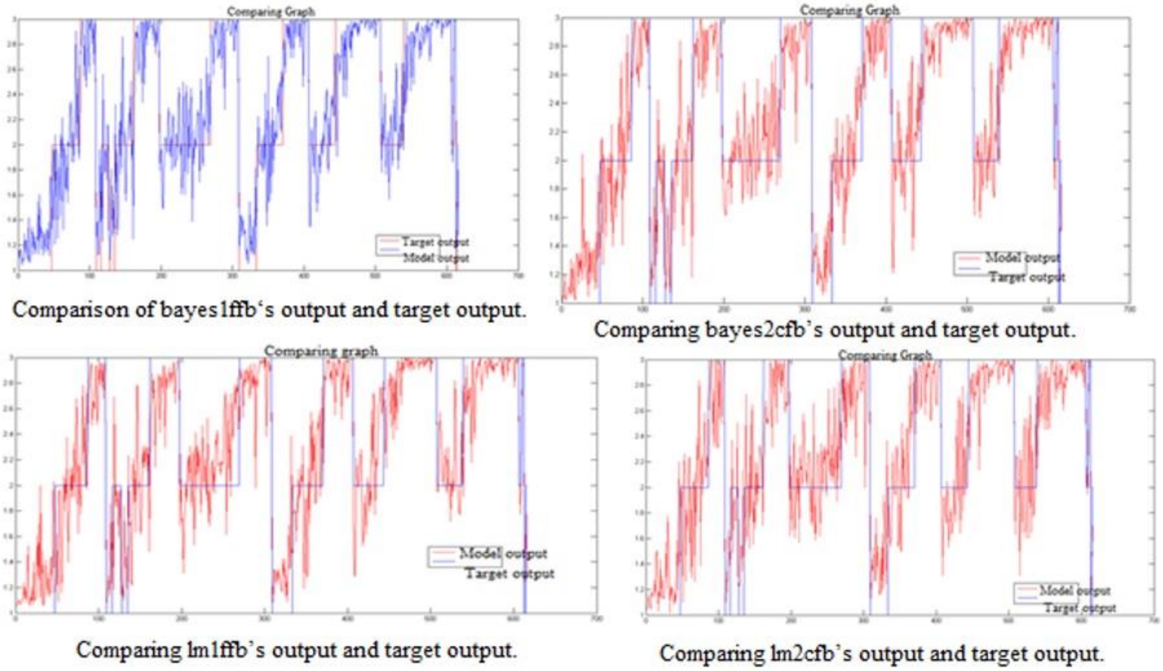


Figure 7: Regression graph of lm2cfb.



Comparison of bayes1ffb's output and target output.

Comparing bayes2cfb's output and target output.

Comparing lm1ffb's output and target output.

Comparing lm2cfb's output and target output.

Figure 8: Comparison of all method's output and target output.

For bayes1ffb model, the output values overlap with the target output values. For bayes2cfb model, the output values agree with target output value at 2 and 3, and the output overlaps with the target output. For lm1ffb model, the output values agree with target output and overlap of target output at 2 and 3. For lm2cfb model, the outputs overlap with the target outputs, and they agree with at 2 and 3.

In this study, we have used a dataset of 616 people for input and output values, which had also been used in [14] to evaluate a learning scheme based on Fuzzy logic. Here, we used 70% of this data for training, 15% for testing and 15% for simulating purposes. The dataset has the following values:

1. Age: The age of starting smoking,
2. Smoke: The frequency of smoking,
3. Genetics: Existance of diagnosed lung cancer in a close relative,
4. Environment: The place where the person lives, and
5. Skin Colour: Swarthy or white skin.

In Table 1, a set of sample data is given. The values are enumerated by a similar manner in [14] to be used in neural network learning steps. Age is a numeric value that shows the starting age of smoking, is crucial because starting to smoke in early ages and smoking for a long period of time enhance the risk of having lung cancer.

Table 1: Sample data in our dataset.

age	smoke	genetics	environment	skin color	output
34	6.05	0	8.9	10	1
30	8.05	0	4.2	50	1
53	30.45	0	10.75	70	2
51	39.35	1.1	10.1	70	3
41	28.25	0.65	9.35	30	3

The Smoke attribute is computed as Smoke = Usage + Passive smoke + (Actual Age – Age). In Table 2, a set of smoke, passive smoke and genetics values are given.

Table 2: Smoke, passive smoke and genetics values.

usage		passive smoke		genetics	
values	means	values	means	values	Means
0.6	no cigarettes	0.30	no passive smoker	0.00	not in known relative
1.1	1-2 cigarettes	0.95	only at home	1.70	parent, brother, sister
1.5	half package	0.95	only at work	1.30	aunt, uncle
1.7	full package	0.95	only with friend	1.30	grand parents
2.4	more than 1 package	1.30	at work and with friend	1.10	cousins or distant relatives
		1.55	more	0.65	do not know

The Environment attribute is computed as Environment = Occupation + Medium of life + Socio-economical status + Nutrition, where Nutrition score is computed via a questionnaire, and the Occupation, Medium of life, and Socio-economical status values are set in Table 3.

Table 3. Occupation, medium of life and socio-economic values.

occupation		medium of life		socio-economic status	
values	means	values	means	values	means
0.25	desk based	0.3	house w garden	0.6	very good
1.55	risky worker	0.5	apartment w garden	1.0	good
1.20	worker in factory	1.0	in city center	1.5	Medium
		1.5	near factory	1.9	bad
				2.4	very bad

The Skin Color attribute is categorized into 5 classes from Swarthy to white skin having values 100-80, 80-60, 60-40, 40-20, 20-0, respectively.

3. Performance Experiments and Analysis

In this performance evaluation study, we used a dataset of 616 people [14] for the risk analysis of lung cancer disease. This dataset was divided into three, where 70% was used for training, 15% was used for testing, and 15% was used for simulating in each ANN model.

Four multi-layered ANN models were considered with regression and validation graphs (see Figures 3, 6, 7, 8), the performance can be ordered as follows:

bayes1ffb > bayes2cfb > lm1ffb > lm2cfb.

The output errors of the proposed 4 ANN models are shown in Figure 9. Box Whisker graph method in bayes1ffb gets less error value than the other models.

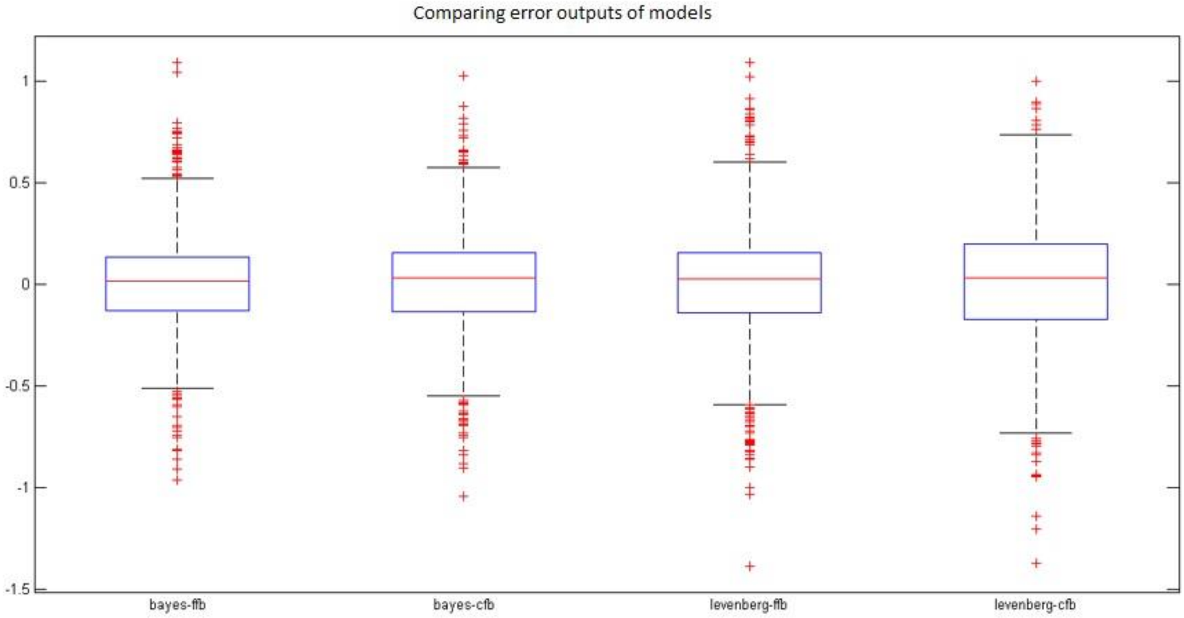


Figure 9: Comparing models output errors.

Another set of experiments is carried out to compare the performance of the proposed ANN models with a fuzzy logic model [14] in terms of output, mean square error sum, output error sum, and validation.

Table 4: Comparing results of models and fuzzy logic model.

	output sum	error square sum	output error sum	validation
bayes1ffb	1402.60	48.032	7.39702	0.919
bayes2cfb	1403.55	49.525	6.44719	0.917
lm1ffb	1409.52	65.553	0.47756	0.836
lm2cfb	1404.63	68.302	5.36495	0.806
fuzzy	1405.07	46.528	4.92389	0.800

Table 4 gives the results of this comparative analysis. Feed forward back propagation with Bayes regulation learning function multi-layered ANN model (bayes1ffb) gets the best result. In addition, bayes1ffb model gets higher accuracy value than the fuzzy logic model in [14], and a multi-layered ANN with Rumelhart and McLelland algorithm [4] in the literature.

4. Conclusion

In this study, we propose a multi-layered artificial neural network to evaluate the risks of lung cancer via utilizing four different learning algorithms. In the analysis of lung cancer risk model, sex, age, skin tone, smoking, age of starting smoking, passive smoking environment, occupational status, living environment, genetic status, economic status, and nutritional habits were determined as factors. We utilized enumeration and normalization formulae for these factors are use as the input for the network structures. We used a dataset of 616 people for the risk analysis of lung cancer disease. The regression and validation graph results are evaluated and as a result, bayes1ffb model would come to the forefront to applying the lung cancer risk analysis. As a future work, we aim to extend the use of our ANN models for different cancer types with an extended set of factors.

References

[1] M. Alvarez, Molecular basis of cancer and clinical applications, Surgical Clinics of North America 80 (2000) 443–457.

[2] V. Gant, S. Rodway, J. Wyatt, Artificial neural networks: Practical considerations for clinical applications, Clinical Applications of Artificial Neural Networks (2001) 329–356.

[3] P. Lisboa, A. Taktak, The use of artificial neural networks in decision sup- port in cancer: A Systematic review, Neural Networks 19 (4) (2006) 408–415.

[4] N. Ganesan, K. Venkatesh, M. Rama, A. Palani, Application of neural networks in diagnosing cancer disease using demographic data, International Journal of Computer Applications 1 (26) (2010) 76–85.

[5] N. V. Zandwijk, Aetiology and prevention of lung cancer, European Respiratory Monograph 17 (2001) 13–33.

[6] J. Khan, J. Wei, M. Ringnr, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, P. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine 7 (6) (2001) 673–679.

[7] S. Ashwin, J. Ramesh, S. Kumar, K. Gunavathi, Efficient and reliable lung nodule detection using a neural network based computer aided diagnosis system, in: Proceedings of the International Conference on Emerging Trends in Electrical Engineering and Energy Management (ICETEEEM 2012), 2012, pp. 135–142.

[8] A. Kathalkar, R. Kawitkar, A. Chopade, Artificial neural network based brain cancer analysis and classification, International Journal of Computer Applications 66 (10) (2013) 40–43.

[9] K. Ahmed, A. Al-Emran, T. Jesmin, R. Mukti, M. Rahman, F. Ahmed, Early detection of lung cancer risk using data mining, Asian Pacific Journal of Cancer Prevention 14 (1) (2013) 595–598.

[10] J. Rajan, C. Chelvan, A survey on mining techniques for early lung cancer diagnoses, in: Proceedings of the International Conference on Green Computing, Communication and Conservation of Energy (ICGCE 2013), 2013, pp. 918–922.

[11] G. Motalleb, Artificial neural network analysis in preclinical breast cancer, Cell Journal (Yakhteh) 15 (4) (2014) 324331.

[12] C. Utomo, A. Kardiana, R. Yuliwulandari, Breast cancer diagnosis using artificial neural networks with extreme learning techniques, International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3 (7) (2014) 10–14.

[13] K. Gorynski, I. Safian, W. Gradzki, M. Marszall, J. Krysinski, S. Gorynski, A. Bitner, J. Romaszko, A. Bucinski, Artificial neural networks approach to early lung cancer detection, Central European Journal of Medicine 9 (5) (2014) 632–641.

[14] A. Yılmaz, K. Ayan, Cancer risk analysis by using fuzzy logic approach and performance status of the model, Turkish Journal of Electrical Engineering and Computer Science 21 (3) (2013) 897–912.