Special Issus: 1st Scientific International Conference, College of Science, Al-Nahrain University, 21-22/11/2017, Part I, pp.196-201

In Silico Model for Lung Cancer Prediction Based on TP53 mutations Using Neural Network

Zahraa Naser Shahweli [1] and Ban Nadeem Dhannoon [2]
Computer Science Department, College of Science, Al-Nahrain University, Baghdad-Iraq.
Corresponding Author: [1] Stcs-zns16@sc.nahrainuniv.edu.iq, [2] bnt@sc.nahrainuniv.edu.iq

Abstract

In silico models have become well known in the current decade because they assist researchers and specialists in organizing and analyzing big data. To complete their work, these models require powerful techniques and algorithms, the most important of which are machine learning algorithms. This work utilizes the Relief F algorithm for feature selection and trains the back propagation neural network (BPNN) algorithm on the UMD TP53 all-2012-R1-US database for lung cancer. Lung cancer is the most commonly diagnosed cancer among women and men, and can be predicted from mutations that occur in the TP53 tumor suppressor gene. Five measures are used to estimate performance: sensitivity and specificity are important dimensions utilized to obtain the receiver operating characteristic (ROC) curve; accuracy and *F* measure are necessary to determine algorithm precision; and Matthews correlation coefficient (MCC), which is the most important measure, provides the right criterion for classification algorithms. The Relief F and BPNN algorithms achieve satisfactory results that reach 99.41 for sensitivity, 95.39 for specificity, 99.04 for accuracy, 99.47 for *F* measure, and 0.93 for MCC. [DOI: 10.22401/ANJS.00.1.26]

Keywords: In silico model, TP53 gene, Lung cancer, BPNN, Relief F feature selection.

Introduction

Cancer cells contain many of the genomic mutations present in almost all or most of the cells in a tumor; these mutations are due to more than one gene. As such, research on gene mutations can enable the prediction of cancer, [1]. The TP53 tumor suppressor gene is a good predictor of cancer in the human body depending on the type and location of the mutation that indicates cancer in a human organ. Lung cancer is most affected by mutations in the TP53 gene, where approximately 40 to 70 of the mutations that cause lung cancer are observed, [2]. Clinical studies since 1930 have shown that exposure to tobacco is a cause of most cancers, including lung cancer, because of the carcinogens in tobacco that lead to mutations in several genes over time; most of the mutations occur in the TP53 gene, [3].

Several types of TP53 mutations have been identified, and these mutations are arranged in databases for ease of use by specialized researchers. One of these databases is the UMD TP53 database, which contains information about all cancer types. The information contained in this database is important for early detection and proper treatment. The organization, analysis, and storage of biological information require the use of high-processing-speed devices, such as computers, because they are faster and have fewer errors than manual handling of humans, [4]. In addition, the significant increase in biological or genetic information requires the use of computational systems. In silico models were developed in the late twentieth century to aid biological and medical specialties, [5].

This study proposes an in silico model for diagnosing and predicting the most common cancers in Iraq see Table (1) and the world see Fig.(1) depending on the cancer-causing mutations that occur in the TP53 gene. Machine learning techniques have been used to help build this model. Relief F feature selection is used as a preprocessing phase to minimize the number of features, followed by back propagation neural network (BPNN) algorithm as a learning phase. This study utilized the UMD TP53 all-2012-R1-US database [6] to extract information related to lung cancer and to train neural networks.

Table (1)
Ten Most Common Cancer Deaths Registered by Site and Gender in Iraq-from the Iraqi Cancer Board 2012.

Site	Total	Male	Female	%	Incidence Rate
Bronchus & Lung	1756	1270	486	17.09	5.13
Brain	1122	576	546	10.92	3.28
Breast	1008	0	1008	9.81	2.95
Leukemia	994	525	469	9.67	2.91
Liver	673	333	340	6.55	1.97
Stomach	539	303	236	5.24	1.58
Urinary Bladder	536	395	141	5.22	1.57
Colorectal	504	280	224	4.90	1.47
Panreasc472	472	268	204	4.59	1.38
Lymph Nodes (Lymphoma)	361	202	159	3.51	1.06
Total Ten	7965	4152	3813	77.50	23.28
All Cancer's Deaths	10278	5346	4932	100.00	30.05

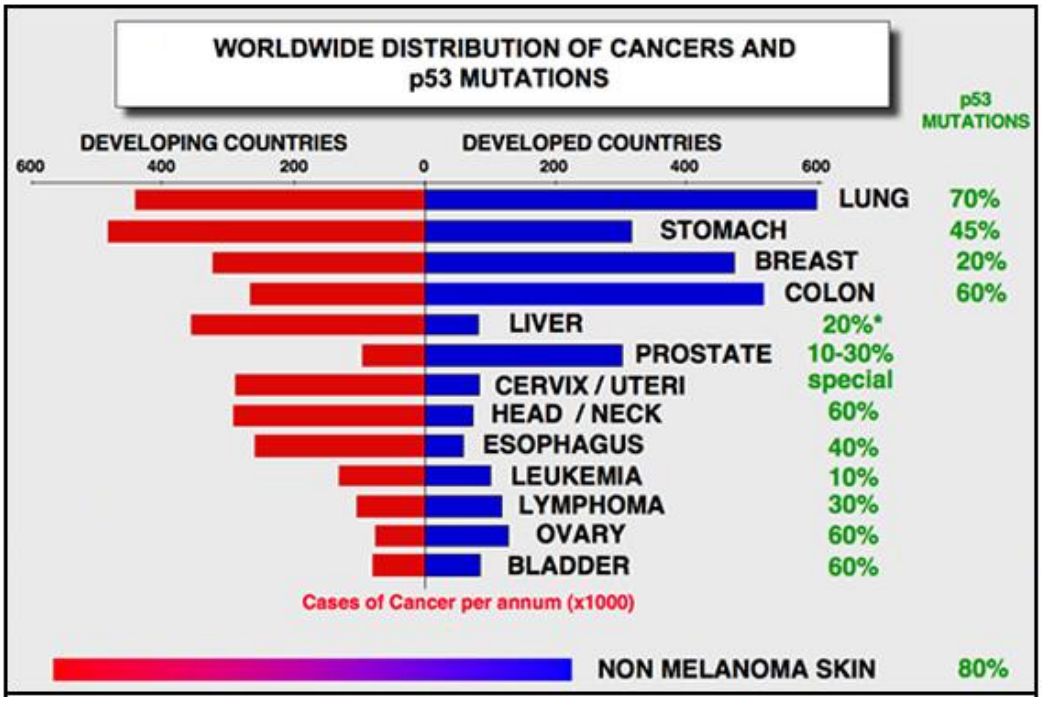

Fig. (1) TP53 Mutations and All Cancer Types, [7].

Related Work

Kawsar Ahmed [8], built prediction system of lung cancer by using significant pattern. First, K- means clustering algorithm for identifying relevant and non-relevant data was used. Then AprioriTid and a decision tree algorithm used as second significant patterns. 400 cancer and non-cancer patients data were collected from different diagnostic centers, pre-processing operations performed on these data, such as deleting duplicate records and completion of missing data.

Ayad Gh. Ismaeel [9], try to diagnose, predict and classified cancers by means of transformations in tumor protein P53 sequence. The procedure of classification particular cancer is performed by utilizing two approaches. The first approach predicts whether the individual has mutations that reason malignancy or not. The second approach that classifies the mutations are acquired from the first way to know which type of disease it caused (malignancy types). Basic local alignment search tool (BLAST)

and multiple sequence alignment (CLUSTALW) are bioinformatics programs utilized as a part of a first approach. BPNN algorithm is utilized as a part of a second approach with the execution mean square error (MSE) spans to (5.6339E-17), and the training rate meets (1). To train BPNN, 12 features from 53 features of UMD_Cell_line_2010 are utilized.

Z. N. Shahweli [10], proposed classification system for breast and prostate cancers using seven data sets contain mutations of TP53 gene. A back propagation neural network used as a learning method with hybrid model of 5-fold cross validation with validation set. Five measures utilized to measure the performance of proposed work and all of which have high values in predicting breast and prostate cancer based on mutations occur in TP53 gene.

Scientific Methods

A. In vivo, In vitro, and In silico

The term "in vivo" refers to the conduct of biological experiments within the body of an organism, including animals, plants, and in several cases, humans, whereas the term "in vitro" refers to the conduct of laboratory experiments outside the body of an organism using samples and laboratory tools, such as drug discovery tests, [11]. Meanwhile, the in silico model is an extension of the in vivo and in vitro models based on the use of a computer and its programs. The expected result is that the computing capacity increases at the lowest cost. The in silico model combines the advantages of the in vivo and in vitro models, and is currently used in large-scale biological systems, such as genomic information and drug testing, [12].

The in silico computational model has two main advantages. First, difficulty in vivo or in vitro tests can be simulated for diagnosis and prediction such that the in silico model can often be used to substitute for the in vivo and in vitro models. Second, the number of experiments and samples can be increased in the case of the lack of experimental methods or early screening systems such that its results can be the basis for future research, [13]. Figure 2 explains the three models; in vivo, in vitro and in silico.

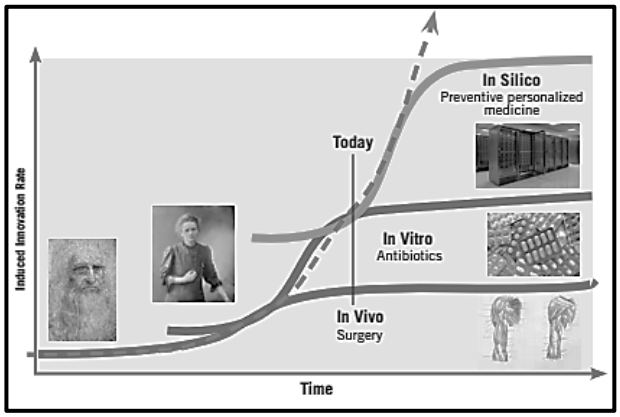

Fig. (2) In vivo, In vitro, and In silico model, [14].

B. Proposed wWork

The UMD TP53 all-2012-R1-US database is used as material in this work. It consists of 18824 rows and 34 columns. After performing needed operations on this data, it will be ready for using by BPNN.

This work is divided into two phases, namely, preprocessing and learning phases. The preprocessing phase has several steps. First, 14 features are selected from the UMD TP53 all-2012-R1-US database based on consultant specializing in genetic engineering; these features are related only to lung cancer. Second, these features are transformed into numerical form for using with the neural network. Finally, the selected features are normalized to configure their use for ReliefF feature selection.

In the learning phase, BPNN with *K*-fold cross-validation was performed. Five measures, namely, sensitivity, specificity, accuracy, harmonic *F* measure, and Matthews correlation coefficient (MCC), were used to evaluate the performance. The results are portrayed using the receiver operating characteristic (ROC) curve.

C. Relief F Algorithm for Feature Selection

Feature selection is an important preprocessing step in machine learning to distinguish the set of features that has a strong role to help the machine to learn better, [15]. In this work, the Relief F algorithm for feature selection was applied.

The Relief F algorithm submitted by Kononenko in 1994 was an update to the original Relief algorithm, which was restricted to two classes only. The Relief F algorithm

works by randomly selecting an instance t_i and searching for k instances of the same class called hit h_j and k instances of different classes called miss m_j. The weights of the features are updated based on the values of t_i, h_j, and m_j, [16]. Algorithm (1) explains the steps of Relief F feature selection.

Algorithm 1. Relief F Feature Selection, [17]:

Input: training data

Output: weight for each attribute

1. set all weights w [A]:=0
2. for i:=1 to m do begin//m: training data size.
3. randomly select an instance t_i;
4. find k-nearest hits h_j;
5. for each class C ≠ class(t_i) do
6. from class C find k nearest misses m_j(c)
7. for A:=1 to a// a: no. of attributes
8. $w[A]=w[A] - \sum_{j=1}^{k} \frac{\mathbf{diff}\ (\mathbf{a}, ti, hj)}{(\mathbf{m}.\mathbf{k})}$
9. $\sum_{c \neq class\ ti} \frac{\frac{p(c)}{1-p(class(ti))} \sum_{j=1}^{k} \mathbf{diff}\ (\mathbf{a}, ti, mj(\mathbf{c}))}{(\mathbf{m}.\mathbf{k})}$
10. end

D. Performance Measures

The performance of the proposed work is measured using accuracy (Acc), sensitivity (Sn), specificity (Sp), *F* measure, and MCC. The ROC curve is plotted to assess the classifier. All these measures depend on the correct and incorrect values that the classifier predicted.

Accuracy determines the rate of correct cases from all test data, sensitivity determines the true positive rate (TPR) from all positive cases, specificity determines the true negative rate from all negative cases, *F* measure is related to the TPR predicted by the classifier, and MCC is the most important measure because it uses all factors in a confusion matrix. Therefore, the MCC is used when the dataset is unbalanced or when negative and positive cases have the same level of importance, [10].

The ROC curve is used to determine the best classifier. The curve of the best classifier is between the low value of the false positive rate (1 – Sp) and the high value of the TPR (Sn). These values shift the points toward the upper left corner of the curve that represents the perfect point, [18].

Results and Discussion

The results of BPNN with 5-fold cross-validation on the selected database that has 14 features were compared with the results of Relief F feature selection as a preprocessing step before performing BPNN. Table (2) contains the results of BPNN without feature selection, whereas Table (3) contains the results of Relief F algorithm as a preprocessing step and BPNN algorithm as a learning step. The ROC curve for the results in Table (3) is illustrated in Fig. (3).

Table (2)
Results of BPNN Before Feature Selection.

Fold no.	Sn.	Sp.	Acc.	F-measure	Mcc
1	99.41	94.73	98.94	99.41	0.93
2	100	86.48	98.67	99.27	0.90
3	99.70	97.05	99.47	99.70	0.96
4	99.42	83.33	98.14	98.99	0.85
5	99.69	96	99.20	99.54	0.95
Average	99.64	91.52	98.88	99.38	0.92
Error for all epochs	0.0111	Run time (m)	02:02	Max Epochs	3000

Table (3)
Results of Relief F algorithm and BPNN.

Fold no.	Sn.	Sp.	Acc.	F-measure	Mcc
1	99.41	97.36	99.20	99.55	0.95
2	99.70	89.18	98.67	99.27	0.91
3	98.83	97.05	98.67	99.27	0.92
4	99.42	93.33	98.94	99.42	0.92
5	99.69	100	99.73	99.84	0.98
Average	99.41	95.39	99.04	99.47	0.93
Error for all epochs	0.0095	Run time (m)	00:14	Max Epochs	1000

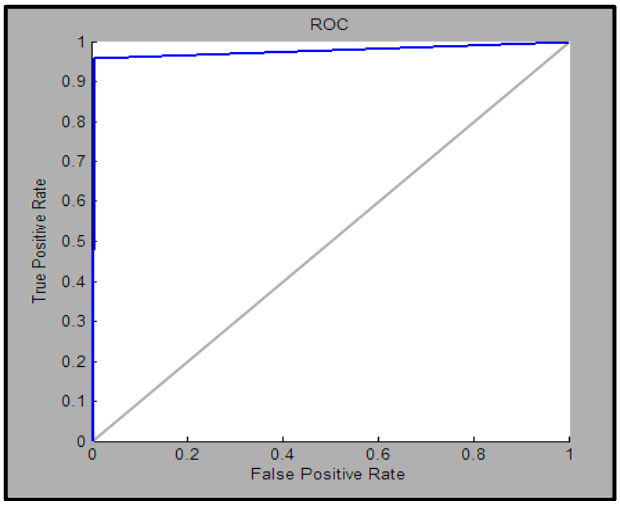

Fig. (3) ROC curve for BPNN and Relief F algorithm.

The results in Table (3) were better than those in Table (2) because of the deletion of unnecessary features by the Relief F algorithm. The deleted features in this work are three features that have duplicate values. Accordingly, BPNN training on 11 features was conducted with fewer epochs and less time. Consequently, feature selection algorithms help neural networks learn better and faster.

Conclusions

Lung cancer, similar to other cancers, has biochemical and genetic reasons. Genetic reasons related to mutations occur in the oncogene and tumor suppressor gene; among these genes is the TP53 gene. In silico models with machine learning techniques aid biologists and doctors in diagnosing and predicting cancer. Prediction is important in determining the appropriate drug to administer at the appropriate time.

This work aims to build an in silico model that predicts lung cancer on the basis of the mutations that occur in the TP53 gene. To achieve this objective, the Relief F algorithm is used to select effective features for diagnosis. These features are inputted into the BPNN for training. *K*-fold cross-validation is used to split the data fivefold. This method allows the use of the data for training and testing.

The Relief F algorithm reduced the features from 14 to 11 after deleting ineffective features. The results of BPNN before and after feature selection are listed in Tables (2) and (3), where the deletion of undesirable features obtained better results and reached on average 99.04 for accuracy, 99.47 for *F* measure, and 0.93 for MCC. Moreover, the ROC curve plotted for the results in Table 3 clarifies the satisfactory outcomes of the classification algorithm.

Acknowledgment

The authors thank Prof. Zahra Mahmoud Al-Khafaji for helping them in selecting features from the databases that relate to cancer classification.

References

[1] Stamatakos, G., N. Graf, and R. Radhakrishnan, *Multiscale cancer modeling and in silico oncology: emerging computational frontiers in basic and translational cancer research.* J Bioeng Biomed Sci, 2013. **3**: p. E114.

[2] Bumroongkit, K., et al., *TP53 gene mutations of lung cancer patients in upper northern Thailand and environmental risk factors.* Cancer genetics and cytogenetics, 2008. 185(1): p. 20-27.

[3] Gibbons, D.L., L.A. Byers, and J.M. Kurie, *Smoking, p53 mutation, and lung cancer.* Molecular Cancer Research, 2014. **12**(1): p. 3-13.

[4] Hapudeniya, M., *Artificial neural networks in bioinformatics.* Sri Lanka Journal of Bio-Medical Informatics, 2010. 1(2).

[5] Edelman, L.B., J.A. Eddy, and N.D. Price, *In silico models of cancer.* Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2010. 2(4): p. 438-459.

[6] France database of TP53 gene. *UMD TP53*. 2012 2012 [cited 2017 1 february]; Available from: http://p53.free.fr.

[7] France database of TP53 gene. *UMD TP53 database*. 2012 [cited 2017 August]; Available from: http://p53.free.fr/Database/p53_cancer/all_cancer.html.

[8] Ahmed, K., et al., *Early detection of lung cancer risk using data mining.* 2013.

[9] Ismaeel, A.G. and D.Y. Mikhail, *Effective Data Mining Technique for Classification Cancers via Mutations in Gene using*

Neural Network. arXiv preprint arXiv:1608.02888, 2016.

[10] Shahweli, Z.N., B.N. Dhannoon, and R.S. Ramadhan, *In Silico Molecular Classification of Breast and Prostate Cancers using Back Propagation Neural Network.* Cancer Biology, 2017. **7**(3).

[11] Mensch, J., et al., *In vivo, in vitro and in silico methods for small molecule transfer across the BBB.* Journal of pharmaceutical sciences, 2009. **98**(12): p. 4429-4468.

[12] Colquitt, R.B., D.A. Colquhoun, and R.H. Thiele, *In silico modelling of physiologic systems.* Best Practice & Research Clinical Anaesthesiology, 2011. 25(4): p. 499-510.

[13] Jeanquartier, F., et al., *In silico modeling for tumor growth visualization.* BMC systems biology, 2016. **10**(1): p. 59.

[14] Marchal, T., *IN VIVO, IN VITRO, IN SILICO!* ANSYS ADVANTAGE, 2015. Volume IX(Issue 1).

[15] Sewell, M., *Feature selection.* Online]. http://machine-learning. martinsewell. com/feature-selection, 2007.

[16] Robnik-Šikonja, M. and I. Kononenko, *Theoretical and empirical analysis of ReliefF and RReliefF*. Machine learning, 2003. 53(1-2): p. 23-69.

[17] Durgabai, R., *Feature selection using Relief F algorithm.* IJARCCE—International Journal of Advanced Research in Computer and Communication Engineering, 2014. 3(10).

[18] Majid, A., *Optimization and combination of classifiers using Genetic Programming.* Faculty of Computer Science, GIK institute, Swabi, 2006.