In Silico Molecular Classification of Breast and Prostate Cancers using Back Propagation Neural Network

Zahraa Naser Shahweli[1], Ban Nadeem Dhannoon[2], Rehab S. Ramadhan[3]

[1.] Computer Science Department, College of Science, Al-Nahrain University, Baghdad, Iraq,
Stcs-zns16@sc.nahrainuniv.edu.iq
[2.] Computer Science Department, College of Science, Al-Nahrain University, Baghdad, Iraq,
bnt@sc.nahrainuniv.edu.iq
[3.] Medical Biotechnology Department, College of biotechnology, Al-Nahrain University, Baghdad, Iraq,
rehabrebah@biotech.nahrainuniv.edu.iq

Abstract: Cancer is a standout amongst the most widely recognized and complex infections of the present century since it happens because of numerous organic and physical responses. One of the amplest and most boundless growths for the ladies today is Breast tumor, while prostate malignancy is a worry for some men. Computational models of disease are being created to help both biological invention and clinical prescription. The In silico models encourage the accumulation and utilization of trials to break down and separate rich organic data from vast natural database. In this study, a total of seven data sets is used, that is, five data sets from the Universal Mutation Database (UMD) TP53 database and two datasets from the International Agency for Research on Cancer (IARC) TP53 database, are used to assess the work. Back propagation neural network with hybrid model of 5-fold cross-validation and validation sets was used to classify and predict breast and prostate cancers in patients based on molecular mutations located in the TP53 gene. The performance of the proposed system in the network testing phase was determined to be satisfactory based on the average values for all folds of five indices (i.e., sensitivity = 97 and 96.5; specificity = 96.6 and 97.3; accuracy = 98 and 96.7; *F*-measure = 98.1 and 97.1; and Matthew's correlation coefficient = 0.93 and 0.91) for breast and prostate cancers, respectively.

[In Silico Molecular Classification of Breast and Prostate Cancers using Back Propagation Neural Network. Zahraa Naser Shahweli, Ban Nadeem Dhannoon, Rehab S. Ramadhan**.** *Cancer Biology* 2017;7(3):1-7]. ISSN: 2150-1041 (print); ISSN: 2150-105X (online). http://www.cancerbio.net. 1. doi:10.7537/marscbj070317.01.

Keywords: Classification; Back Propagation Neural Network; Cross Validation; Validation Set; TP53 gene; in silico; Breast Cancer; Prostate Cancer.

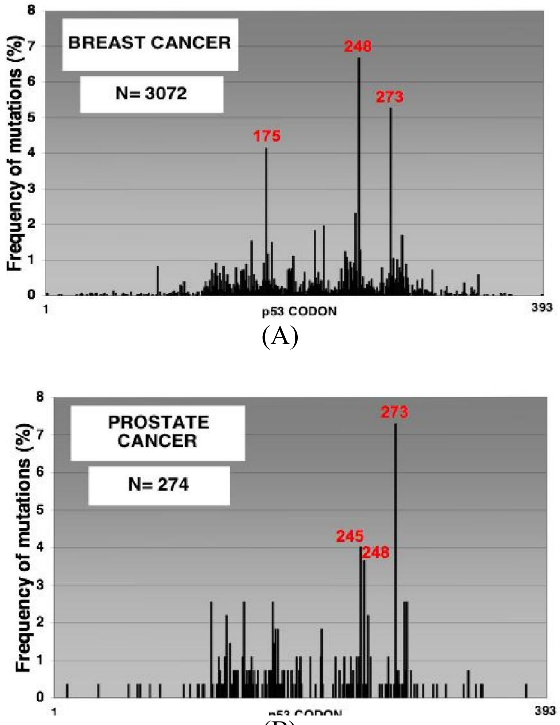
1. Introduction

Breast cancer is the most repetitive obtrusive tumor in women (Altobelli et al., 2017), though prostate growth is the most well-known ailment in men. high rates of dreariness and mortality caused by breast growth and prostate malignancy (Stuart et al., 2004).

Breast and prostate cancers, like different tumors, are illnesses with complex hereditary and biochemical reasons. No single condition, that is, genomic or metabolic, Can be considered as a factor of event and development. In any case, a couple of key players have been distinguished, and among them as hereditary factor, the TP53 tumor suppressor gene is ordinarily transformed in breast, prostate, and different cancers. The P53 protein, encoded by the TP53 tumor suppressor gene, is one of the ace sub-atomic chiefs of stress response in human cells (Walerych et al., 2012) and is included in around half of all individual disease cases (Oren et al., 2016). Affirmed tests demonstrates that mutant p53 coming about because of TP53 mutations has lost its wild-type p53 tumor suppressor activity and adds to harmful progression (Muller and Vousden, 2014). Figure 1 (A and B) demonstrates the P53 changes in breast and prostate malignancies individually (France database of TP53 gene, 2012a).

Examinations of these mutations have been profitable for enhancing learning on the structure–function relationships within the TP53 protein and the high level of heterogeneity of various TP53 mutants in human tumor (Leroy et al., 2014). Therefore, numerous databases contain data on TP53 transformations that prompt tumors. These databases have rich datasets covering a wide range of changes that reason growths. The Universal Mutation Database (UMD) p53 transformation database is one of these databases. The International Agency for Research on Cancer (IARC) TP53 database is likewise utilized as a part of this work. Foreseeing the result of these datasets is a standout amongst the most requesting and fascinating errands in creating information mining applications. With the utilization of automated systems in a computer, bigger volumes of natural information are being gathered and made accessible to medical and biological research gatherings. Accordingly, Knowledge Discovery in Databases, which incorporates information mining procedures, has turned into an exploration instrument for biological and natural specialists to break down and analyze examples and connections among countless put away in vast databases (Gupta et al., 2011).

Information mining systems, for example, machine learning, are utilized to encourage and redesign the procedure of research and prediction. Prediction is a kind of classification utilized as a part of information mining and assumes a noteworthy part in distinguishing fundamental prevention and treating malignancy. The expectation of transformations in genes needs analysis and arrangement, which depend on adequately extensive database to achieve enough right outcomes (Ismaeel and Mikhail, 2016).



(A)

(B)

Figure 1. (A) TP53 mutations in breast cancer (B) TP53 mutations in prostate cancer.

This study displays an in silico neural network (NN) model, which classify breast and prostate tumors in view of the UMD TP53 and IARC TP53 databases.

The in silico NN display is an alluring methodology that has impressive potential for diminishing the quantity of observational examinations sought for choosing and enhancing the achievement rate. Moreover, prediction can be made on virtual situations (Fredsted et al., 2007).

No cancer classification algorithm has yet been viewed as the best and most regular; each algorithm has its own peculiarity and additionally its own particular advantages and disadvantages (YOUSIF et al., 2015). The model proposed in this study is manufactured utilizing one of the classification algorithms, The backpropagation neural network (BPNN) algorithm utilizes 14 features of TP53 mutations to predict breast and prostate cancer risk.

2. Material and Methods

The proposed work consists of two phases, namely, preprocessing and learning phases. In the preprocessing phase, the features in the databases were converted from string to numeric and three features in each database (i.e., mutation position, exon/intron number, and protein variant) were normalized between [0,1]. Thus, extensive estimations of the features couldn't blend with small values. In the learning phase, a BPNN with five fold cross-validation was applied and this technique divides the dataset into five sets, with four of the five datasets used to train the model and the remaining dataset used to test the model. In each fold, training data are split into 80% for training and 20% for validation. Training is stopped when the validation error reached the desired threshold. This procedure is repeated five times where all data is tested. In each fold, the sensitivity, specificity, accuracy, harmonic *F*-measure, and Matthew correlation coefficient (MCC) are reported.

Input Databases

In this work, the performance of the proposed scheme has been evaluated using five datasets from the UMD TP53 mutation database (latest version in 2012) (France database of TP53 gene, 2012b) and two datasets from the International Agency for Research on Cancer (IARC) (latest version in 2016) (International Agency of Research on Cancer, 2016) The names of these datasets are listed in Table 1. From these datasets, 14 features of breast, prostate, and normal tissues were extracted and converted to numeric data as input to the NN. The 14 features from the large features selected based on the proposal of a specialist in biology as a sufficient features to classify cancer. These features contain mutations associated with breast cancer in women and in few men and mutations associated with prostate cancer in men. Each dataset is divided into two datasets, that is, one for breast and normal tissues and the other for prostate and normal tissues, such that binary classification is best for therapeutic/organic application particularly malignancy order and prediction.

Features Selection

The 14 features selected from TP53 UMD and IARC databases explained in table 2.

Table 1: name of the databases used in the proposed work

No.	Database name	Breast/ normal	Prostate/ normal
1	UMDTP53_all_2012_R1_US	4362/1401	334/274
2	UMDTP53_uncurated_2012_R1_US	3460/403	222/40
3	UMDTP53_curated_2012_R1_US	4260/1045	317/128
4	UMDTP53_germline_2012_R1_US	34/38	___
5	UMD_Cell_line_2010	68/64	68/52
6	Germline Mutation Data IARC TP53 Database, R18	268/587	___
7	Somatic Mutation Data IARC TP53 Database, R18	696/252	75/54

Table 2: name of features used in proposed work

No.	Feature name	Feature description	No.	Feature name	Feature description
1	Mutation position	Position of mutation in P53	8	Protein variant	Mutation on protein function
2	Exon	No. of exon in which mutation took place	9	Variation- type	Effect of mutation on variation
3	codon	The no. of codon in which mutation took place	10	Event	Base pair changed in DNA
4	WT codon	Wild type codon (codon before mutation)	11	Type	Type of mutation
5	Mutant codon	Codon with mutation	12	CPG	Effect of mutation on CPG
6	WT AA	Wild type amino acid	13	origin	Source of sample
7	Mutant AA	Mutant amino acid	14	Multiple mutation	No. of mutations in this sample

Model Development Phase

In this phase, visual C# 2010 was used to perform BPNN. The backpropagation algorithm cycles through two distinct passes, that is, a forward pass followed by a backward pass through the layers of the network. The algorithm relays between these passes several times as it scans the training data.

Forward Pass**:** calculating the outputs of all the neurons in the network.

• The algorithm starts with the first hidden layer using the independent variables of a case from the training dataset as input values.

• The neuron outputs are computed for all neurons in the first hidden layer by performing the relevant sum and activation function computations.

• These outputs are the inputs to neurons in the second hidden layer. The relevant sum and activation function computations are again performed to compute the outputs of second layer neurons. The activation function used in this work was the sigmoid function.

Backward Pass**:** propagation of the error and adjustment of weights.

• This phase begins with the computation of the error at each neuron in the output layer. A well-known error function is the squared difference between Ok the output of node k and Yk the target value for that node.

• The target value is only 1 for the output node corresponding to the class of the exemplar and 0 for other output nodes.

• The new value of the weight W_j^k of the connection from node j to node k is derived as:

$W_j^k new = W_j^k old + \eta\ Oj\ \delta_k$,…. (1)

where η is an important tuning parameter that is selected by trial and error by repeated runs on the training data. Typical values for η are in the range 0.1 to 0.9.

• The backward propagation of the weight adjustments along these lines continues until the NN training phase reaches the input layer.

• At this time, a new set of weights will be obtained, from which a new forward pass could be done when presented on 14–8–2 NN with a training data observation (Ganatra et al., 2011).

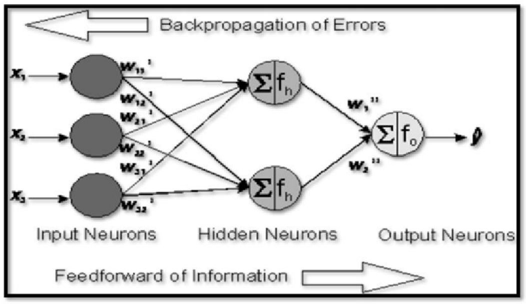


Figure 2. Back propagation architecture (MEKIĆ and MEKIĆ, 2014)

The BPNN algorithm is designed to minimize the mean square error between the desired output and the actual output of the multilayer feed forward perceptron. Figure 2 explain the architecture of back propagation neural network.

BPNN in the proposed work utilized the Fisher–Yates shuffle and random permutation of the input matrix to optimize the work. The Fisher–Yates shuffle (named after Ronald Fisher and Frank Yates) is an algorithm used to generate a random permutation of a finite set.

The mechanism used in the shuffle apply the following merits:

– It is unbiased, such that every permutation is equally likely.

– No additional storage space is needed. It requires only time proportional to the number of items being shuffled. Thus, the method is efficient. Although the algorithm has a dynamic shuffling nature, the implementation enhances the time complexity from O (n^2) to O (n) (Ade-Ibijola, 2012). The steps of Fisher Yates shuffle is showed in figure 3.

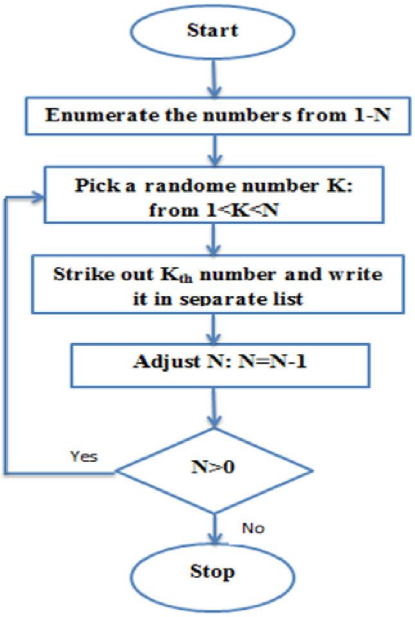

Start
Enumerate the numbers from 1-N
Pick a randome number K: from 1<K<N
Strike out K_th number and write it in separate list
Adjust N: N=N-1
Yes
N>0
No
Stop

Figure 3. steps of fisher yates shuffle(Ade-Ibijola, 2012)

Data Splitting

One of the principle objectives is to manufacture computational models with a high capacity to generalize well the separated information. When preparing BPNN, poor generalization is regularly described by overtraining. A typical technique to abstain from overtraining is the hold-out cross-validation (early stopping). Another model is the *k*-fold cross-validation that uses a blend of more tests to pick up a steady gauge of the model blunder. The dataset T is partitioned into k parts of the same size. One part forms the validation (testing) set T_v, and the other parts form the training set T_{tr}. This process is repeated for each part of the data (Reitermanov, 2010). The proposed work used five fold cross-validation, and in each iteration, one fold is used for testing, and the remaining folds are divided into 80% for training and 20% for validation. Meanwhile, the validation set T_v is periodically used to evaluate the model performance during the training to avoid overtraining. The training is stopped when the performance on T_v is sufficiently good enough or when the last epoch ends. The K-fold cross validation with Hold-out-cross validation algorithm is used in the proposed system as shown in algorithm 1. It separates the dataset T (of size m) into K disjoint subsets, one subset for testing T_t of size m_t and other for training and validation T_{trv} of size m_{trv} then divide T_{trv} into two disjoint, training T_{tr} and validation T_v of sizes m_{tr} and m_v successively. Figure 4 illustrates the proposed mechanism used for 5-folds cross validation and validation set.

Algorithm 1: K-Fold Cross Validation with Hold-Out Cross Validation

```
Input: dataset T, dataset size m, number of folds k.
Output: performance function error.
Begin
Step 1: Divide T into k disjoint subsets $T_1 \ldots T_K$ of the
        same size.
Step 2: For i = 1 to k                                   // k=5
    2.1: $T_t \leftarrow T_i$ , $T_{trv} \leftarrow T - T_i$     //$T_i$ is 1 from 5 folds
         $M_t$ = dataset / k :
         $M_{trv} = m - m_t$
      2.2: Divide $T_{trv}$ into two disjoint subsets $T_{tr}$
           (80%) and $T_v$ (20%).
    2.3: For j=1 to $m_{trv}$
        2.3.1: Train the model $L_j$ on $T_{tr}$
            2.3.2: stop training when error based on $T_v$
                   is satisfied
                   $E_v^j(i) = error(L_j(T_v))$
    2.4: For j=1 to $m_t$
        2.4.1: evaluate the performance of the $K_{th}$ model
               on $T_t$ :
               $E_t^j(i) = error(L_j(T_t))$
Step 3: evaluate the performance of the models by:
        $E = \frac{1}{k}\sum_{i=1}^{k} E_t^j(i)$
End.
```

Performance Measurements

The execution of the proposed scheme is assessed utilizing accuracy (Acc), sensitivity (Sn), specificity (Sp), *F*-measure (harmonic *F*), Matthew correlation coefficient (MCC), and receiver operating characteristic (ROC). These measures are based on the correct and incorrect predicted values of the classifier.

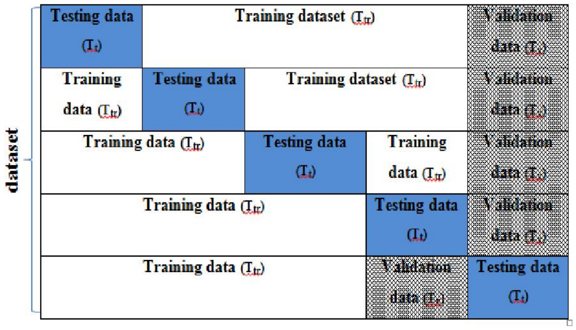


Figure 4. 5-folds cross validation with validation set

– Accuracy is the extent of the quantity of accurately recognized cases in the aggregate number of test cases:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \ldots. (2)$$

where TP is the true positives, TN is the true negatives, FP is the false positives, and FN is the false negatives.

- Sensitivity measures the proportion of positives, which are effectively recognized by the classifier. Numerically, sensitivity is the number of TP results divided by the sum of TP and FN results:

$$SN = \frac{TP}{TP+FN} \ldots\ldots\ldots\ldots (3)$$

- Specificity measures the proportion of negatives, which are effectively recognized by the classifier. Numerically, specificity is the number of TN results divided by the sum of TN and FP results:

$$SP = \frac{TN}{TN+FP} \ldots\ldots\ldots \text{(4)(Zhu et al., 2010)}$$

- *F*-measure is a combination of precision and sensitivity. A high value of *F*-measure shows a high value of precision and sensitivity.

– Precision is the number of TP over the number of TP plus the number of FP:

$$prc = \frac{TP}{TP+Fp} \ldots\ldots\ldots\ldots (5)$$

$$F = \frac{2*PRC*SN}{(PRC+SN)} \ldots\ldots\ldots\ldots \text{(6) (Powers, 2007)}$$

- MCC is a factual measure used to evaluate the nature of learning algorithm. To express a confusion matrix perfectly by a single number, MCC is viewed as one of the best measures on the grounds that different measures, for example, accuracy, are not useful when the dataset is unbalanced. The MCC returns values in the range [−1, 1], where 1 represents a perfect prediction, 0 represents an average random prediction, and −1 represents an inverse prediction:

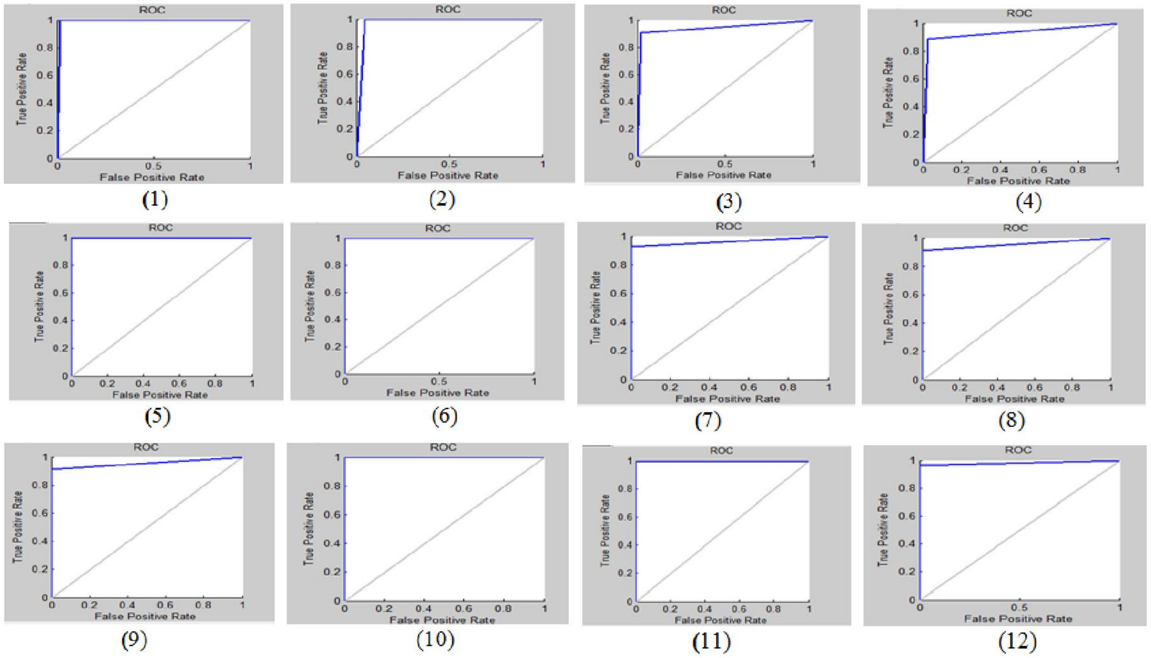
$$MCC = \frac{TP.TN-FP.FN}{\sqrt{(TP+FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)}}$$ (Baldi et al., 2000)

- ROC curves are gainful in the choice of the best classifier under certain choice criteria. The curve represents low values of FPR (1 − Sp) and high values of TP rate (Sn). These values help the points shift toward the upper left corner of the ROC, thus showing better decision. This kind of behavior is desired in applications where the cost of FPR is important (Majid, 2006).

3. Results

Table 3. Result of all data sets

No	Dataset name	Sn	Sp	Acc	F- measure	MCC	Run time (m)
1	UMDTP53all/breast	99.9	95.6	98.9	99.2	0.95	4:30
2	UMDTP53all/prostate	100	94.1	97.3	97.6	0.9	0:28
3	UMDTP53uncurated /breast	99.0	81.1	97.1	98.4	0.82	3:02
4	UMDTP53uncurated /prostate	91.5	92.7	91.9	93.0	0.79	0:46
5	UMDTP53curated /breast	99.8	100	99.8	99.9	0.99	1:20
6	UMDTP53curated /prostate	100	100	100	100	1	0:12
7	UMDTP53germline /breast	90.1	100	95.7	94.6	0.91	0:10
8	UMD Cell_line /breast	91.2	100	95.3	95.2	0.91	0:14
9	UMD Cell_line /prostate	93.6	100	95.8	96.6	0.91	0:08
10	Germline IARC /breast	99.6	100	99.8	99.8	0.99	0:21
11	Somatic IARC /breast	99.7	100	99.7	99.8	0.99	0:04
12	Somatic IARC /prostate	97.5	100	98.5	98.7	0.96	0:02
Average value of breast cancer		97	96.6	98	98.1	0.93	
Average value of prostate cancer		96.5	97.3	96.7	97.1	0.91	



(1) (2) (3) (4)

(5) (6) (7) (8)

(9) (10) (11) (12)

Figure 4. ROC curves for all datasets in table 3 from (1) to (12).

A total of 14 features from 7 datasets of TP53 gene databases are used. The results are computed using five fold cross-validation with validation set in each fold technique for these datasets. The average values for all fold of each dataset were calculated and tabulated in Table 3.

The average values of accuracy and *F*-measure for breast cancer are 98% and 98.1%, respectively. The average values for accuracy and *F*-measure for prostate cancer are 96.7% and 97.1%, respectively. Figure 4 show that ROC curve for all datasets in table 3 ordered from 1 to 12.

4. Discussion and Conclusion

Breast cancer, as other malignancy, is related with various sorts of somatic hereditary mutations, such as, transformations in oncogenes and tumor suppressor genes. The most continuous positions of gene changes are in the TP53 gene with around 20-30% of Breast malignancies, based on tumor size and stage of the tumor, it would be expected that p53 would be a useful biomarker for the prediction of Breast tumorigenesis (Bertheau et al., 2013).

While prostate cancer is the second tumor as far as overall occurrence among men (Chen and Zhao, 2013). as a model, Mutation databases for the TP53 gene are used, it's contain largest collection of somatic mutations or germline mutations.

The objective of this study is to create an effective neural model to perform the acceptable classification and prediction of breast and prostate cancers from large databases such as UMD and IARC TP53 based on these mutations.

As appeared in Table 3, BPNN gives the least time for learning and testing, this appear in the last column in table 3.

Values of F- measure recorded in the table 3 indicates that the network involved an acceptable level of reliability in classifying the cases. Also the system executed in this study is more productive than other artificial neural systems because of its high performing velocity and great generalizability. MCC used to assess the work explain that positive prediction and in some time perfect prediction also sensitivity and specificity gave high values, this is very clear in ROC curve plotted for each database where ROC curve depending on sensitivity and (1- specificity). One of the reasons for the high sensitivity and specificity of the network in this study could be credited to the determination of the fitting elements and the suitable choice of features and system Structural.

Acknowledgements:

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

Corresponding Author:
Zahraa Naser Shahweli
Department of Computer Science
College of Science, Al-Nahrain University Baghdad, 10006, Iraq
Telephone: 771-274-8183

E-mail: Stcs-zns16@sc.nahrainuniv.edu.iq

References

1. Ade-Ibijola AO. A Simulated Enhancement of Fisher-Yates Algorithm for Shuffling in Virtual Card Games using Domain-Specific Data Structures. International Journal of Computer Applications. 2012;54.
2. Altobelli E, Rapacchietta L, Angeletti PM, Barbante L, Profeta FV, Fagnano R. Breast Cancer Screening Programmes across the WHO European Region: Differences among Countries Based on National Income Level. Int J Environ Res Public Health. 2017;14:452.
3. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16:412-24.
4. Bertheau P, Lehmann-Che J, Varna M, Dumay A, Poirot B, Porcher R, et al. p53 in breast cancer subtypes and new insights into response to chemotherapy. The Breast. 2013;22: S27-S9.
5. Chen F-z, Zhao X-k. Prostate cancer: current treatment and prevention strategies. Iranian Red Crescent medical journal. 2013;15:279.
6. France database of TP53 gene. UMD TP53. 2012a.
7. France database of TP53 gene. UMD TP53. 2012b.
8. Fredsted B, Brockhoff PB, Vind C, Padkjær SB, Refsgaard HH. In Silico Classification of Solubility using Binaryk-Nearest Neighbor and Physicochemical Descriptors. QSAR & Combinatorial Science. 2007;26:452-9.
9. Ganatra A, Kosta Y, Panchal G, Gajjar C. Initial classification through back propagation in a neural network following optimization through GA to evaluate the fitness of an algorithm. International Journal of Computer Science and Information Technology. 2011;3:98-116.
10. Gupta S, Kumar D, Sharma A. Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian Journal of Computer Science and Engineering (IJCSE). 2011;2:188-95.
11. International Agency of Research on Cancer. IARC TP53 database. 2016.
12. Ismaeel AG, Mikhail DY. Effective Data Mining Technique for Classification Cancers via Mutations in Gene using Neural Network. arXiv preprint arXiv:160802888. 2016.
13. Leroy B, Anderson M, Soussi T. TP53 mutations in human cancer: database reassessment and prospects for the next decade. Hum Mutat. 2014;35:672-88.
14. Majid A. Optimization and Combination of Classifiers Using Genetic Programming: Ghulam Ishaq Khan Institute of Engineering Sciences & Technology, Swabi; 2006.
15. MEKIĆ E, MEKIĆ E. Application of Ann in Australian Credit Card Approval. Teacher education and professional development. 2014.
16. Muller PA, Vousden KH. Mutant p53 in cancer: new functions and therapeutic opportunities. Cancer cell. 2014;25:304-17.
17. Oren M, Tal P, Rotter V. Targeting mutant p53 for cancer therapy. Aging (Albany NY). 2016;8:1159-60.
18. Powers DM. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, School of Informatics and Engineering, Flinders University, Adelaide, Australia. TR SIE-07-001, Journal of Machine Learning Technologies 2: 1 37-63. https://dl-web. dropbox. com/get/Public/201101-Evaluation_JMLT_Postprint-Colour. pdf; 2007.
19. Reitermanov Z. Data splitting. WDS2010. p. 31-6.
20. Stuart RO, Wachsman W, Berry CC, Wang-Rodriguez J, Wasserman L, Klacansky I, et al. In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proc Natl Acad Sci U S A. 2004;101:615-20.
21. Walerych D, Napoli M, Collavin L, Del Sal G. The rebel angel: mutant p53 as the driving oncogene in breast cancer. Carcinogenesis. 2012;33:2007-17.
22. YOUSIF SA, SAMAWI VW, ELKABANI I, ZANTOUT R. Enhancement of Arabic Text Classification Using Semantic Relations with Part of Speech Tagger. W transactions Advances In Electrical And Computer Engineering. 2015:195-201.
23. Zhu W, Zeng N, Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland. 2010:1-9.

7/24/2017