International Research Journal of Computer Science (IRJCS)
Issue 12, Volume 4 (December 2017)
ISSN: 2393-9842
www.irjcs.com

NEURAL NETWORK WITH NEW RELIEF FEATURE SELECTION FOR PREDICTING BREAST CANCER BASED ON TP53 MUTATION

Zahraa N. Shahweli
Computer Science Department, College of Science,
Al-Nahrain University, Baghdad, Iraq
Stcs-zns16@sc.nahrainuniv.edu.iq

Ban N. Dhannoon
Computer Science Department, College of Science,
Al-Nahrain University, Baghdad, Iraq
bnt@sc.nahrainuniv.edu.iq

Manuscript History
Number: **IRJCS/RS/Vol.04/Issue12/DCCS10080**
DOI: 10.26562/IRJCS.2017.DCCS10080
Received: 08, November 2017
Final Correction: 23, November 2017
Final Accepted: 02, December 2017
Published: December 2017
Citation: Shahweli, Z. N. & Dhannoon, B. N. (2017). NEURAL NETWORK WITH NEW RELIEF FEATURE SELECTION FOR PREDICTING BREAST CANCER BASED ON TP53 MUTATION. **International Research Journal of Computer Science, Volume IV, 07-12. doi: 10.26562/IRJCS.2017.DCCS10080**
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract- TP53 gene is an effective predictor of all cancer types, including breast cancer. The use of neural network algorithms facilitates the detection and diagnosis of cancer through its training on mutations that occur in genes, such as the TP53 gene. Moreover, feature selection techniques are important for detecting the necessary training features. In this work, an improved method of the Relief algorithm called ReliefK was proposed and yielded satisfactory results. This method also used with Back Propagation Neural Network (BPNN) to predict and classify cancer depending on the mutations that were recorded in the International Agency for Research on Cancer (IARC) TP53 database (somatic and germline mutations). Five measures, including sensitivity (Sn), specificity (Sp), accuracy (Acc), F-measure, and Matthew correlation coefficient (MCC), were used to evaluate the work. The proposed method of feature selection ReliefK and BPNN yielded MCC of 1 and 0.88 for IARC TP53 somatic and germline mutations, respectively.

Keywords- BPNN; Breast Cancer; IARC TP53 Database; Relief Feature Selection; ReliefK Feature Selection;

I. INTRODUCTION

Breast cancer remains one of the most common cancers in women. It was found to be similar to other cancers considering genetic and biochemical reasons. Among these reasons is that mutations occur in genes, and among these genes is the TP53 tumor suppressor gene [1].Studies and experiments have shown that TP53 is responsible for encoding tumor suppressor protein, and medical experiences have shown that the inactivity of this suppressor protein is the cause of most cancers.

Thus, TP53 gene and its P53 protein have a major role in predicting cancer, as well as identifying early treatment and tumor inhibition[2]. Numerous sites provide large databases of mutant TP53 gene, which have been used in many scientific studies on somatic and germline mutations. The International Agency for Research on Cancer (IARC) database is one of these databases. Managing these databases is difficult because they require storage, processing, and analysis[3].

Accordingly, these vast amounts of data are used by neural network techniques to recognize and observe cancer. In the present work, Back Propagation Neural Network (BPNN) was used to predict and prognosis breast cancer based on TP53 mutations contained in the IARC TP53 somatic mutation and IARC TP53 germline mutation databases. The most broadly utilized neural network today is BPNN because it has a few advantages, such as potent ability of classification, fast and proper discrimination [4]. The neural network in this work learned on data by using five-fold cross-validation. Cross-validation is a practical technique for assessing and contrasting the performance of learning algorithms. That is, cross-validation is used to quantify the generalizability of algorithms [5]. To determine that the preferable subset of features has the least number of dimensions that affect accuracy [6], the new feature selection method called ReliefK algorithm was used. It is an update of the Relief and ReliefF algorithms. Three types of feature selection techniques are filters, wrappers, and embedded techniques, which have an interactive role with machine learning algorithms [7]. ReliefK algorithm is of a filtering method type, where it separated from the applied machine learning algorithm.

II. MATERIAL AND METHODS

A. Input Data Base

In this work, the performance of the proposed scheme was evaluated using the IARC TP53 somatic mutation and IARC TP53 germline mutation databases (latest version in 2016; IARC, 2016). The IARC TP53 somatic mutation database contains 67 features and 29895 records for all types of cancer, and the IARC TP53 germline mutation database contains 56 features and 3173 records for all types of cancer. This work extracted 14 features and the records related with breast cancer and normal samples from each database.New Relief feature selection was used as preprocessing step for minimizing the 14 features before training them by BPNN.

B. Proposed Work

The proposed work has two phases, namely, preprocessing and learning.
The preprocessing phase comprises the following three steps.

1. Balance the minority class (normal patients).

This step is important due to few numbers of records of normal samples compared with that of patients with breast cancer. Therefore, the number of records of normal samples in somatic mutation database was increased manually from 54 to 252, whereas the number of records of patients with breast cancer is 696. The germline mutation database has 587 records for normal samples and 268 records for patients with breast cancer; thus, balancing for this database was not performed.

2. The normalization of all features was performed as follows:

$$\frac{dF-MinF}{MaxF-MinF} \quad (1)$$

where *dF* is the feature value; *MinF* and *MaxF* are the minimum and maximum values that the feature ϕ can obtain, respectively; and ϕ is the normalized value of feature ϕ.

3. Use the new Relief feature selection.

The proposed ReliefK algorithm used in this work is only an update of the original Relief algorithm and is simpler than ReliefF algorithm; however, it is only used with binary problems similar with the original Relief. The new proposed Relief algorithm called ReliefK randomly selects an instance r_i and then searches for the K-nearest neighbor for the same class called Hit h_j and the K-nearest neighbor for the second class called Miss m_j. ReliefK updates the quality estimation W [a] for all attributes a depending on their values for r_i, *m*, and *h*. ReliefK is demonstrated in Algorithm 1.

Algorithm 1: ReliefK Algorithm for feature selection

Input: All training data
Output: Weight for each feature
Begin
Set W[a]=0 for each attribute a

For i=1 **to** m **do** // m: training data size
Select sample r_i from training data at random
For j=1 **to** k **do**// k: no. of hit and miss used to update weight in each step

Find nearest hit h_j and nearest miss m_j
For a=1 **to** A **do**// A: no. of features
dif1 = dif1+ diff(a,r_i,h_j) //dif1 and dif2 initially zero
dif2 = dif2+ diff(a,r_i,m_j)
End for
dif1 = dif1/(m.k)
dif2= dif2/(m.k)
W[a]=W[a]+ dif2- dif1
End for
End for

The proposed algorithm provided better results compared with the original Relief algorithm and its extension ReliefF algorithm when performing on the IARC TP53 somatic mutation and germline mutation databases, where K in this study was suggested to be 10. After completing the preprocessing steps, the learning phase with BPNN begins. BPNN used five-fold cross-validation to perform enhanced learning and generalization. Several performance measures, including accuracy, F-measure, Matthew correlation coefficient (MCC), sensitivity and specificity (for plotting an ROC curve), were used to evaluate the work.

III. RESULTS AND DISCUSSION

A. Results before Balancing

The results of learning BPNN with five-fold cross-validation on 14 features selected from each database before balancing on minority class are shown in Table 1, where UD denotes undefined numbers.

TABLE 1.TheResult of the IARC TP53 Somatic Mutation Database before Balancing.

Database Name	Fold no.	Sn.	Sp.	Acc.	F-measure	Mcc
IARC TP53 Somatic mutation	1	100	100	100	100	1
	2	100	0.00	92	95.83	UD
	3	100	0.00	90	94.73	UD
	4	100	100	100	100	1
	5	99.29	100	99.33	99.64	94.53
	Average	99.85	60	96.26	98.04	UD

From Table 1, acceptable results were obtained when sensitivity, accuracy, and F-measure were considered because these measures depend on the majority class (positive). By contrast, specificity depends on the minority class (negative), thereby obtaining conflicting results. Nevertheless, MCC is the best measure for binary classification when the majority and minority classes have similar importance due to its equation, which depends on all the factors of confusion matrix (true positive, true negative, false positive, and false negative); MCC does not work correctly if the data are unbalanced.

B. Results after Balancing

The results of learning with BPNN using five-fold cross-validation on selected databases before using feature selection and after balancing on minority class are shown in Table 2.

TABLE 2. The Result of the IARC TP53 Somatic Mutation Database after Performing Balancing and before Using Feature Selection Algorithm.

Database Name	Fold no.	Sn	Sp.	Acc.	F-measure	Mcc
IARC TP53 Somatic mutation	1	99.29	100	99.47	99.64	0.98
	2	100	100	100	100	1
	3	100	100	100	100	1
	4	100	100	100	100	1
	5	99.31	100	99.48	99.65	0.98
	Average	99.72	100	99.79	99.86	0.99

From Table 2,one can conclude the effect of balancing between two classes in binary classification results.

C. Results of Using ReliefK Algorithm

Before learning with BPNN, ReliefK feature selection was used to minimize the number of features. First, original Relief algorithm was performed, followed by ReliefF algorithm. The results for both algorithms were compared with ReliefK to detect the best algorithm for binary classification. Table 3 shows the results of BPNN with five-fold cross-validation after Relief algorithm removed 1 feature from the 14 features of the databases. Table 4 shows the results after ReliefF algorithm removed 4 features from the 14 features of the databases.

TABLE 3. Results after Performing Relief Algorithm

Database Name	Fold no.	Sn	Sp.	Acc.	F-measure	Mcc
IARC TP53 Somatic mutation	1	100	100	100	100	1
	2	100	100	100	100	1
	3	100	100	100	100	1
	4	100	100	100	100	1
	5	99.3	100	99.47	99.65	0.98
	Average	99.86	100	99.89	99.93	0.99
IARC TP53 Germline mutation	1	92.59	94.01	93.56	90.09	0.73
	2	98.18	98.27	98.24	97.29	0.92
	3	98.14	94.87	95.9	93.8	0.80
	4	96	95.04	95.32	92.3	0.78
	5	98.18	96.55	97.07	95.57	0.86
	Average	96.62	95.75	96.02	93.81	0.82

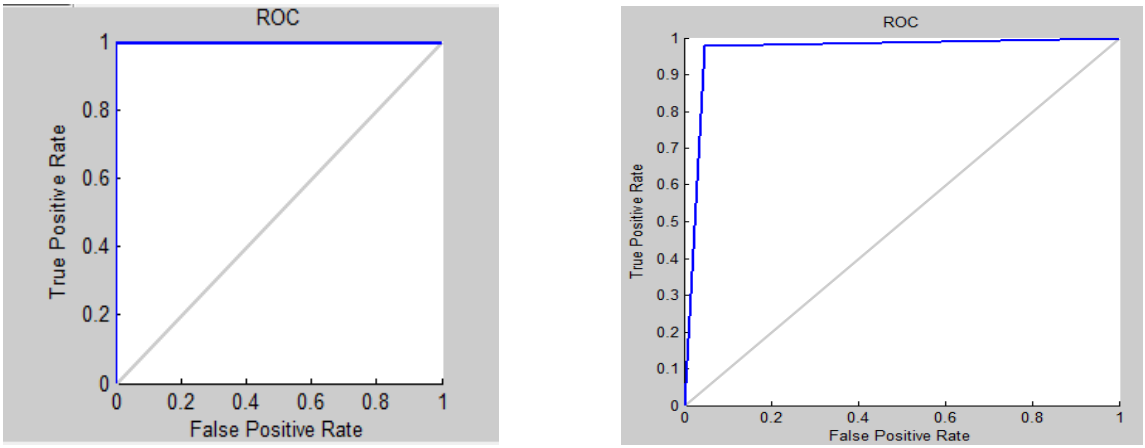
TABLE 4. Results after Performing ReliefF Algorithm

Database Name	Fold no.	Sn	Sp.	Acc.	F-measure	Mcc
IARC TP53 Somatic mutation	1	99.29	100	99.47	99.64	0.98
	2	100	100	100	100	1
	3	100	100	100	100	1
	4	100	100	100	100	1
	5	99.3	100	99.47	99.65	0.98
	Average	99.71	100	99.79	99.85	0.99
IARC TP53 Germline mutation	1	90.74	90.59	90.64	85.96	0.62
	2	87.27	97.41	94.15	90.56	0.81
	3	90.74	94.01	92.98	89.09	0.72
	4	96	88.42	90.64	85.71	0.56
	5	89.09	96.55	94.15	90.74	0.79
	Average	90.76	93.4	92.51	88.41	0.70

Tables 3 and 4 show acceptable results and prove the efficiency of Relief and ReliefF algorithms. However, the proposed ReliefK algorithm providedmore satisfactory results than both previous algorithms after removing one feature that is different from the one that Relief and ReliefF removed(Table 5). AReceiver Operating Characteristic (ROC) curve wassketched for the results of BPNN after performing the ReliefK algorithm, as shown in Figure 1.

TABLE5. Results after Performing New Relief Algorithm (ReliefK).

Database Name	Fold no.	Sn.	Sp.	Acc.	F-measure	Mcc
IARC TP53 Somatic mutation	1	100	100	100	100	1
	2	100	100	100	100	1
	3	100	100	100	100	1
	4	100	100	100	100	1
	5	100	100	100	100	1
	Average	100	100	100	100	1
IARC TP53 Germline mutation	1	92.59	99.14	97.07	95.23	0.91
	2	85.45	99.13	94.73	91.26	0.86
	3	100	98.29	98.83	98.18	0.93
	4	96	96.69	96.49	94.11	0.84
	5	96.36	96.55	96.49	94.64	0.85
	Average	94.08	97.96	96.72	94.68	0.88



(A) (B)

Fig 1. ROC curve for BPNN with ReliefK algorithm, (A) for IARC TP53 somatic mutation database, (B) for IARC TP53 germline mutation database.

IV. CONCLUSIONS

The following are the results obtained from this study:

1. Balancing between minority and majority class has a considerable effect on the results of the classification algorithm.
2. Normalization for all features is a mandatory step before performing all types of Relief algorithm.
3. The proposed ReliefK algorithm proved its effectiveness through the obtained results, which were more satisfactory than the results of Relief and ReliefF.
4. BPNN with five-fold cross-validation works well with the proposed ReliefK algorithm in predicting and classifying breast cancer based on mutations stored in the IARC TP53 somatic and germline mutation databases.

ACKNOWLEDGMENT

Teaching staff of the Faculty of Science, Department of Computer Science at the University of Al Nahrain thanks and gratitude to give useful scientific advice to complete work on what it is.

REFERENCES

1. 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.
2. Maurici D, Hainaut P. TP53 gene and p53 protein as targets in cancer management and therapy. BIOTECHNOLOGY. 2001;12.
3. Edlund K, Larsson O, Ameur A, Bunikis I, Gyllensten U, Leroy B, et al. Data-driven unbiased curation of the TP53 tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors. Proceedings of the National Academy of Sciences. 2012;109:9551-6.
4. Kaensar C. Analysis on the parameter of back propagation algorithm with three weight adjustment structure for hand written digit recognition. Service Systems and Service Management (ICSSSM), 2013 10th International Conference on: IEEE; 2013. p. 18-22.
5. Kale S, Kumar R, Vassilvitskii S. Cross-validation and mean-square stability. In Proceedings of the Second Symposium on Innovations in Computer Science (ICS2011: Citeseer; 2011.
6. Sewell M. Feature selection. Online] http://machine-learning martinsewell com/feature-selection. 2007.
7. Stańczyk U, Jain LC. Feature selection for data and pattern recognition: Springer; 2015.

