American J. of Engineering and Applied Sciences 4 (4): 576-585, 2011
ISSN 1941-7020

Cross Validation Evaluation for Breast Cancer Prediction Using Multilayer Perceptron Neural Networks

[1]Shirin A. Mojarad, [1]Satnam S. Dlay,
[1]Wai L. Woo and [1,2]Gajanan V. Sherbet
[1]School of Electrical, Electronic and Computer Engineering,
Newcastle University, England, UK
[2]Institute for Molecular Medicine, Huntington Beach, CA, USA

Abstract: Problem statement: The presence of metastasis in the regional lymph nodes is the most important factor in predicting prognosis in breast cancer. Many biomarkers have been identified that appear to relate to the aggressive behaviour of cancer. However, the nonlinear relation of these markers to nodal status and also the existence of complex interaction between markers have prohibited an accurate prognosis. **Approach:** The aim of this study is to investigate the effectiveness of a Multilayer Perceptron (MLP) for predicting breast cancer progression using a set of four biomarkers of breast tumors. The biomarkers include DNA ploidy, cell cycle distribution (G0G1/G2M), steroid receptors (ER/PR) and S-Phase Fraction (SPF). A further objective of the study is to explore the predictive potential of these markers in defining the state of nodal involvement in breast cancer. Two methods of outcome evaluation viz. stratified and simple k-fold Cross Validation (CV) are studied in order to assess their accuracy and reliability for neural network validation. Criteria such as output accuracy, sensitivity and specificity are used for selecting the best validation technique besides evaluating the network outcome for different combinations of markers. **Results:** The results show that stratified *2*-fold CV is more accurate and reliable compared to simple k-fold CV as it obtains a higher accuracy and specificity and also provides a more stable network validation in terms of sensitivity. Best prediction results are obtained by using an individual marker-SPF which obtains an accuracy of 65%. **Conclusion/Recommendations:** Our findings suggest that MLP-based analysis provides an accurate and reliable platform for breast cancer prediction given that an appropriate design and validation method is employed.

Key words: Breast cancer, k-fold cross validation, Multilayer Perceptron (MLP), predictive analysis

INTRODUCTION

Breast cancer has been identified as the most widespread cancer amongst women and also the major cause of female cancer death all over the world (Etchells and Lisboa, 2006). An important factor influencing the breast cancer mortality rate is the efficacy of treatment intervention which in turn is influenced by the stage and accuracy of prognosis. Hence, accurate prognosis in patients with early stage breast cancer is of significant importance to reduce mortality rate.

Several prognostic factors including patient age, tumor size, tumor grade, DNA content (ploidy) and receptor status have been identified for nodal metastasis prediction with the hope to avoid axillary lymph node dissection (Lyman *et al*., 2005). However, no individual or combination of these prognostic factors has replaced nodal dissection for node status determination (Giuliano *et al*., 1997).

Amongst prognostic markers, those that can be obtained via minimally invasive methods are preferred for determining nodal status and survival prediction so to minimize patient morbidity along with mortality. Several studies have investigated different prognostic factors in an effort to define the prognostic value of these markers and find an optimal combination of markers which can be used as an accurate and reliable predictor for breast cancer prognosis. However, the complex interaction of these markers with nodal status and survival rate besides the existence of inter-relation between the markers has prevented accurate predictions using these markers.

Multivariate statistical methods have been widely used to investigate the prediction significance of prognostic factors. These multivariate models mainly include logistic regression (Hosmer and Lemeshow, 2000). However, there are several inadequacies in these

Corresponding Author: Shirin A. Mojarad, School of Electrical, Electronic and Computer Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom Fax: +44 (0) 191 222 81802

methods which present doubts in their reliability. The study conducted by Concato *et al*. (1993) on the deficiencies of these statistical methods has investigated the present problems of multivariate analysis in medical research. Some of the reported problems include over fitting of data, not considering the inter-relation between markers and unknown method of selection among candidate markers which necessitates the need for improvement in medical research using these multivariate statistical methods. Multivariate regression methods are also prone to over-optimistic results which lead to misleading interpretation in defining the prognostic value of the investigated markers (Altman and Lyman, 1998).

Another approach that has been widely used for the aim of cancer prognosis is Artificial Neural Network (ANN) (Schwarzer *et al*., 2000; Ahmed, 2005; Kaur and Wasan, 2006; Ashidi *et al*., 2007). ANNs are parallel processing structures consisting of basic processing units (neurons) which are interconnected by weighted links. ANNs have the ability to learn patterns existing in data and hence perform classification and prediction for new data. There are different types of ANNs depending on their structure and learning process. The connections between neurons can be formed in different directions. In feed forward ANNs, all connections are set up in one direction from network's input towards the output. In addition, the learning process can be supervised or unsupervised depending on whether the input data is associated with known outputs during learning or not.

ANN has been confirmed as a robust method for the aim of cancer prognosis (Burke *et al*., 1994). It is also superior to conventional methods employed for breast cancer prediction such as Tumor, Node, Metastasis (TNM) staging system and logistic regression (Burke *et al*., 1997). One of the main advantages of ANNs over conventional methods is their ability in capturing the complex and nonlinear interaction between prognostic markers and the outcome to be predicted. They also enable taking into account the inter-relation between markers which can significantly improve the prognosis in oncology.

An ANN can have different structures based on the type of its input-output data and also its application. Among available structures, Multilayer Perceptron (MLP) has been widely used for the aim of cancer prediction and prognosis (Schwarzer *et al*., 2000). MLP is a class of feed forward neural networks which is trained in a supervised manner to become capable of outcome prediction for new data (Haykin, 2009).

In this study, three cellular markers including DNA ploidy, S-Phase Fraction (SPF) and cell cycle distribution in addition to a molecular marker-the state of steroid receptors including Estrogen and Progesterone Receptors (ER/PR) have been employed for nodal status prediction in breast cancer. The aim of the study is to employ a MLP neural network as a platform to predict the state of nodal involvement based on the four cellular and molecular biomarkers. This study also investigates the predictive accuracy of individual biomarkers in order to define their impact on outcome prediction in breast cancer. Besides, the relation between the mentioned cellular and molecular markers will be explored. We will also illustrate the capability of MLP in capturing both the linear and nonlinear relationship between the above markers and breast cancer outcome. In addition, the efficiency of stratified and simple k-fold Cross Validation (CV) in validating the MLP outcome for cancer prediction is investigated.

The study is organized as follows: the next section explains the breast cancer dataset used in this study and the roles of the biomarkers. Materials and methods include the MLP structure employed for cancer prediction, the validation method for assessing the designed network and also a brief description of Pearson's correlation coefficient which its results are later used to compare and validate those results obtained by the MLP. Following that, results and discussion are elaborated. Finally, the findings of the study are presented in the conclusion.

Breast cancer dataset: The data utilised for nodal involvement analysis contains the information corresponding to four cellular and molecular breast tumor biomarkers pertaining to 46 patients who had been diagnosed with a carcinoma or benign breast tumor. The biomarkers include DNA ploidy, cell cycle distribution (G0G1/G2M), Steroid Receptors (ER/PR) and S-Phase Fraction (SPF). Nodal status in terms of cancer metastasis to regional lymph nodes has been defined as an outcome for all 46 patients.

DNA aneoploidy is a state in which abnormal sets of chromosomes exist within the nucleus and is considered as an indicator of tumor malignancy. The degree of DNA ploidy is calculated based on the Integrated Nuclear Density (IND) measurement which is obtained by staining the aspirated tumor cells. Many studies have investigated the role of DNA ploidy of cancer cells in cancer prognosis. The results demonstrate that this marker is highly associated with relapse of the disease (Yuan *et al*., 1992), reduced survival time (Azua *et al*., 1997), metastasis to regional lymph nodes and early death (Gilchrist *et al*., 1993). In addition, aneuploidy has been identified as a significant prognostic biomarker for breast, prostate and endometrial cancer prognosis (Moureau-Zabotto *et al*., 2005; Suehiro *et al*., 2008; Pretorius *et al*., 2009).

However, some studies have found DNA ploidy uncorrelated with breast cancer prognosis (Naguib *et al*., 1999). Moreover, some studies suggest DNA ploidy as a consequence of premature 3 cells entering the S-phase and therefore the close correlation between aneoploidy and size of SPF. Nevertheless, Sherbet and Lakshami (Naguib and Sherbet, 2001) have found them totally uncorrelated.

The pattern of cell cycle distribution is defined by the G0G1/G2M ratio (ratio of the number of the cells in G0G1 phase over the number of the cells in G2M phase) which is measured by ICM (Anderson *et al*., 2003). The fraction of cycling cells in the tumor has proven to be an effective factor in the response of the carcinoma to chemotherapy (Remvikos *et al*., 1989). The size of the proliferative fraction is also known to be a good prognostic feature (Kallioniemi *et al*., 1988). Cell cycle distribution can be measured from DNA profiles derived from flow cytometry which is also considered as a reliable method for SPF measurement (Naguib *et al*., 1999). The reliability of SPF estimation depends upon the differentiation clarity of the G0G1 and G2M parts of the cell cycle distribution diagram.

In many studies, it has been proved that the status of hormone receptors of breast cancer cells can be used as useful information for cancer prognosis and treatment (Anderson *et al*., 2003; Grey *et al*., 2003; Esteva and Hortobagyi, 2004). The steroid receptors considered in this study include Estrogen Receptor (ER) and Progesterone Receptor (PR). Estrogen is a hormone with growth stimulating ability in a variety of target tissues. It binds to estrogen receptors which are then transmitted to the nucleus where they instigate responsive genes transcription and lead to appropriate psychological function. Estrogen and progesterone hormones can initiate the transcription of some target genes related with cell differentiation and proliferation (Phippard *et al*., 1996).

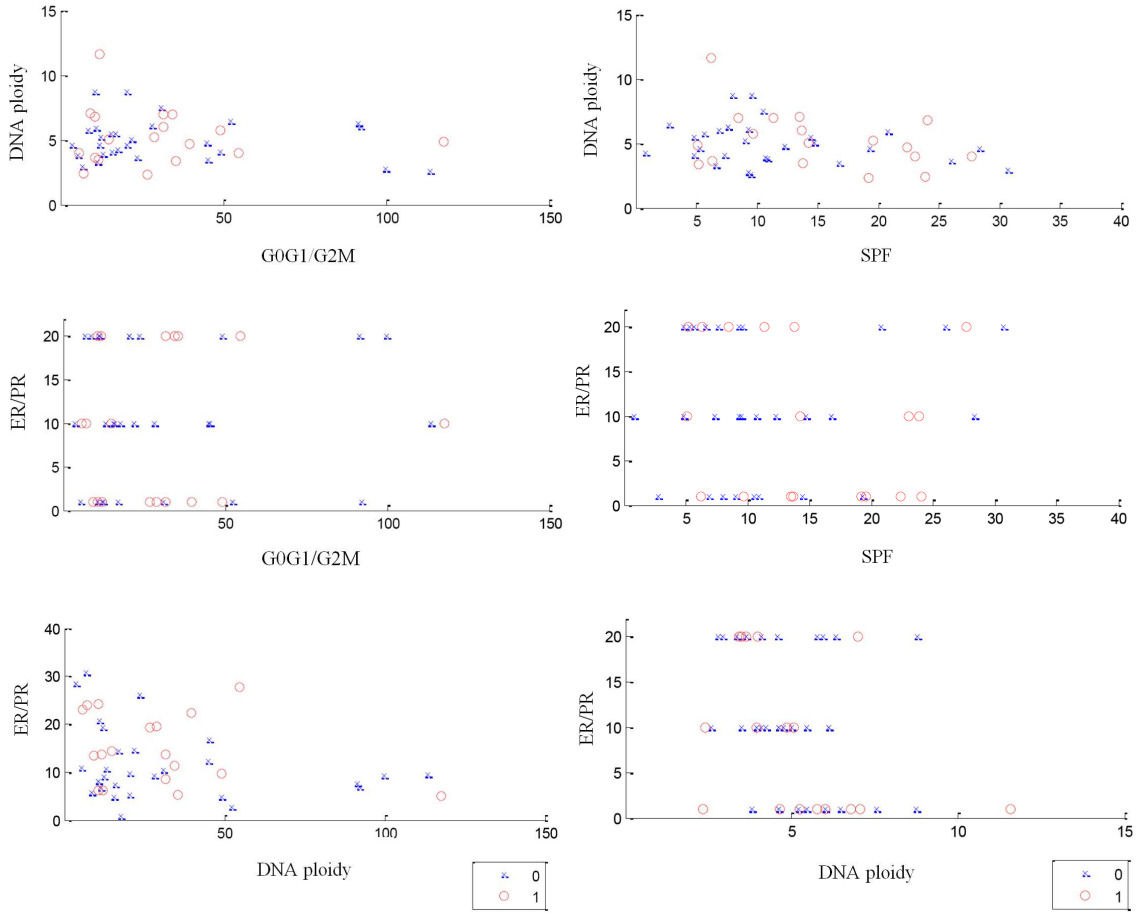

Fig. 1: Scatter plot of the data feature vectors dichotomized by output groups (data with nodal status= 0 are marked by “x” and data with nodal status = 1 are marked by “o”)

Table 1: Descriptive statistics for continuous markers used in this study

	Mean	Standard deviation	Minimum	Maximum	Range
DNA Ploidy	5.094	1.841	2.33	11.58	9.25
SPF	12.830	7.550	0.76	30.70	29.94
G0G1/G2M	31.000	29.090	3.56	117.60	114.04

Table 2: Descriptive statistics for discrete markers and output used in this study

	Values	Proportion of value 0 (%)	Proportion of value 1 (%)	Proportion of value 2 (%)
DNA Ploidy	0, 1, 2	35	30	35
SPF	0, 1	57	43	-

Tumors that are receptor positive respond well to treatment with anti-estrogens. So the absence of ER in breast cancer is considered as a sign of poor prognosis since these patients cannot benefit from anti-estrogen therapy. ER absence in breast cancer is caused by ER gene silencing resulting from hypermethylation (Grey *et al*., 2003). The role of the PR positivity in breast cancer is less significant. Normally, ER positive cancers are also PR positive, but there would be a poor prognosis for PR positive tumors that are not ER positive.

The size of the SPF indicates the percentage of cells in the stage of DNA replication in cell cycle and it is a validated marker for estimating the proliferative rate of tumor cells (Clark *et al*., 1989). SPF is also recognized as an independent prognostic factor in breast cancer (Bae *et al*., 2007; Gazic *et al*., 2008).
A complete procedure of SPF measurement is described by Naguib *et al*. (1999).

Except for ER/PR, which takes discrete values, other markers are continuous within different ranges. Nodal status is defined as either 0 or 1 for the case of no node involved or metastasis to the regional lymph nodes, respectively. Table 1 and 2 show some descriptive statistics for continuous and discrete markers respectively.

All the mentioned markers are established as effective markers in breast cancer prognosis in medical context. However, the efficiency of the combination of these markers and also their inter-relation is further investigated in this study. In addition, the data feature vectors for the two output groups are plotted in the form of scatter plots in Fig. 1 Each scatter plot in Fig. 1 shows two feature vectors on two axes with the two output groups shown by “x” and “o” for no nodal metastasis and nodal metastasis respectively.

The scatter plots in Fig. 1 show that the data feature vectors are not linearly separable in the 2-dimensional space.

METERIALS AND METHODS

MLP: ANNs are a class of artificial intelligence methods commonly used for classification and pattern recognition. A MLP is a type of ANN which consists of a set of interconnected artificial neurons connected only in a forward manner to form layers. One input, one or more hidden and one output layer are the layers forming a MLP. A MLP with one hidden layer and its connections is illustrated in Fig. 2.

An artificial neuron is the basic processing element of a neural network, which consists of a linear combiner followed by a transfer function. The neuron’s output (o) is computed by weighting the summation of the neuron’s inputs which is then passed through a transfer function φ (.). This can be formulated in the Eq. 1 as:

$$o = \varphi\left(\sum_{i=1}^{m} w_i v_i + b_i\right) \quad (1)$$

where, υ_i is defined as the external input, m is the total number of inputs of the neuron and w_i and b_i are the weight and bias corresponding to the connection linking the i_{th} input to the neuron. A hyperbolic tangent transfer function has been chosen in this paper for its special properties such as symmetry and monotonicity

A hyperbolic tangent transfer function can be represented in the Eq. 2 as:

$$\varphi(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \quad (2)$$

The simplest form of trainable neural network, first developed (Rosenblatt, 1959), composed of two layers of nodes namely input and output layer. A mapping between the input and output data could be established by assigning weights to the input numerical data during training. More complicated MLPs which are commonly used consist of some hidden layers in addition to the input and output layers. These hidden layers enable the MLP to extract higher order statistics from a set of given data and hence, capture the complex relationship between input-output data. Therefore, MLPs commonly consist of an input layer for which the number of nodes are defined by size of input vector, one or more hidden layers which can have variable number of nodes depending on the application and an output layer which has one or more nodes depending on the number of output classes. Connections between these layers are defined by weights which are assigned in a supervised learning process so that the neural network would respond correctly to new data.

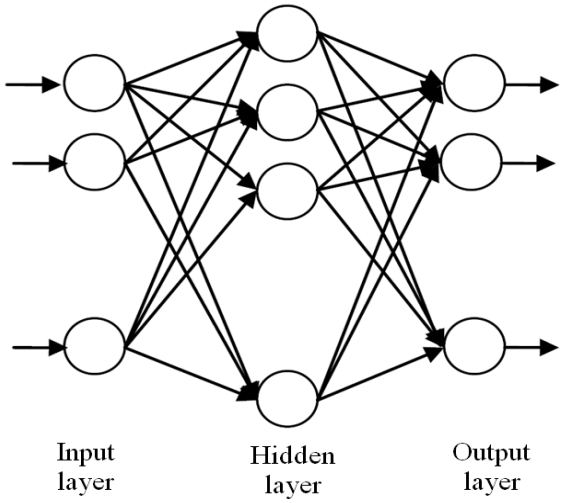

Fig. 2: The structure of a feed-forward MLP with one hidden layer

This can be done via a training algorithm, in which a cost function is computed by comparing the network's output and the desired output and is then minimized with respect to the network parameters.

In this study, Scaled Conjugate Gradient (SCG) algorithm is employed as a supervised training algorithm for the MLP. SCG algorithm, proposed by Moller (1993), is a class of conjugate gradient optimization techniques applied for training feed forward neural networks. Conjugate gradient techniques consist of iterative algorithms for optimization in which the minimum of an error function is located by proceeding in a direction on error surface which is conjugate to the previous step. This is advantageous to standard back propagation in which the algorithm proceeds only in a downward direction on error surface and therefore one step is partially undone by the next step.

SCG, like other training algorithms in feed forward networks, consists of a forward and backward pass. In the forward pass, an error is computed by comparing the network's output and the desired output which is then fed to a cost function. A Mean Square Error (MSE) cost function is chosen in this work, defined in the Eq. 3 as:

$$MSE = \frac{1}{2N}\sum_{j=1}^{N}(t_j - O_j)^2 \quad (3)$$

where, the MSE cost function is the mean of squared-error of the total number of patterns denoted by N. t_j and o_j are the desired output and the network's output respectively using the p^{th} input pattern p_j-O_{pj}.

During the backward pass, the network parameters i.e., the weights and biases are updated by computing the second order partial derivative of the cost function. This derivative is called Hessian matrix and is computed in the Eq. 4 as:

$$H = \frac{\delta^2 \xi(W)}{\delta W^2} \quad (4)$$

where, vector W indicates the network parameters. Using second order derivatives enable the network to predict the next input pattern more accurately. The Hessian matrix provides additional information related to the curvature of the cost function and hence results in faster and more accurate convergence to the minimum compared to first order techniques such as standard back propagation that uses first order derivatives only. The network parameters (i.e., weight and bias) update is then performed by changing the weight vector length and direction by Eq. 5:

$$wl+1 = w_l + a_l d_l \quad (5)$$

where, α_l and dl define the step size and search direction at step l respectively. The search direction at each step is chosen such that it does not have any component parallel to the previous search direction. The step size in each step is defined in the Eq. 6 as:

$$a_l = -\frac{d_l^T g_l}{d_l^T H d_l} \quad (6)$$

where, the error surface gradient at step l is defined as. In conjugate gradient algorithm, the Hessian matrix is computed by performing a line-search (Bishop, 1995). However, the high computational cost of line-search is an issue in conjugate gradient algorithm. In order to reduce this computational cost in SCG, is computed by evaluating. This is viable by online estimation of the Hessian matrix eigenvectors (LeCun *et al*., 1993). In this approach, the product of Hessian matrix with an arbitrary vector dm is computed without computing the full Hessian in each step. To ensure that Hessian in Eq. 6 is a positive definite matrix, it can be replaced by a modified version which is defined in the Eq. 7 as:

$$\tilde{H} = H + \beta I \quad (7)$$

β is a positive coefficient defined such that the new Hessian $\tilde{H}$ would be positive definite. In Eq. 7, I represent a unit matrix.

The training process is formed by several passes of information through the network called training iterations. Training may only complete when one of the predefined stopping criteria has occurred. These criteria are varied depending on the type of network and the training algorithm. In this study, a minimum amount of gradient performance and a maximum number of iterations are employed in conjunction as the network's stopping criteria to avoid over fitting and providing a good generalization performance for the network.

K-fold crosses validation: After training, the network's performance is evaluated by a test process through which the network's classification outcome is computed using a new set of data fed to the input layer. Hence, the available dataset is initially divided into two parts which will be used for training and test independently. Random division of the data into two parts is commonly used for the training/test data division. However, this might not result in a reliable evaluation of the network for a small dataset as a part of the data is only reserved for the test purpose. Moreover, the random division might bring about training/test datasets with different proportions of output classes. This especially happens in dataset with imbalanced output classes.

In k-fold CV, the dataset is divided into k independent folds where k-1 folds are used to train the network and the remaining one is reserved for the test purpose. This procedure is then repeated until all folds are used once as a test set. The final output of the network is then computed by averaging over the obtained accuracy from each test set. We will refer to k-fold CV as "simple k-fold CV" to differentiate it from the stratified k-fold CV.

Stratified k-fold CV is a special type of k-fold CV where the data folds are chosen such that each fold contains nearly the same proportion of the output data. Both stratified and simple k-fold CV is evaluated in this stuyd using different number of data folds to find an optimum evaluation method for the in-hand dataset.

Correlation coefficient: Correlation coefficient is a measure of dependence between two variables. In this study, Pearson's correlation coefficient (r) is used as a measure of linear relationship between different markers and the cancer outcome. Pearson's correlation coefficient can be obtained for two variables A and B by normalizing their covariance with respect to their standard deviation σ_A and σ_B as in the Eq. 8:

$$\gamma_{A,B} = \frac{\text{cov}(A,B)}{\sigma_A \sigma_B} = \frac{E\langle (A-\mu_A)(B-\mu_B)\rangle}{\sigma_A \sigma_B} \quad (8)$$

where, μ_A and μ_B are the expected values of two random variables A and B and E is the expected value of the random variable. Pearson's correlation coefficient assigns a number between -1 to +1 for the measure of linear dependence between variables. A positive value represents a positive linear relationship while a negative one implies negative linear relationship and 0 suggests no linear relation between variables.

RESULTS

The designed MLP in this study consists of an input layer and one hidden layer with variable number of nodes depending on the number of input markers and an output layer with one neuron. The network is fed with different combination of markers in each run to investigate the predictive significance of each marker. Hence, the number of input neurons is defined by the number of markers and the number of hidden neurons is optimized for each marker combination. The network is then trained using SCG algorithm and validated with k-fold CV.

The network's outcome is classified into four groups depending on the desired output. A True Positive (TP) outcome denotes a cancer case classified correctly while a False Negative (FN) implies a cancer case classified as normal incorrectly. Accordingly, True Negative (TN) and False Positive (FP) stand for the normal cases classified correctly and incorrectly respectively. The network is thus evaluated by computing its accuracy, sensitivity and specificity edfined in the Eq. 9-11 as:

$$\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (9)$$

$$\text{sesitivity} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (11)$$

The results obtained by running stratified and simple k-fold CV are first obtained by the designed network to predict the outcome using all input markers. These results are then analyzed to choose the best validation method to further investigate network prediction accuracy and markers' significance in outcome prediction.

Results of K-fold cross validation analysis: The output accuracy of the designed network using different number of folds for stratified and simple k-fold CV are illustrated in Fig. 3-5.

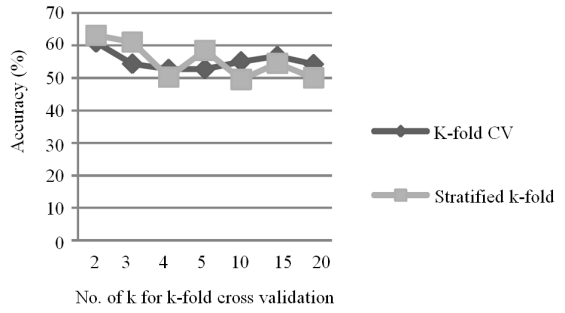

Fig. 3: Network accuracy using different values of k for k-fold cross validation

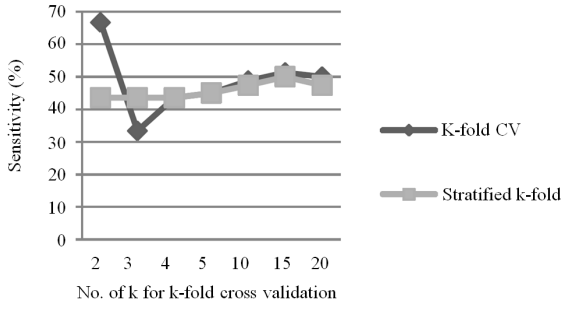

Fig. 4: Network sensitivity using different values of k for k-fold cross validation

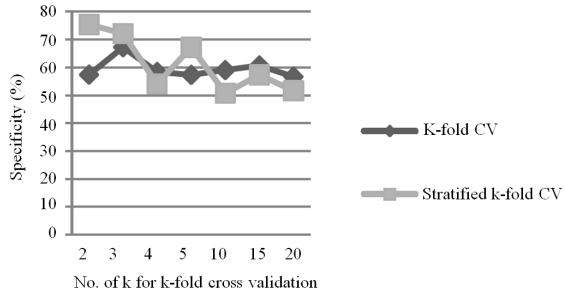

Fig. 5: Network specificity using different values of k for k-fold cross validation

Considering the network accuracy, sensitivity and specificity using different stratified and simple k-fold CVs illustrated in Fig. 3-5, stratified CV is preferred over a simple CV as it obtains better and more reliable results. Moreover, investigating the output results for different values of k for k-fold CV shows that 2-fold CV is a better choice for network validation with the in-hand dataset. Hence, the MLP results are evaluated using a stratified *2*-fold CV.

Table 3: Pearson's correlation coefficients computed for all 2-member possible combinations of the set including input markers and the output

	ER/PR	DNA Ploidy	SPF	G0G1/G2M	Nodal Status
ER/PR	1.00	-0.29	-0.03	0.07	-0.10
DNA Ploidy	-0.29	1.00	-0.27	-0.11	0.06
SPF	-0.03	-0.27	1.00	-0.27	0.21
G0G1/G2M	0.07	-0.11	-0.27	1.00	-0.04
Nodal Status	-0.10	0.06	0.21	-0.04	1.00

Table 4: Best MLP results for nodal status prediction using different number of markers

Marker combination	Accuracy (%)	Sensitivity (%)	Specificity (%)
All 4 markers	63	44	75
ER/PR, SPF and G0G1/G2M	63	44	75
ER/PR and G0G1/G2M	63	33	82
SPF	65	33	85

Results of correlation coefficient and MLP analysis: The results for Pearson's correlation coefficient computed for all 2-member possible combinations of the set including the input markers and the output are presented in Table 3. The cross section of each row and column in Table 3 shows the coefficient between the associated variables. The table illustrates a symmetric matrix with a diagonal of 1 as the Pearson's correlation coefficient is the same between variable A and B and vice versa and is 1 for two identical variables.

Results from Table 3 suggest significant linear relation between DNA ploidy and ER/PR ($p = 0.05$). The degree of linear dependence of SPF and DNA ploidy ($p = 0.06$) and G0G1/G2M and SPF ($p = 0.07$) is also noticeable. Nonetheless, there is no significant linear relation between other markers and the output ($p>0.1$). These results however, do not necessarily provide any indication about the existence of any nonlinear interaction between the different markers and the output.

The MLP results are obtained using different combination of the mentioned four markers in the form of 3, 2 and 1-member marker sets and also for the full marker set. The best classification results based on inputs including groups of 4, 3, 2 and 1 biomarkers are included in Table 4. First column in Table 4 shows the markers used in the combination while the other columns represent the obtained sensitivity, specificity and accuracy in percentage.

Results in Table 4 show that the prediction accuracy obtained using all markers remains virtually unchanged despite using a 3 or 2-marker set. This can be explained by the interaction between the markers. Removing DNA ploidy from the set of all markers results in the same accuracy, sensitivity and specificity. In addition, removing SPF from the set including ER/PR, SPF and G0G1/G2M results in a higher specificity at the cost of reduced sensitivity but the accuracy remains unchanged.

From Table 3, no significant linear relation could be found between the individual markers and the output. However, in spite of the lack of linear relation between them, the higher predictive accuracy provided by SPF alone compared to other combinations proves the existence of strong nonlinear relation between SPF and output captured by the MLP.

DISCUSSION

A good deal of research conducted in the field of breast cancer prognosis has led to the identification of many new prognostic markers. However, besides exploring novel markers, finding the relationship between the new markers to those previously used along with the additional information they can provide is of great importance. Therefore, a reliable prediction system capable of predicting cancer progression on the basis of the tumor markers and which can also define the predictive accuracy of these markers is highly demanded. In the search of the best prediction models, many research studies have confirmed ANN as a good modeling approach for cancer diagnosis and prognosis (Hudson and Cohen, 2000).

This study has presented an artificial neural network based method to define the predictive accuracy of the features or subsets of features in breast cancer prognosis in terms of nodal status prediction. The final network structure is a three-layered network trained using a SCG algorithm. Although a single perception can perform nonlinear classification, there is no evidence that it can realize optimal decision boundary and has poor ability to generalize to unseen data. On the other hand, MLP has been proven to realize the optimal decision boundary and has the ability to generalize well to unseen data (Hornik *et al*., 1989). Finally, the designed network is evaluated using different number of folds in stratified and simple k-fold CV.

The results show that stratified 2-fold CV is a more accurate and reliable method as it obtains a higher accuracy and specificity and also provides a more stable network validation in terms of sensitivity. This can be explained by the same proportion of the output data existing in each group (fold) in stratified CV. when simple CV is used to partition the data into k folds, one fold may contain only one output data. This gives rise to biased output accuracy as the network is tested with only one group of outputs in the test set.

This is rectified in stratified CV by having a balanced number of output groups in each fold.

The low variance and high accuracy of stratified 2-fold CV in small sample sizes has been confirmed for k-nearest neighbor classifiers (Weiss, 1991). This is also proved for the MLP used for the breast cancer data in this study as the stratified 2-fold CV obtains higher accuracy and specificity compared to simple CV and other number of folds in k-fold CV.

In addition, stratified CV shows more consistent results compared to simple CV especially for sensitivity. Although the sensitivity achieved by the simple *2*-fold CV is higher than that of stratified *2*-fold CV, the later is chosen as it is more reliable.

All the three marker combinations including 4, 3 and 2 markers include ER/PR. This shows the important role of including ER/PR as an individual marker in nodal involvement prediction. Amongst 3-marker input combinations, the arrangement including ER/PR, SPF and cell cycle distribution results in the best output accuracy which indicates the efficiency of this pattern for accurate prediction of nodal involvement. Between 2-marker combinations of ER/PR with other markers, the amalgamation with steroid receptors ends in the same accuracy achieved in the case of including all 4 biomarkers in the input which verifies the previous assumption about the efficiency of this combination for accurate prediction.

Pearson's correlation coefficient shows almost no linear relation between G0G1/G2M and nodal status outcome. ER/PR and G0G1/G2M are also hardly correlated linearly, based on the correlation coefficient results. However, the combination including ER/PR and G0G1/G2M provides a prediction as accurate as those results obtained by using all markers. These findings confirm the ability of the designed MLP in capturing nonlinear relations between these markers and the nodal status outcome.

Leaving DNA ploidy out from the network inputs does not cause any variation in classification accuracy. This can be explained by the close relation between G0G1/G2M and DNA ploidy. Since DNA ploidy is determined based on the percentage of cells being in G0G1 phase of cell cycle, it can be considered as an aspect of cell cycle distribution. Therefore, the inclusion of cell cycle distribution seems to compensate for the lack of DNA content information. It is worthy of note however that best prediction results are obtained by using only one marker-SPF. This confirms the predictive significance of this marker and also the negative correlation of markers in some cases which results in a lower predictive outcome using all the available markers.

CONCLUSION

This study presents an evaluation of four cellular and molecular breast cancer markers for the purpose of nodal status predication using a MLP neural network.

The main aim of the study is to investigate the neural network ability in capturing nonlinear interaction of these markers and nodal status in breast cancer. We have also assessed the effectiveness of stratified and simple k-fold CV for MLP outcome evaluation in case of having breast cancer dataset containing limited number of data. The results confirm the superiority of stratified 2-fold CV over the simple k-fold CV especially for a limited number of data. The ability of neural network in extracting the complex patterns existing in breast cancer tumor markers is further confirmed in this study.

ACKNOWLEDGMENT

The researchers thank Dr C. Bartoli and Professor F. Cajone of University of Milan for clinical collaboration.

REFERENCES

Ahmed, F.E., 2005. Artificial neural networks for diagnosis and survival prediction in colon cancer. Mol Cancer, 4: 29-29. DOI: 10.1186/1476-4598-4-29

Altman, D.G. and G.H. Lyman, 1998. Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res. Treatment, 52: 289-303. DOI: 10.1023/A:1006193704132

Anderson, M., S.S. Dlay and G.V. Sherbet, 2003. Oestrogen and progesterone receptor expression influences DNA ploidy and the proliferation potential of breast cancer cells. Anticancer Res., 23: 3029-3039.

Ashidi, N., M.I. Esugasini, S.M. Yusoff, M.N. Hayati and Othman, 2007. Fine needle aspiration cytology evaluation for classifying breast cancer using artificial neural network. Am. J. Applied Sci., 4: 999-1008.

Azua, J., P. Romeo, M. Serrano, M.D. Tello and J. Azua Jr, 1997. Prognostic value from DNA quantification by static cytometry in breast cancer. Anal Quant Cytol Histol, 19: 80-86. PMID: 9051190

Bae, J.H., J.W. Bae, S.U. Woo, C.W. Kim and J.B. Lee *et al*., 2007. S-phase fraction as an independent prognostic factor in invasive breast carcinoma -a study of long-term follow-up. J. Breast Cancer, 10: 36-42.

Bishop, C.M., 1995. Neural networks for pattern recognition. University Press, Oxford.

Burke, H.B., D.B. Rosen and P.H. Goodman, 1994. Comparing artificial neural networks to other statistical methods for medical outcome prediction. Proceedings of the IEEE International Conference on IEEE World Congress on Computational Intelligence, Neural Networks, Jun. 27-Jul. 2, IEEE Xplore Press, Orlando, pp: 2213-2216. DOI: 10.1109/ICNN.1994.374560

Burke, H.B., P.H. Goodman, D.B. Rosen, D.E. Henson and J.N. Weinstein, 1997. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, 79: 857-862. DOI: 10.1002/(SICI)1097-0142(19970215)79:4<857::AID-CNCR24>3.0.CO;2-Y

Clark, G.M., L.G. Dressler, M.A. Marilyn, A. Owens and G. Pounds, 1989. Prediction of relapse or survival in patients with node-negative breast cancer by DNA flow cytometry. New England J. Med., 320: 627-633. DOI: 10.1056/NEJM198903093201003

Concato, J., A.R. Feinstein and T.R. Holford, 1993. The risk of determining risk with multivariable models. Annals Int. Med., 118: 201-210.

Esteva, F.J. and G.N. Hortobagyi, 2004. Prognostic molecular markers in early breast cancer. Breast Cancer Res., 6: 109-18. DOI: 10.1186/bcr777

Etchells, T.A. and P.J.G. Lisboa, 2006. Orthogonal Search-based Rule Extraction (OSRE) for trained neural networks: A practical and efficient approach. Neural Netw. IEEE Trans., 17: 374-384. DOI: 10.1109/TNN.2005.863472

Gazic, B., J. Pizem, M. Bracko, T. Cufer and S. Borstnar *et al*., 2008. S-phase fraction determined on fine needle aspirates is an independent prognostic factor in breast cancer – a multivariate study of 770 patients. Cytopathology, 10: 294-302. DOI: 10.1111/j.1365-2303.2007.00528.x

Gilchrist, K.W., R. Gray, A.M.J. Driel-Kulker, W.E. Mesker and J.J. Ploem-Zaaijer, 1993. High DNA content and prognosis in lymph node positive breast cancer. A case control study by the University of Leiden and ECOG. Breast Cancer Res. Treat, 28: 1-8. DOI: 10.1007/BF00666350

Giuliano, A.E., R.C. Jones, M. Brennan and R. Statman, 1997. Sentinel lymphadenectomy in breast cancer. J. Clin. Oncol., 15: 2345-2350.

Grey, S.R., S.S. Dlay, B.E. Leone, F. Cajone and G.V. Sherbet, 2003. Prediction of nodal spread of breast cancer by using artificial neural network-based analyses of S100A4, nm23 and steroid receptor expression. Clin. Exper. Metastasis, 20: 507-514. DOI: 10.1023/A:1025846019656

Haykin, S.S., 2009. Neural Networks and Learning Machines. 3rd Edn., Prentice Hall, New York, ISBN: 9780131471399, pp: 906.

Hornik, K., M. Stinchcombe and H. White, 1989. Multilayer feedforward networks are universal approximators. Neural Netw., 2: 359-366. DOI: 10.1016/0893-6080(89)90020-8

Hosmer, D.W. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edn., Wiley, New York, ISBN: 0471356328, pp: 373.

Hudson, D.L. and M.E. Cohen, 2000. Neural Networks and Artificial Intelligence for Biomedical Engineering. 1st Edn., IEEE Press, New York, ISBN: 0780334043, pp: 306.

Kallioniemi, O.P., G. Blanco, M. Alavaikko, T. Hietanen and J. Mattila *et al*., 1988. Improving the prognostic value of DNA flow cytometry in breast cancer by combining DNA index and S-phase fraction: A proposed classification of DNA histograms in breast cancer. Cancer, 62: 2183-2190. DOI: 10.1002/1097-0142(19881115)62:10<2183::AID-CNCR2820621019>3.0.CO;2-B

Kaur, H. and S.K. Wasan, 2006. Empirical Study on Applications of Data Mining Techniques in Healthcare. J. Comput. Sci., 2: 194-200.

LeCun, Y., P.Y. Simard and B. Pearlmutter, 1993. Automatic learning rate maximization by on-line estimation of the hessian's eigenvectors. Morgan Kaufmann, 5: 156-163.

Lyman, G.H., A.E. Giuliano, M.R. Somerfield, A.B. Benson and D.C. Bodurka *et al*., 2005. American society of clinical oncology guideline recommendations for sentinel lymph node biopsy in early-stage breast cancer. J. Clin. Oncol., 23: 7703-7720. DOI: 10.1200/jco.2005.08.001

Moller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw., 6: 525-533. DOI: 10.1016/S0893-6080(05)80056-5

Moureau-Zabotto, L., C. Bouchet, D. Cesari, S. Uzan and J.P. Lefranc, 2005. Combined flow cytometry determination of S-phase fraction and DNA ploidy is an independent prognostic factor in node-negative invasive breast carcinoma: Review of a series of 271 patients with stage I and II breast cancer. Cancer Radiother, 91: 61-71. PMID: 15868432

Naguib, R.N.G. and G.V. Sherbet, 2001. Artificial Neural Networks in Cancer Diagnosis, Prognosis and Patient Management. 1st Edn., CRC Press, Boca Raton, ISBN: 0849396921

Naguib, R.N.G., H.A.M. Sakim, M.S. Lakshmi, V. Wadehra and T.W.J. Lennard *et al*., 1999. DNA ploidy and cell cycle distribution of breast cancer aspirate cells measured by image cytometry and analyzed by artificial neural networks for their prognostic significance. Inform. Technol. Biomed. IEEE Trans., 3: 61-69. DOI: 10.1109/4233.748976

Phippard, D.J., S.J. Weber-Hall, P.T. Sharpe, M.S. Naylor and H. Jayatalake *et al*., 1996. Regulation of Msx-1, Msx-2, Bmp-2 and Bmp-4 during foetal and postnatal mammary gland development. Development, 122: 2729-2737.

Pretorius, M.E., H. Waehre, V.M. Abeler, B. Davidson and L. Vlatkovic *et al*., 2009. Large scale genomic instability as an additive prognostic marker in early prostate cancer. J. Analytical Cellular Pathol., 31: 251-259. DOI: 10.3233/CLO-2009-0463

Remvikos, Y., P. Beuzeboc, A. Zajdela, N. Voillemot and H. Magdelenat *et al*., 1989. Correlation of pretreatment proliferative activity of breast cancer with the response to cytotoxic chemotherapy. J. Natl. Cancer Inst., 81: 1383-1387. DOI: 10.1093/jnci/81.18.1383

Rosenblatt, F., 1959. The perceptron: A probabilistic model for information storage and organization in the brain. Psych. Rev., 65: 386-407. DOI: 10.1037/h0042519

Schwarzer, G., W. Vach and M. Schumacher, 2000. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Stat. Med., 19: 541-561.

Suehiro, Y., T. Okada, T. Okada, K. Anno and N. Okayama *et al*., 2008. Aneuploidy predicts outcome in patients with endometrial carcinoma and is related to lack of CDH13 hypermethylation. Clin. Cancer Res., 14: 3354-3354. DOI: 10.1158/1078-0432.CCR-07-4609

Weiss, S.M., 1991. Small sample error rate estimation for k-NN classifiers. Patt. Anal. Mach. Intell. IEEE Trans. 13: 285-289. DOI: 10.1109/34.75516

Yuan, J., C. Hennessy, A.L. Givan, I.P. Corbett and J.A. Henry *et al*., 1992. Predicting outcome for patients with node negative breast cancer: A comparative study of the value of flow cytometry and cell image analysis for determination of DNA ploidy. Br J Cancer, 65: 461-465. DOI: 10.1038/bjc.1992.93