Breast Cancer Prediction and Cross Validation Using Multilayer Perceptron Neural Networks

S.A. Mojarad, S.S. Dlay, W.L.Woo, and G.V. Sherbet
School of Electrical, Electronic and Computer Engineering, Newcastle University, England, UK
{shirin.mojarad, s.s.dlay, w.l.woo, gajanan.sherbet}@ncl.ac.uk

Abstract—**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 has prohibited an accurate prognosis. The aim of this paper is to investigate the effectiveness of a multilayer perceptron (MLP) for the aim of predicting breast cancer progression using a set of four biomarkers of breast tumours. 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 also studied in order to assess their accuracy and reliability for neural network validation. We used output accuracy, sensitivity and specificity for selecting the best validation technique besides evaluating the network outcome for different combinations of markers. Findings suggest that ANN-based analysis provides an accurate and reliable platform for breast cancer prediction given that an appropriate design and validation method is employed.**

Index Terms—**Breast cancer, k-fold cross validation, multilayer perceptron (MLP), predictive analysis.**

I. 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 [1]. An important factor influencing the breast cancer caused 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 in patients with breast cancer.

Axillary lymph node status is the most significant factor for prognosis in early stage breast cancer [2]. Several prognostic factors including patient age, tumour size, tumour grade, DNA content (ploidy) and receptor status have been identified for nodal metastasis prediction. [3]. However, no individual or combination of these prognostic factors has replaced nodal dissection for node status determination [4].

Amongst prognostic markers, those which can be obtained via a minimally invasive method are preferred for the aim of nodal status and survival prediction 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 existing inter-relation between the markers has prohibited an accurate prediction 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 [5], proportional hazards regression (Cox regression) and Kaplan-Meier Curve [6]. However, there are several inadequacies in these methods which present doubts in their reliability. The study conducted by Concato et al. [7] on the deficiencies of these statistical methods has investigated the present problems of multivariate analyses in medical research. Some of the reported problems include overfitting 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 might be misleading in defining the prognostic value of the investigated markers [8].

Another approach that has been widely used for the aim of cancer prognosis is artificial neutrals network (ANN) [9, 10]. ANN has been confirmed as a robust method for the aim of cancer prognosis [11]. It is also superior to conventional methods employed for breast cancer prediction such as TNM (Tumor, Node, Metastasis) staging system and logistic regression [12]. 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 more widely used for the aim of cancer prediction and prognosis [9]. 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 [13].

978-1-86135-369-6/10/$25.00 ©2010 IEEE

In this paper, 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 paper 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 paper 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 paper is organized as follows: In section 2, some information about the breast cancer dataset used in this paper is provided. Methods in section 3 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. Results, discussion and conclusion are then followed in sections 4-6.

II. Breast cancer Dataset

The data utilised for nodal involvement analysis contains the information corresponding to four cellular and molecular breast tumour biomarkers pertaining to 46 patients who had been diagnosed with a carcinoma or benign breast tumour. 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 have been defined as an outcome for all 46 patients.

Except for ER/PR, which takes only three discrete values of 0-2, other markers are continuous within different ranges. These ranges are 2.33 - 11.58 for DNA ploidy, 0.76 - 30.7 for SPF and 3.56 - 117.6 for G0G1/G2M. Nodal status is defined as either 0 or 1 for the case of no node involved or metastasis to the regional lymph nodes respectively. All the above 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.

III. Methods

A. MLP

A MLP neural network consists of a set of interconnected artificial neurons connected only in a forward manner to form layers. 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 $\varphi(.)$. This can be formulated as

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

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 as

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

The simplest form of trainable neural network, first developed in 1959 [14], 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. Hence, 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. 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 paper, scaled conjugate gradient (SCG) algorithm is employed for training the MLP neural network. SCG algorithm, proposed by Moller [15], is a class of conjugate gradient optimization techniques which 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, formulated as

$$MSE = \frac{1}{2N}\sum_{j=1}^{N}\left(t_j - o_j\right)^2 \tag{3}$$

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.

During backward pass, the network parameters – weights and biases are updated by computing the second partial derivative of the cost function. This derivative, also called a Hessian matrix, is computed with respect to the network parameters – weights and biases to achieve an optimum value for these parameters. This enables the network to predict the next input pattern more accurately. Using 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 derivatives only.

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 paper, a minimum amount of gradient performance and a maximum number of iterations are employed in conjunction as the network's stopping criteria to avoid overfitting and providing a good generalization performance for the network.

B. K-fold cross 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 data to 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 a data with imbalanced output classes.

In k-fold CV, the data 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 a 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 are evaluated in this paper using different number of data folds to find an optimum evaluation method for the in-hand dataset.

C. Correlation coefficient

Correlation coefficient is a measure of dependence between two variables. In this paper, Pearson's correlation coefficient 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

$$r_{A,B} = \frac{cov(A,B)}{\sigma_A \sigma_B} = \frac{E\langle (A-\mu_A)(B-\mu_B)\rangle}{\sigma_A \sigma_B} \quad (4)$$

where μ_A and μ_B are the expected values of two random variables A and B. Pearson's correlation coefficient assigns a number between -1 and 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 relation between variables.

VI. 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 cross validation.

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 with respect to the above outcome classes as

$$accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (5)$$

$$sesitivity = \frac{TP}{TP+FN} \quad (6)$$

$$specificity = \frac{TN}{TN+FP} \quad (7)$$

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.

A. 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. 1-3.

Considering the network accuracy, sensitivity and specificity using different stratified and simple k-fold CVs illustrated in Fig. 1-3, 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.

B. 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 1. The cross section of each row and column in Table 1 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.

These results suggest some degree of linear relation between Nodal status outcome and SPF. However, there is no or little linear relation between other markers and the output. The degree of linear dependence of SPF with DNA ploidy and G0G1/G2M is also noticeable. These results however do not present any information about the existence of any nonlinear interaction between different markers and the output.

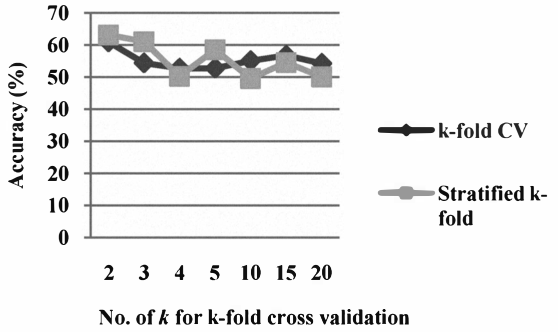

Fig. 1: Network accuracy using different values of k for k-fold cross validation

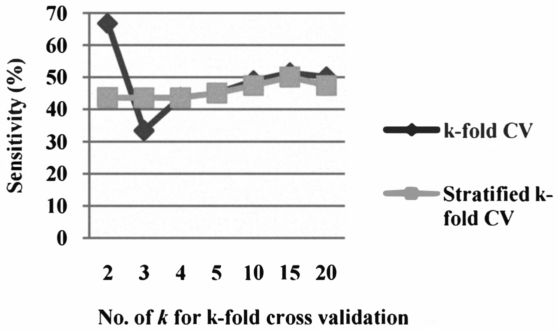

Fig. 2: Network sensitivity using different values of k for k-fold cross validation

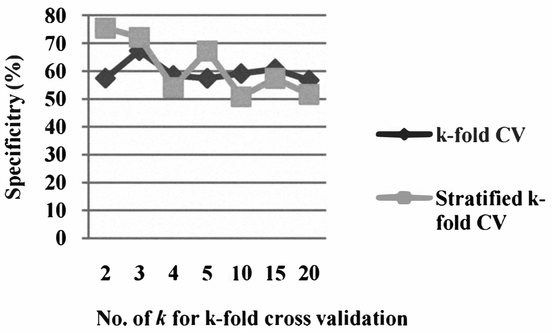

Fig. 3: Network specificity using different values of k for k-fold cross validation

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 Tables 2. First column in Table 2 shows the markers used in the combination while the other columns represent the obtained sensitivity, specificity and accuracy in percentage.

V. 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

TABLE I
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	-0.29	-0.03	0.07	-0.10
DNA Ploidy	-0.29	1	-0.27	-0.11	0.06
SPF	-0.03	-0.27	1	-0.27	0.21
G0G1/G2M	0.07	-0.11	-0.27	1	-0.04
Nodal Status	-0.10	0.06	0.21	-0.04	1

TABLE II
BEST MLP RESULTS FOR NODAL STATUS PREDICTION USING DIFFERENT NUMBER OF MARKERS

Marker combination	Accuracy (%)	Sensitivity (%)	Specificity (%)
All 4 markers	63.04	43.58	75.40
All Markers except DNA ploidy	63.04	43.58	75.40
ER/PR and G0G1/G2M	63.04	33.33	81.96
SPF	65.21	33.33	85.24

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 [16].

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.

The network is then evaluated using different number of folds in stratified and simple k-fold CV. The results show that stratified 2-fold cross validation 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.

Furthermore, the results obtained by the MLP for different marker combinations demonstrate the same outcome for the set including and excluding DNA ploidy from the other three markers. This shows little or no relation between this marker and the status of nodal involvement which can be also confirmed from those results obtained by the Pearson's correlation coefficient.

The combination including ER/PR and G0G1/G2M also provides a prediction as accurate as those results obtained by using all markers. However, 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. These findings confirm the ability of the designed MLP in capturing nonlinear relations between these markers and the nodal status outcome.

It is noticeable that however, best prediction results are obtained by using only one marker – SPF. This confirms the predictive significance of this marker in cancer prediction and also the negative correlation of markers in some cases which results in a lower predictive outcome using all the available markers.

VI. Conclusion

This paper 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 paper was 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 k-fold CV over the simple case of 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 paper.

Acknowledgment

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

References

[1] T. A. Etchells and P. J. G. Lisboa, "Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach," *Neural Networks, IEEE Transactions on,* vol. 17, pp. 374-384, 2006.

[2] A. Luini, G. Gatti, B. Ballardini, S. Zurrida, V. Galimberti, P. Veronesi, A. R. Vento, S. Monti, G. Viale, G. Paganelli, and U. Veronesi, "Development of axillary surgery in breast cancer," *Ann Oncol,* vol. 16, pp. 259-262, 2005.

[3] G. H. Lyman et al., "American Society of Clinical Oncology Guideline Recommendations for Sentinel Lymph Node Biopsy in Early-Stage Breast Cancer," *J Clin Oncol,* vol. 23, pp. 7703-7720, 2005.

[4] A. E. Giuliano, R. C. Jones, M. Brennan, and R. Statman, "Sentinel lymphadenectomy in breast cancer," *J Clin Oncol,* vol. 15, pp. 2345-2350, 1997.

[5] D. Hosmer and S. Lemeshow, "Model-building strategies and methods for logistic regression," in *Applied Logistic Regression*: John Wiley & Sons, Inc, New York, NY, 2000, pp. 91-142.

[6] B. Efron, "Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve," *Journal of the American Statistical Association,* vol. 83, pp. 414-425, 1988.

[7] J. Concato, A. R. Feinstein, and T. R. Holford, "The Risk of Determining Risk with Multivariable Models," *Annals of Internal Medicine,* vol. 118, pp. 201-210, February 1993.

[8] D. G. Altman and G. H. Lyman, "Methodological challenges in the evaluation of prognostic factors in breast cancer," *Breast Cancer Research and Treatment,* vol. 52, pp. 289-303, 1998.

[9] G. Schwarzer, W. Vach, and M. Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statistics in Medicine,* vol. 19, pp. 541-561, 2000.

[10] F. E. Ahmed, "Artificial neural networks for diagnosis and survival prediction in colon cancer," *Mol Cancer,* vol. 4, p. 29, 2005.

[11] H. B. Burke, D. B. Rosen, and P. H. Goodman, "Comparing artificial neural networks to other statistical methods for medical outcome prediction," *IEEE World Congress on Computational Intelligence, IEEE International Conference on*, pp. 2213-2216, 1994.

[12] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, J. J. R. Marks, D. P. Winchester, and D. G. Bostwick, "Artificial neural networks improve the accuracy of cancer survival prediction," *Cancer,* vol. 79, pp. 857-862, 1997.

[13] S. Haykin, *Neural Networks and Learning Machines*, Third ed.: Prentice Hall, 2009.

[14] F. Rosenblatt, "The perceptron: a probabilistic method for information storage in the brain," *Psych Rev,* vol. 65, pp. 386-407, 1959.

[15] M. F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks,* vol. 6, pp. 525-533, 1993.

[16] D. Hudson and M. E. Cohen, "Neural Networks and Artificial Intelligence for Biomedical Engineering," *Piscataway, NJ: IEEE Press,* 2000.