USING ARTIFICIAL NEURAL NETWORKS TO PREDICT MALIGNANCY OF OVARIAN TUMORS

C. Lu[1], J. De Brabanter[1], S. Van Huffel[1], I. Vergote[2], D. Timmerman[2]
[1]Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium
[2]Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium

Abstract-**This paper discusses the application of artificial neural networks (ANNs) to preoperative discrimination between benign and malignant ovarian tumors. With the input variables selected by logistic regression analysis, two types of feed-forward neural networks were built: multi-layer perceptrons (MLPs) and generalized regression networks (GRNNs). We assess the performance of the models using the Receiver Operating Characteristic (ROC) curve, particularly the area under the ROC curves (AUC), and statistically compare the cross-validated estimate of the AUC of different models.**

Keywords- **Ovarian tumor, multi-layer perceptron, generalized regression network, logistic regression, ROC, cross-validation.**

I. INTRODUCTION

Ovarian masses are a common problem in gynecology. A reliable test for preoperative discrimination between benign and malignant ovarian tumors is of considerable help for clinicians in choosing appropriate treatments for patients. Conservative management or less invasive surgery suffices for patients with a benign tumor; in contrast, those with suspected malignancy should be timely referred to an oncological surgeon.

There have already been several attempts to automate the classification process, such as the risk of malignancy index and logistic regression [2][3]. This paper discusses the development of supervised ANNs, both MLPs and GRNNs, to predict the malignancy of ovarian tumors. Statistical data analysis and input selection are first described. Then the issues related to network design and training, especially how to avoid overfitting, are addressed. The use of AUC as performance measure of the models, and the statistical comparison of the overall performance of the models by means of cross-validation, are outlined. The results and conclusions are presented at the end of the paper.

II. THE DATA

The data set includes the information of 425 patients who were referred to the University Hospital Leuven, Belgium, between 1994 and 1999. Among the available 425 cases, 291 patients had benign tumors, whereas 134 had malignant tumors. Firstly we performed a statistical analysis of the data. Input variables for model building were selected by means of a multivariate logistic regression analysis.

Univariate data analysis: The original data set contains 25 features. Some feature values have been transformed prior to further analysis, e.g. CA 125 serum level was rescaled by taking its logarithm. Table I lists most important variables that were considered.

TABLE I
Demographic, serum marker, color Doppler imaging and morphologic variables

	Variable (symbol)	Benign	Malignant
Demographic	Age (*age*)	45.6 ± 15.2	56.9 ± 14.6
	Postmenopausal (*meno*)	31.0 %	66.0 %
Serum marker	CA 125 (log) (*l_ca125*)	3.0 ± 1.2	5.2 ± 1.5
CDI	High color score (*colsc4*)	19.0%	77.3 %
Morphologic	Abdominal fluid (*asc*)	32.7 %	67.3 %
	Bilateral mass (*bilat*)	13.3 %	39.0 %
	Unilocular cyst (*un*)	45.8 %	5.0 %
	Multiloc/solid cyst (*mulsol*)	10.7 %	36.2 %
	Solid (*sol*)	8.3 %	37.6 %
	Smooth wall (*smooth*)	56.8 %	5.7 %
	Irregular wall (*irreg*)	33.8 %	73.2 %
	Papillations (*pap*)	12.5 %	53.2 %

Note: for the continuous variables the mean and standard deviation for each class are reported; for binary variables, the last two columns give the relative presence of the feature in both classes of benign and malignant tumors, e.g. 67.3% of the patients with a malignant tumor had abdominal fluid.

Multivariate data analysis: To get a first idea of the important predictors, we performed a factor analysis using the principal components as factors. Fig. 1 shows the biplot in a 2-dimensional space generated by (FACTOR1, FACTOR2). The

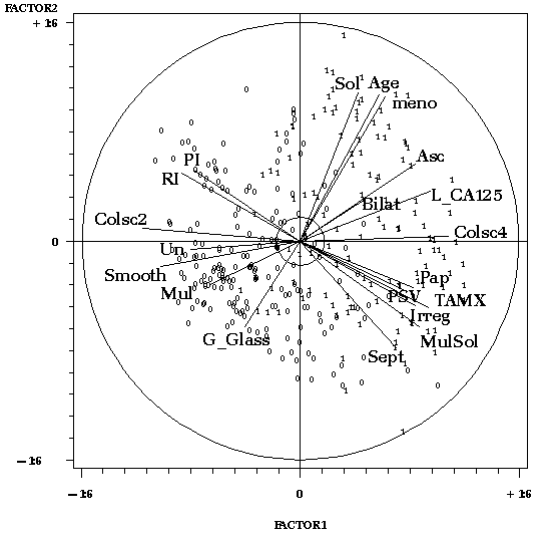


Fig. 1. Biplot of Ovarian Tumor data. The observations are plotted as points (0=benign, 1=malignant), the variables are plotted as vectors from the origin, i.e. taking the respective factor loadings as the coordinates.

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction (IUAP P4-02 and P4-24), initiated by the Belgian State, Prime Minister's Office - Federal Office for Scientific, Technical and Cultural Affairs, of the European Community TMR Programme, Networks, project CHRX-CT97-0160, of the Concerted Research Action (GOA) projects of the Flemish Government MEFISTO-666, of the IDO/99/03 project (K.U.Leuven).

biplot visualizes the correlation between the variables, and the relations between the features and classes. E.g. the variables with small angles like (*age, meno*) are highly correlated; the observations of malignant tumors (1) have relatively high values for variables *sol, age, meno, asc, l_ca125, colsc4, pap, irreg*, etc; but relatively low values for the variables *colsc2, smooth, un, mul*, etc.

Input Selection: important predictors could be selected from stepwise multivariate logistic regression analysis of the whole data set. They could also be obtained by fixing several of the most significant variables, then varying combinations with the other predictive variables. Different logistic regression models with different subsets of input variables have been built and validated. In the end, two subsets of variables were selected according to their predictive performance on the training set and test set. The subset with eight variables is just the result of the stepwise logistic regression, and is noted as MODEL1. The other subset is called MODEL2, containing seven variables (see Fig. 2).

III. METHODS

A. Network Design and Training

Generalization is a central issue both in network design and training. A properly trained NN should have the capability to extract the unknown relationships from the training data and have the generalization capability towards unseen cases from the same distribution. Overfitting occurs when the error on the training set is driven to a very small value, but when new data is presented to the network the error is large. Also, the more complex the neural network, the higher the risk for overfitting.

Multi-layer Perceptrons:

The most commonly used neural network structure for classification tasks, are multi-layer perceptrons (MLPs). Fig. 2 illustrates the architecture of the one-hidden-layer, one output variable network and its mapping function, which we use in the experiments. The activation functions $g(.)$ could vary from layer to layer; the typical ones are the logistic sigmoidal function, tanh activation function and threshold function.

In the experiments, the number of hidden neurons is chosen as 3. The activation functions for both layers are logistic sigmoidal functions, which transform all the output values into the interval [0, 1]. We denote the MLP that takes MODEL1 as input variables by MLP1; the one that takes MODEL2 as input variables is called MLP2.

The training of the feed-forward NN is often done by an iterative backpropagation procedure, until the discrepancy between the target output t_k and actual response y_k is minimized. The commonly used error function which reflects this discrepancy within a set of N data, is the sum of squared error (*sse*) function

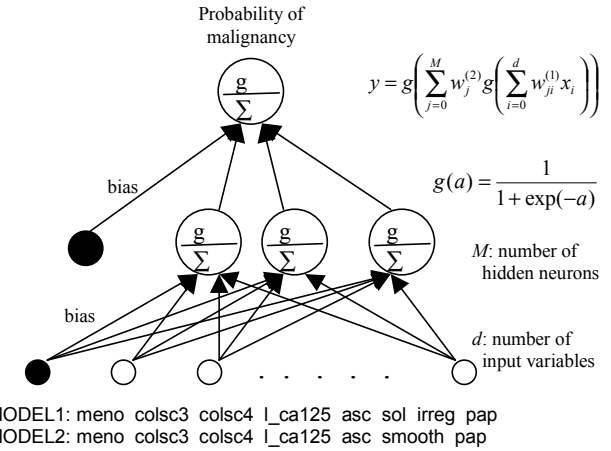


Fig. 2. Architecture of MLPs for Predicting Malignancy of Ovarian Tumors

$$sse = \sum_{k=1}^{N} (t_k - y_k)^2 \tag{1}$$

This type of error function is continuous and differentiable, so it can be used in gradient based optimization techniques such as steepest descent, quasi-Newton or Levenberg-Marquardt methods. There are also several alternatives in order to avoid overtraining. One is early stopping, the others are regularization techniques which favor smoother network mapping by adding a penalty term to the error function.

In this study, the weight and bias values were updated according to the Levenberg-Marquardt optimization. The error function to be minimized is a combination of *sse* and *ssw* (the sum of squared weights):

$$E_{reg} = \alpha sse + \beta ssw \tag{2}$$

where α and β are the regularization hyperparameters which are determined by a Bayesian approach [1].

The initial values of the weights and biases are randomly chosen from a normal distribution with mean zero and variance one. The training is repeated 100 times with different initializations; the parameters of the MLP which exhibit the best performance, i.e. the one with the highest AUC on the test set, will be taken as the final parameters of the MLP.

Generalized Regression Networks:

The generalized regression neural networks (GRNNs) are the paradigms of radial basis function (RBF) networks, often used for function approximations. It's another term for Nadaraya-Watson kernel regression, and has the following form for the function mapping [1].

$$y(\mathbf{x}) = \frac{\sum_k t_k \exp\{-\|\mathbf{x} - \mathbf{x}_k\|^2 / 2h^2\}}{\sum_k \exp\{-\|\mathbf{x} - \mathbf{x}_k\|^2 / 2h^2\}} \tag{3}$$

GRNNs share a special property, namely that they do not require iterative training; the hidden-to-output weights are just the target values t_k, so the output $y(\mathbf{x})$, is simply a weighted average of the target values t_k of training cases $\mathbf{x}_k$ close to the given input case $\mathbf{x}$. It can be viewed as a normalized RBF network in which there is a hidden unit centered at every training case. These RBF units are called

"kernels" and are usually probability density functions such as the Gaussians considered in (3). The only weights that need to be learned are the widths of the RBF units h. These widths (often a single width is used) are called "smoothing parameters" or "bandwidths" and are usually chosen by cross-validation. GRNN is a universal approximator for smooth functions, so it should be able to solve any smooth function-approximation problem given enough data. The main drawback of GRNNs is that, like kernel methods in general, they suffer seriously from the curse of dimensionality. GRNNs cannot ignore irrelevant inputs without major modifications to the basic algorithm.

We denote the GRNN with 8 input variables of MODEL1 as GRN1, the one with 7 input variables of MODEL2 as GRN2.

B. Performance Measure

The most commonly used performance measure of a classifier or a model is the classification accuracy, or the rate of correct classification, under the assumptions of equal misclassification costs and constant class distribution in the target environment. Both assumptions are not satisfied in real-world problems. Unlike classification accuracy, ROC is independent of class distributions or error costs and has been widely used in the biomedical field. Let's give a brief description about the ROC curves.

Assume a dichotomic classifier $y(\mathbf{x})$, which is the output value of the classifier given input $\mathbf{x}$. Then the ultimate decision is taken by comparing the output $y(\mathbf{x})$ with a certain cutoff value. The *sensitivity* or true positive rate of a classifier is then defined as the proportion of malignant cases that are predicted to be malignant, and *specificity* as the proportion of benign cases that are predicted to be benign. The false positive rate is 1-*specificity*. When varying the cutoff value, the sensitivity and specificity will change. A ROC curve is constructed by plotting the sensitivity versus the false positive rate, or 1-*specificity*, for varying cutoff values. The AUC is a one-value measure of the accuracy of a test. It can be statistically interpreted as the probability of the classifier to correctly classify malignant cases and benign cases. The higher AUC, the better the test. In this study, the area under the ROC curves was obtained by a non-parametric method based on the Wilcoxon statistic, using the trapezoidal rule, to approximate the area. This method also gives a standard error that can be used for comparing two different ROC curves [4].

C. Overall performance estimate from cross-validation

The commonly used method for estimating the generalization error in neural networks is cross-validation. *Holdout method* is one often used cross-validation method, which partitions the data into two mutually exclusive subsets, namely a training set and a test set. This is what we will do in the first experiment.

One can also repeat the holdout method k times. Each time a different partition is chosen. Then the estimated AUC is derived by averaging over all the runs. However, in medical practice, the holdout method makes inefficient use of the data set, which is usually smaller than desired. For example, one third of the data set is not used for training the classifier.

K-fold cross-validation is a variant of cross-validation. The data set is randomly divided into k (k>1) mutually exclusive subsets (k folds) of approximately equal size. The model is trained on all the subsets except for one, and the validation AUC is measured by testing it on the subset left out. This procedure is repeated k times, each time using a different subset for validation. The performance of the model is assessed by averaging the AUCs under validation over the k estimates. Repeating the k-fold cross-validation for multiple runs can provide a better statistical estimate.

The cross-validation estimate is a random number that depends on the division of the data set. We hope that the estimates have low bias and low variance. Leave-one-out is a special k-fold cross-validation, in which the number of folds equals the number of available data. This method is almost unbiased, but has high variance, leading to unreliable estimates. When choosing the number of folds, we would like to tradeoff bias for low variance.

IV. RESULTS

The above 4 neural networks, which encompass 2 kinds of architectures and 2 sets of input variables, are constructed and trained with the neural network toolbox of Matlab 6. Function *trainbr* is used to train the MLPs. Function *grnn* is called to create GRNNs, the optimal values for the width of the radial basis function are found by searching in the interval [0.5, 5].

The overall data, both for input variables and output variable, are first preprocessed: the continuous variable *l_ca125* is standardized and the binary variables {0,1} are transformed to {-1,1}, since the two algorithms perform best on data within [-1, 1].

A. AUC from Holdout Cross-Validation

We take the data of the most recently treated 160 patients as test set, the remaining 265 as the training set. The proportions of malignant tumors in the training set and test set are both about 1/3. Table II reports the AUCs and their standard error (according to Hanley and McNeil's method [3]) of the four neural networks, both on the training set and test set. The performance of the Risk of Malignancy Index (RMI) and two logistic regression (LR) models LR1 and LR2, using respectively MODEL1 and MODEL2 as inputs, are also shown for comparison.

TABLE II
Area Under the ROC curve (AUC) and its standard error

	Training		Test	
Model	AUC	SE	AUC	SE
RMI	0.898	0.024	0.861	0.034
LR1	0.972	0.013	0.904	0.029
LR2	0.966	0.014	0.908	0.029
MLP1	0.975	0.012	0.924	0.026
MLP2	0.964	0.015	0.917	0.027
GRN1	0.966	0.015	0.911	0.028
GRN2	0.968	0.014	0.905	0.029

We can observe from this table, that LRs, NNs have significantly higher AUCs than RMI. However this difference is not significant on the test set. Till now, one might ask, how much confidence we can get from these results? How representative is the test set we choose in this way? To answer this question, we will perform a *k*-fold cross-validation as introduced above.

B. AUC from K-fold Cross-validation

As we have a moderately sized data set (*N*=425) and two classes, we developed a stratified 7-fold cross-validation. Stratification forces an equal proportion of malignant cases (32%) in each fold. For each subset of the data, a model is developed with around 365 data in the training set and 60 in the test set. This procedure is repeated 30 times, by randomly dividing the data set into seven stratified folds.

The estimated AUC for each trial of 7-fold cross-validation is the mean of AUCs, denoted by mAUCs, over all the 7 validations. Then the mean and variance of the 30 mAUCs can be computed. The 30 mAUCs for each model are shown in the boxplot of fig. 3. Fig. 4 shows the expected ROC curves for models with input variables MODEL1, which are obtained by *averaging* [5].

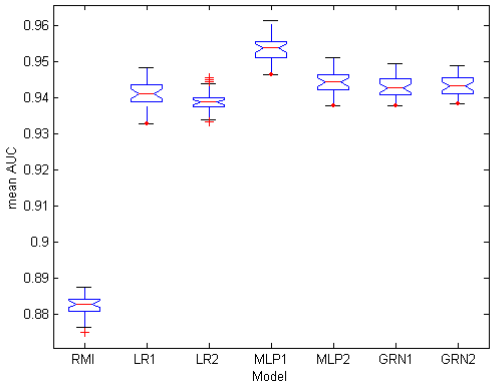


Fig. 3. Boxplot of mean AUCs. The line in the middle of the notched "box" is the sample median, the lower and upper lines of the "box" are the 25th and 75th percentiles of the sample.

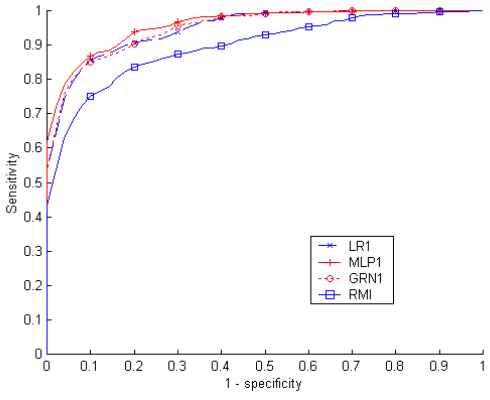


Fig. 4. Expected ROC curves of model MLP1, GRN1, LR1, and RMI

We then conduct a one-way ANOVA followed by Tukey multiple comparison [6]. The mean and variance of mAUCs of different models are listed in table III ordered by the mean. The subsets of adjacent means that are not significantly different at 95% confidence level are shown, and are indicated by drawing a dashed line under the subsets. We conclude that: all the models including LRs, MLPs and GRNs, have higher expected AUCs than the risk of malignancy index (RMI); the multi-layer perceptrons have higher expected AUC than the models generated from the other methods.

TABLE III
Rank ordered significant subgroups from multiple comparison on mean AUC

Models	RMI	LR2	LR1	GRN1	GRN2	MLP2	MLP1
mean mAUC	0.882	0.939	0.941	0.943	0.944	0.944	0.954
SD	0.003	0.003	0.004	0.003	0.003	0.003	0.003

V. Discussion And Conclusions

Our experiments confirm that neural network classifiers have the potential to give a more reliable prediction of the malignancy of ovarian tumors based on patient data. Multivariate statistical analysis could be of great help in obtaining an overview of the data set and in selecting predictive input variables. During network design and training, some techniques can be applied in order to avoid overfitting. Area under the ROC curves is the advocated performance measure of different models.

However, neural network models are black-box models. A hybrid methodology, which combines them with the advantages of white-box models (e.g. Bayesian network models), might be more promising.

References

[1] C. Bishop, *Neural Networks for Pattern Recognition,* Oxford University Press, 1996.

[2] D. Timmerman, et al., "A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: The development of a new logistic regression model," *Am J Obstet Gynecol*, vol 181, no. 1, pp. 57-65, 1999.

[3] D. Timmerman, et al., "Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses," *Ultrasound Obstet Gynecol*, 13:17-25, 1999.

[4] J.A. Hanley and B. McNeil, "The meaning and use of the area under a Receiver Operating Characteristic curve," *Diagnostic Radiology*, vol. 143, no. 1, pp.29-36, 1982.

[5] F. Provost, T. Fawcett, R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," *Proceedings of the 15th International Conference on Machine Learning (IMLC-98)*, Madison, WI, 1998.

[6] J. Neter, M.H. Kutner, C.J. Nachtsheim, W. Wasserman, *Applied Linear Statistical Models*, fourth edition. WCB/McGraw-Hill, 1996.