MISSING DATA ESTIMATION FOR CANCER DIAGNOSIS SUPPORT

Witold Jacak[(a)], Karin Proell[(b)]

[(a)]Department of Software Engineering
Upper Austria University of Applied Sciences Hagenberg, Softwarepark 11, Austria
[(b)]Department of Medical Informatics and Bioinformatics
Upper Austria University of Applied Sciences Hagenberg, Softwarepark 11, Austria

[(a)] Witold.Jacak@fh-hagenberg.at [(b)] Karin.Proell@fh-hagenberg.at

ABSTRACT

The paper presents a heterogeneous neural network based system that can be used for estimation of missing tumor marker values in patient data and in a second step for calculating the possibility of a cancerous disease.
For estimation of missing values we use different approaches: neural network based estimation of a specific marker depending on existing values of a related marker and neural network based estimation of missing tumor markers depending on standard blood parameter measurements. Finally we compare the results for calculating the possibility of a cancerous disease using different methods for missing value estimation of patient data.

Keywords: neural network, tumor marker prediction, missing values in biomedical data

1. INTRODUCTION

Tumor markers are substances produced by cells of the body in response to cancer but also to noncancerous conditions. They can be found in body liquids like blood or in tissues and can be used for detection, diagnosis of some types of cancer. For different types of cancer different tumor markers can show abnormal values and the levels of the same tumor marker can be altered in more than one type of cancer.

Blood examination tests only a few tumor marker values and for this reason the usage of such incomplete data for cancer diagnosis support needs estimation of missing marker values. Neural networks are proven tools for prediction tasks. For example neural networks were applied to differentiate benign from malignant breast conditions bases on blood parameters (Astion et Wilding 1992), for diagnosis of different types of liver disease (Reibnegger et al. 1991), for early detection of prostate cancer (Djavan et al. 2002; Matsui et al. 2004), for studies on blood plasma (Liparini et al. 2005) or for prediction of acute coronary syndromes (Harrison et al. 2005).

In this work we present a heterogeneous neural network based system that can be used for tumor marker value estimation and for prediction of cancer possibility.

We combine different neural networks, which calculate the possibility of a cancerous disease in different ways,

- without estimation of missing marker values,
- with neural network based estimation of missing marker values depending on existing values of other markers, and
- with neural network based estimation of missing marker values depending on standard blood parameters.

2. GENERAL CANCER DIAGNOSIS SUPPORT SYSTEM

We focus our considerations on the design of a complex decision support system for the calculation of the possibility of cancerous diseases.

The cancer prediction system is based on data coming from vector $\boldsymbol{C} = (C_1, \ldots, C_m)$ of tumor marker values. We use several parallely coupled neural networks to calculate the possibility of general cancer occurrence.

The basic problem in such approaches is data incompleteness of training data, which leads to problems in training neural networks. Data completeness can be achieved in two ways.

First we can use the existing values of markers in vector $\boldsymbol{C} = (C_1, \ldots, C_m)$ of tumor marker values to estimate the missing values of other markers in the same vector.

Second we can estimate the missing values of markers by using supporting data - in this case a blood parameter vector $\boldsymbol{P} = (P_1, \ldots, P_n)$ of each patient is used. Frequently also this vector is incomplete too. For that reason it is necessary to develop a system using complete or partially complete blood parameters for directly estimating values for missing tumor markers or for a classification of them. The structure of the system is presented in Figure 1.

3. TUMOR MARKER VALUES BASED CANCER DIAGNOSIS SUPPORT SYSTEM

Cancer diagnosis support uses parallel working systems ($Cancer^k$), with the same structure of networks trained for different types of cancer. The input of each $Cancer^k$ system is the complete or incomplete vector $\boldsymbol{C}$ of tumor marker specific for the chosen type of cancer, and the output represents the possibility (values between 0 and 1) of a cancerous disease. Output values of the network system greater than 0,5 are treated as cancer occurrence.

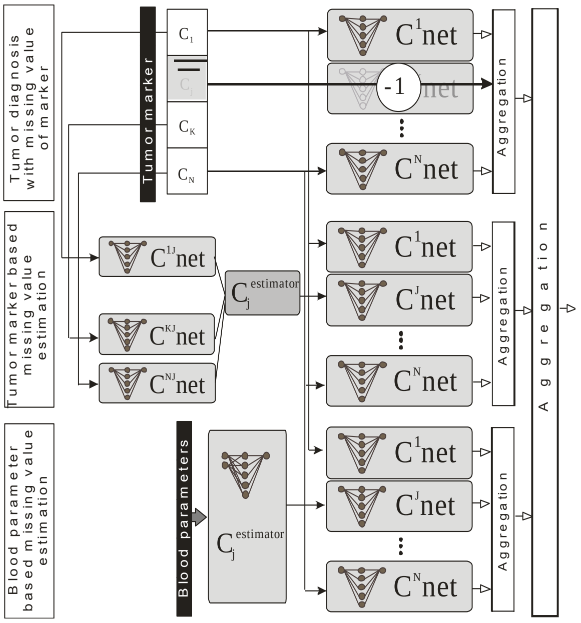

Figure 1. Architecture of Data Driven Cancer Diagnosis Support System

Each *Cancer*^{k} system consists of many different groups of neural networks (see Figure 1.).

- Group of neural networks (C^k_{net}) for individual marker C_i*: i=1,..,m.*
- Feed forward neural network (C^{Group}_{net}) for a whole vector of marker ***C***, with complete or incomplete values.
- Group of neural networks (C^{kj}_{net}) for estimation of marker value of marker C_j based on marker C_k
- Feed forward neural network (P^{Group}_{net}) for estimation of maker value *Cj*, depending on blood parameters
- Cascaded coupled aggregation method for final calculation of cancer plausibility.

4. GROUP OF SEPARATE NEURAL NETWORKS FOR INDIVIDUAL MARKER (C^K_{NET})

The first group of neural networks contains parallel coupled neural networks, which are individually trained for different tumor markers. Each neural network is of type feed forward with one hidden layer having 6 -10 neurons, activation functions tan/sigmoid, further one input (normalized tumor marker value) and one output (diagnosis: 0 – no cancer (healthy) and 1– cancer (ill)). The networks were trained independently of type of cancer disease (i.e. for all types of cancer diseases).

The values of markers are further categorized to four intervals (Classes). The first interval includes all values less than a *Normal Value* of marker, the second interval includes all values between the *Normal Value* and an *Extreme Normal Value* of marker, third interval includes values between the *Extreme Normal Value* and a still *Plausible Value* of marker and fourth interval includes all values greater than *Plausible Value*.

The input values of each network for each training and testing process are normalized using the respective upper bound of *Plausible Value*. Each value of marker, which extends that upper bound, obtains the value 1. The individually trained networks represent a generalized cancer occurrence prediction, disregarding specific type of cancer and based only on one specific tumor marker.

4.1. Case study: Breast Cancer – C125, C153, C199 and CEA marker Group

One example of a trained neural network (with 6 neurons in hidden layer) for tumor marker C125 is presented in Figure 2. The x-axis represents the normalized values of tumor marker C125 and the y-axis represents cancer possibility. Network-output values greater than 0,5 (middle line) are interpreted as cancer occurrence.

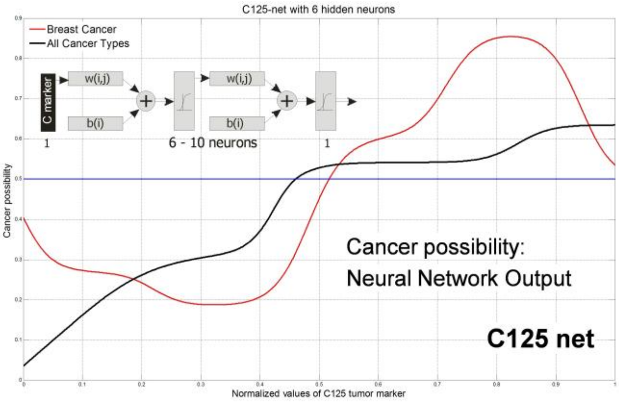

Figure 2. Output of individual trained neural network for tumor markers C125.

The networks were trained with 2598 datasets for C125 marker, with 2442 datasets for C153 marker, with 4519 dataset for C199 marker and with 7153 dataset for CEA marker. The datasets contain data with different cancer types from C00 to C96 ICD 10 code (44%) and data without cancer occurrence (56%).

Additionally, we trained these networks with smaller sets of data for one specific type of cancer disease. Figure 2 presents an example of trained C125 representing breast cancer.

Generally the network trained for all cancer types is more pessimistic, it means this network predicts cancer possibility greater than 0,5 for smaller values of tumor marker as the network trained for one special cancer type. In the example in Figure 2 the threshold points of networks trained for all cancer types is 0,46 (75,9 U/ml) for C125 marker. For marker C153 the threshold was 0,32 (34,8 U/ml), for C199 0,55 (73,1 U/ml) and or CEA 0,25 (14,1 ng/ml). The threshold points of networks trained for breast cancer are 0,52 (85,5 U/ml) for C125, 0,44 (48,2 U/ml) for C153, 0,42 (55,8 U/ml) for C199 and 0,37 (20,4 ng/ml) for CEA.

(Values for Markers C153, C199 and CEA are not shown in Figure 2).

The regressions between network outputs trained for all cancer types and breast cancer type are 0.69, 0.88, 0.88, and 0.91 for C125, C153, C199 and CEA, respectively. The influence of networks trained in this way on the final diagnosis prediction will be discussed in the next section.

The input of parallel coupled C_{nets} is the vector of tumor marker $\boldsymbol{C} = (C_1, \ldots, C_m)$, where some C_i are missing. When the tumor marker value in vector $\boldsymbol{C}$ is available, then the adequate C_{net} calculates the predicted cancer possibility. When a marker value in vector $\boldsymbol{C}$ is not available, then the output of C_{net} is set to -1. The individually calculated output values of C_{nets} can be aggregated in many different ways. We compare three methods of aggregation:

1. Maximum value of all individual network outputs: $C_{net}(\boldsymbol{C}) = \max\{ C^{i}_{net}(C_i) | i = 1, .., m\}$
2. Average value of all individual network outputs, without missing values: $C_{net}(\boldsymbol{C}) = \text{avg}\{ C^{i}_{net}(C_i) | i = 1, .., m \;\&\; C_i \neq -1\}$
3. $\text{Net}_{aggregation}$ - neural network trained on individual networks outputs (this neural network can be trained with data of only one chosen cancer type $Cancer^k$). $C_{net}(\boldsymbol{C}) = \text{net}_{aggregation}(C^{i}_{net} | i = 1, .., m)$

Other interesting possibility is using the thresholded values of outputs of individual networks (i.e. if $C_{net}(C) < 0{,}5$ then $C_{net}(C) = 0$, else $C_{net}(C) = 1$) as input for the perceptron type network.

We use one aggregation type in the full system. In case of max aggregation: If only one marker of the marker group shows a greater value than the aggregation has yielded this value is taken.

The diagnosis prediction based on aggregation of separately cancer predictions of individual marker networks C_{net} is not sufficient for generalization of cancer occurrence. It is necessary to reinforce the information coming from data of whole group of markers. Therefore two neural networks with cumulative marker groups are added. These networks will be trained only for a specific cancer type. When the tumor marker value in vector $\boldsymbol{C}$ is not available, then this value is set to -1. Based on this assumption we can generate training sets for a specific cancer type ($Cancer^k$) and train the neural network: A feed forward neural network with 16-20 hidden neurons and tansig/linear activation functions (C^{group}_{net}) with a vector $\boldsymbol{C}$ on input and diagnosis of tumor occurrence on output. This network can be used additionally to the individually trained networks for diagnosis prediction without and with estimated missing values of vector $\boldsymbol{C}$.

5. SYSTEM FOR PREDICTION OF MISSING TUMOR MARKER VALUE

5.1. Estimation of missing tumor marker values based on other tumor marker

Estimation of missing values of tumor markers can be done by values of other tumor markers from tested marker group $\boldsymbol{C}$. It can be accomplished by preparation of pair wise trained neural networks C^{ij}_{net} where the pattern set includes values of C_i marker as input and values of C_j marker as output in case both marker values are availibale. Not all tumor marker values can be predicted with sufficient level of plausibility.

In our case study we use feed forward neural networks with 5-10 hidden neurons for the estimation of a missing value of marker C_j based on a known value of marker C_i.

Figure 3 presents the prediction of C125 based on marker C153 and the prediction of marker value of CEA based on marker C199.

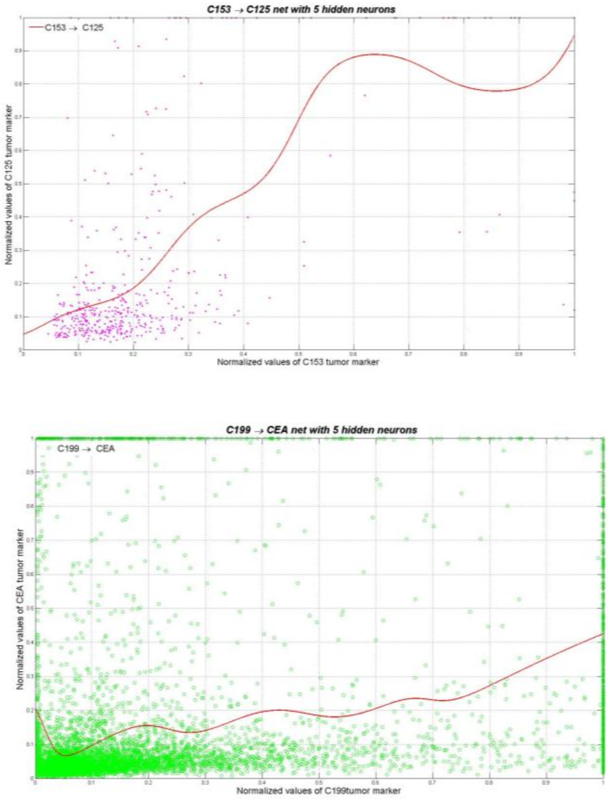

Figure 3. Outputs of neural networks for tumor markers C125 based on C153, and CEA based on C199.

In the first case we can observe clear dependency of marker C125 on marker C153 and in the second case a dependency does not exist.

In cases more than one value exists in vector $\boldsymbol{C}$, then the estimator for a missing value C_k can be computed as:

$$C_k = max\{C^{ik}_{net}(C_i) | i \neq k\} \quad \text{or} \quad C_k = avg\{C^{ik}_{net}(C_i) | i \neq k\}$$

where C_i represent the existing values markers in vector $\boldsymbol{C}$. The quality of the estimation of a missing value will highly depend on existing values in vector $\boldsymbol{C} = (C_1, \ldots, C_m)$.

5.2. Estimation of missing tumor marker values based on blood parameters.

Typically in labor blood examination 27 blood parameters such as HB, WBC, HKT, MCV, RBC, PLT, KREA, BUN, GT37, ALT, AST, TBIL, CRP, LD37, HS, CNEA, CMOA, CLYA, CEOA, CBAA, CHOL, HDL, CH37, FER, FE, BSG1, TF and tumor markers such as AFP, C125, C153, C199, C724, CEA, CYFRA, NSE, PSA, S100, SCC, TPS etc. are measured. For each parameter and marker are experimentally established upper and lower bounds of values. We divide the values range of marker $\boldsymbol{C}$ and blood parameter $\boldsymbol{P}$ into k non-overlapping intervals, called classes.

In our case study we define four classes (k = 4). *Class 1* includes all values less than *Normal Value* of marker or blood parameter, *Class 2* includes all values between *Normal Value* and *Extreme Normal Value* of marker or blood parameter, *Class 3* includes values between *Extreme Normal Value* and *Plausible Value* of marker or blood parameter and *Class 4* include all values greater than *Plausible Value*. These classes and their limits are used in normalizing process of parameter and marker values. In normalizing process, we replace the missing value with the value -1.

The system consists of three heterogeneous parallel-coupled artificial neural networks and a decision-making system based on aggregation rules (Jacak et al., 2010a).

The input and output values of each network for training and testing are normalized using the respective upper bound of *Plausible Value*. Each value of parameter or marker, which is greater than this upper bound, obtains the normalized value 1.

The general marker value estimation system contains three neural networks.

- Feed forward neural network (*FF*) with p inputs (normalized values of blood parameter vectors $\boldsymbol{P}$) and one output, normalized values of marker C_i
- Pattern recognition neural network (*PR*) with p inputs (normalized values of blood parameter vectors $\boldsymbol{P}$) and k outputs, k-dimensional binary vector coding classes of marker C_i
- Combined feed forward neural network (*FC*) with p inputs (normalized values of blood parameter vectors $\boldsymbol{P}$) and two outputs: normalized values of marker C_i (as in network FF), and normalized classes of marker C_i :

All neural networks have one hidden layer and tan-sigmoid or log-sigmoid transfer function. The output values of neural networks belong usually to interval [0, 1].

Based on the neural networks calculated estimation of marker value we can establish four hypotheses x_1, x_2, x_3, x_4 concerning the class of marker. For each hypothesis x_1, x_2, x_3, x_4 the possibility value is calculated too. These hypotheses are to be verified for finding the maximal possible prediction. (Jacak et al., 2010b)

The empirical test shows that the best results are achieved with networks having 40-60 neurons of hidden layer.

It can be expected that not all markers can be predicted with good quality. The examples of regression between blood parameters test data and estimation system output for tumor markers C153 (regression 0,71) and CEA (regression 0,53) are presented in Figure 4.

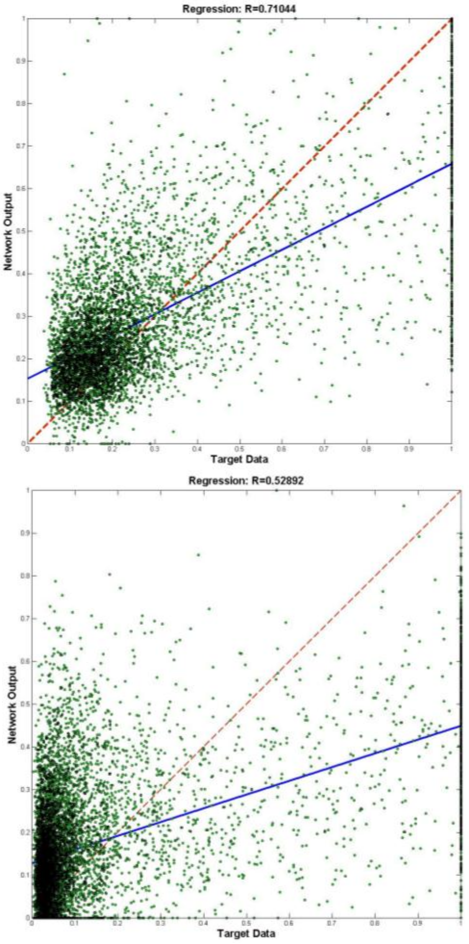

Figure 4: Regression between test data and predicted tumor marker values for markers C153 (first) and CEA (second).

In this method of value estimation, the quality of estimation will highly be dependent on missing values of marker in vector $\boldsymbol{C} = (C_1, \ldots, C_m)$.

The two methods presented have different properties and it will be necessary to combine both estimations results for a final prediction of cancer occurrence. We aggregate the outputs of the three diagnosis prediction networks (without estimation of missing values, with estimation of missing values based on other existing markers values and with estimation of missing values based on blood parameters) by the applying the maximum function. The result of prediction quality is presented in the next section.

6. RESULTS AND COMPARISON OF PREDICTION QUALITY OF DIFFERENT APPROACHES

For comparison between previously described methods of breast cancer prediction based on marker group ***C***= (C125, C153, C199, CEA) we have prepared a test data set containing 695 positive cases (diagnose coded as 1) and 765 negative cases (diagnose coded as 0). By assumption that the general probability of positive and negative cancer occurrence is 0,5. We can estimate the probability P(1/1) (true positives) and P(0/0) (true negatives) for the various systems.

The results of prediction are presented in table 1.

Table 1: Prediction of cancer occurrence

System	***P (correct)***	***P (1/1)***	***P (0/0)***
Individually trained networks without estimation of missing values	0,63	0,29	0,93
FF network with **C** vector, no estimation of missing values	0,67	0,45	0,89
Individually trained networks with estimation of missing values, based on existing additional values in vector **C** (max as aggregation function)	0,63	0,35	0,90
FF network with **C** vector, with estimation of missing values based on existing additionalr values in vector **C** (max as aggregation function)	0,66	0,57	0,74
Individual trained networks with estimation of missing values based on blood parameters values	0,66	0,57	0,73
FF network with **C** vector, estimation of missing values based on blood parameters	0,63	0,42	0,80
Aggregated prediction based on three network outputs (maximum aggregation function applied)	0,70	0,77	0,63

Diagnosis prediction performed without estimation of missing values of marker values works increases the probability P(0/1) of false negatives (see Figure 5). Prediction based on missing value estimation decreases false negatives rate but increases false positives rate. All confusion matrices of chosen experiments are presented in Figures 5-8. The whole system increases the probability of correct cancer diagnosis and decreases the false positives rate. The confusion matrix of the overall system for breast cancer diagnosis is presented in Figure 8.

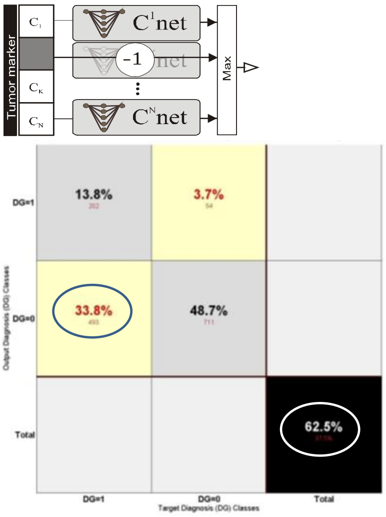

Figure 5. Confusion matrix of breast cancer diagnosis based on C125, C199, C153 and CEA marker group without estimation of missing values. False positives rate is 34%.

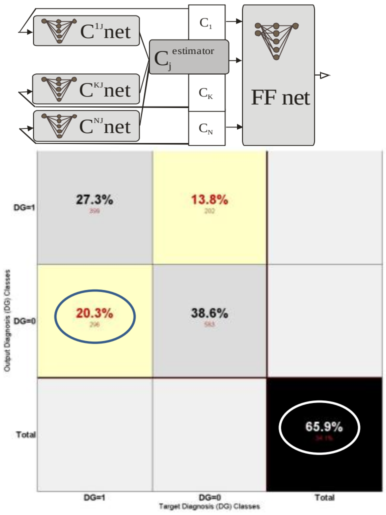

Figure 6. Confusion matrix of breast cancer diagnosis based on C125, C199, C153 and CEA marker group with marker based estimation of missing values of markers. The false positives rate is 20%.

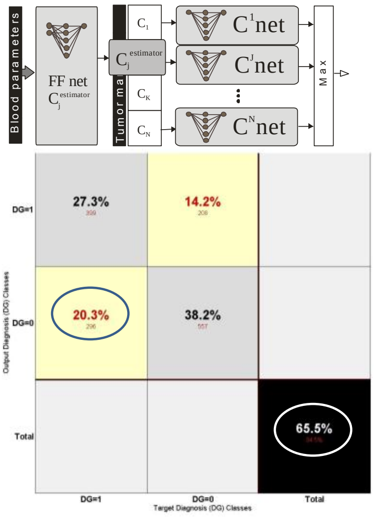

Figure 7. Confusion matrix of breast cancer diagnosis based on C125, C199, C153 and CEA marker group with blood parameters based estimation of missing values of markers. False positives rate is 20 %.

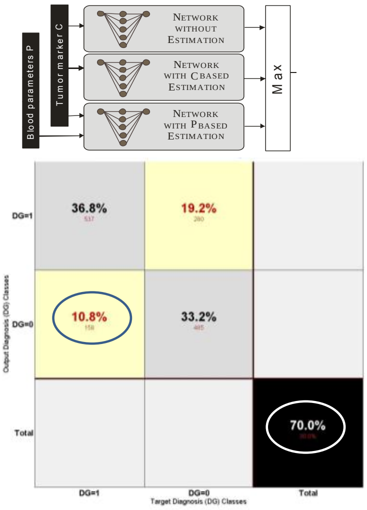

Figure 8. Confusion matrix of breast cancer diagnosis based on aggregated outputs of 3 networks with different estimation of missing values. The false positive rate is 11%.

REFERENCES

Astion, M.L., Wilding P., 1992, Application of neural networks to the interpretation of laboratory data in cancer diagnosis. *Clinical Chemistry*, Vol 38, 34-38.

Djavan, et al., 2002. Novel Artificial Neural Network for Early Detection of Prostate Cancer. Journal of Clinical Oncology, Vol 20, No 4, 921-929

Djavan, B., Remzi, M., Zlotta, A., Seitz, C., Snow, P., Marberger, M., 2002, Novel Artificial Neural Network for Early Detection of Prostate Cancer. *Journal of Clinical Oncology*, Vol 20, No 4, 921-929

Harrison, R.F., Kennedy, R.L., 2005, Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. *Ann Emerg Med*.; 46(5):431-9.

Jacak W., Proell K., 2010a, Data Driven Tumor Marker Prediction System, *Proceedings of EMSS 2010*, Fes, Marokko

Jacak W., Proell K., 2010b, Neural Network Based Tumor Marker Prediction, *Proceedings of BroadCom 2010*, Malaga, Spain

Liparini, A., Carvalho, S., Belchior, J.C., 2005, Analysis of the applicability of artificial neural networks for studying blood plasma: determination of magnesium on concentration as a case study. *Clin Chem Lab Med*.; 43(9):939-46