Neural Network Based Tumor Marker Prediction

Witold Jacak(a), Karin Proell(b)

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

***Abstract* - In this paper a system for the prediction of tumor marker values based on standard blood examination is presented. Several neural networks are used to learn from blood examination measurements and predict tumor markers in case these values are missing. In a post processing step the predicted values are evaluated in a fuzzy logic like style against different hypotheses and the best hypothesis is used to optimize the predicted values and its plausibility. The predicted values of markers are compared to outputs of the second system. This system is called cascade neural network. It uses the outputs of several networks as input for new cascaded connected neural networks. These predicted values can then be used as input for a second system to support decision making in cancer diagnosis. A variety of experiments with tumor marker C153 shows that we can get a quite good prediction accuracy. Our experiments are based on hundreds of samples of up to 27 different features (blood parameters) per vector. We try to predict distinct values, classes of values and a combination of classes and values for specific marker types.**

Keywords: neural network, tumor marker prediction, decision support system

I. INTRODUCTION

Tumor markers are substances produced by cells of the body in response to cancerous but also to noncancerous conditions. They can be found in body liquids like blood or in tissues and can be used for detection, diagnosis and treatment 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. Examples of tumor markers include CA 125 (in ovarian cancer), CA 153 (in breast cancer), CEA (in ovarian, lung, breast, pancreas, and gastrointestinal tract cancers), and PSA (in prostate cancer). Although an abnormal tumor marker level may suggest cancer, tumor markers are not sensitive or specific enough for a reliable cancer diagnosis. But abnormally altered tumor marker values indicate a need for further medical examination.

During blood examination only a few tumor marker values are tested 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 on medical data [7]. For example neural networks were applied to differentiate benign from malignant breast conditions base on blood parameters[1], for diagnosis of different types of liver disease [7], for early detection of prostate cancer [2,5], for studies on blood plasma [4] or for prediction of acute coronary syndromes [3].

In this work we present a novel heterogeneous neural network based system that can be used for tumor marker value prediction. We use an n-dimensional vector of blood parameter values of a several hundred patients as input and train three neural networks in parallel to predict distinct values, classes of values and a combination of classes and values for specific marker types. In a post-processing step the outputs of all networks are adjusted by a fuzzy logic like decision system to obtain the most possible prediction. Unfortunately neural networks are unable to work properly with incomplete data; missing values however are a common problem in medical datasets. It may be that a specific medical procedure was not considered necessary in a particular case or that the procedure was taken in a different laboratory with the values not available in the patient record, or that the measurement was taken but not recorded due to time constraints.

II. GENERAL CANCER DIAGNOSIS SUPPORT SYSTEM

We focus our considerations on the design of a complex decision support system for early recognition of possibility of cancerous diseases. The system consists of two components; a Tumor Marker Prediction System and a Diagnosis Support System (see Fig.1). Both systems use several heterogeneous artificial neural networks in parallel. The Tumor Marker Prediction System is in support of the Diagnosis Support System, which uses input data coming from the vector of tumor marker values $C = (C_1, \ldots, C_m)$ and calculates the possibility of presence of a cancerous disease in general and the possibility of a specific tumor type in particular (tumor types are coded according to ICD 10 system).

The output values are evaluated against different hypotheses and the best hypothesis is used to optimize the predicted diagnosis and its plausibility.

An important issue for the Diagnosis Support System is data incompleteness. We need thousands of vectors of tumor markers for training and evaluation of the neural networks. Those vectors do not only contain distinct values of various marker types but also ranges of marker values, so called

classes (e.g. one class for normal values, one for extreme normal values, one for beyond normal but plausible values and one for extreme values). Many of the available marker vectors consist of just a few measurements and cannot be used as training data for neural networks without further processing.

One approach to overcome this problem is to restrict the analysis only to vectors with complete data but this leads to very small sample sets. Another option is to extend the number of input values to all parameters of the blood examination, thus including also non- marker values, using a whole blood parameter vector $p = (p_1, \ldots, p_n)$ as input. Frequently also this vector is incomplete too. For this reason the Tumor Marker Prediction System is connected ahead the Diagnosis Support System, which uses complete or partial complete blood parameter vectors p of patients to train a couple of neural networks for estimating values or classes of values for tumor markers. The output values are evaluated against different hypotheses and the best hypothesis is used to optimize the predicted value. Additionally a possibility value for estimated marker value is calculated. This Marker Value Prediction System could also be used as a stand-alone system for a rapid estimation of marker values.

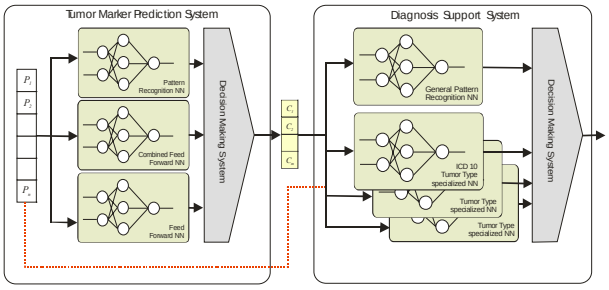

Figure 1. Architecture of Data Driven Cancer Diagnosis Support System

III. SYSTEM FOR PREDICTION OF TUMOR MARKER VALUES

Typically, the blood examination results in maximal 27 values of 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 about 35 tumor markers such AFP, C125, C153, C199, C724, CEA, CYFRA, NSE, PSA, S100, SCC, TPS etc. Different labs may consider slightly different marker levels to be normal or abnormal. This can depend on a number of factors, including a person's age and gender, which test kit the lab uses, and how the test is done. For each parameter and marker we use reference ranges [4]. We divide the value range of marker C and blood parameter 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 the 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 the Plausible Value.

For each *Class i* of marker values we calculate the average value μ_i and the standard deviation σ_i. For example the respective values of marker C 153 calculated from patient data are presented in Table I.

TABLE I. EXAMPLE OF BLOOD PARAMETER RANGES

Code	*Class*	*μ*	*σ*	*Min*	*Max*	*dmax*
C153	1	15,54	5,02	2	25	100
C153	2	33,59	6,73	26	50	100
C153	3	68,48	13,78	56	100	100
C153	4	162,20	321,22	101	10000	100

A. Architecture of marker value prediction system

We consider two systems for marker value prediction. The first system consists of three heterogeneous artificial neural networks connected in parallel and a decision-making system based on aggregation rules. The second one contains four heterogeneous artificial neural networks connected in a cascaded .

Normalizing: The input and output values for training and testing of each network are normalized using the respective upper bound of Plausible Value. Each value of parameter or marker, which is greater than its upper bound, obtains the normalized value 1. Such a normalizing process guarantees that all values are mapped to interval [0, 1].

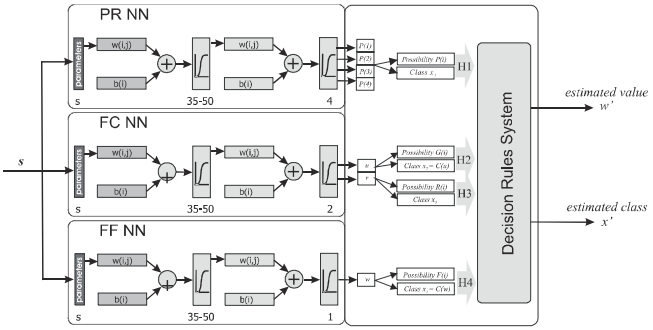

Figure 2. Rules based tumor marker value prediction system

B. Rules Based Prediction System Structure

The general marker value estimation system contains of three neural networks (see Figure 2).

- Feed forward neural network (FF) with p inputs (normalized values of blood parameter vectors p) and one output, normalized values of marker C
- Pattern recognition neural network (PR) with p inputs (normalized values of blood parameter vectors p) and k outputs, k-dimensional binary vector coding classes of marker C
- Combined feed forward neural network (FC) with p inputs (normalized values of blood parameter vectors p) and two outputs: normalized values of

marker *C* (as in network FF), and normalized classes of marker *C* as:

$$NormClass^{j}(C) = j/k, \quad for = 1,...,k \tag{1}$$

All neural networks have one hidden layer and a tan-sigmoid or linear transfer function. The output values of neural networks usually range within the interval [0, 1].

For a given parameter vector *p* the three networks calculate different output data.

- The pattern recognition neural network PR produces the *k*-value vector $(P(i) \mid i = 1,..,k)$, where *P(i)* describes the possibility (in sense of fuzzy logic) of *Class i* of *C* marker connected to input parameter vector *p*. The supposed class of marker *C* is

$$x_1 = arg(max\{P(i) | i = 1,...,k\} \tag{2}$$

- The feed forward neural network FC uses value and class as inputs and generates two outputs: a value *r* interpreted as possibility of class of marker *C* and a value *u* representing the prediction of the normalized value of marker *C*. Value *u* and the known limits of the marker values are used for determination of class $x_2 = Class(u)$ indicating the class to which the output *u* belongs. Separately we calculate the possibility vector $(G(i) \mid i = 1,...,k)$ as indirect distance to the average value of each class of marker *C*

$$G(i) = 1-(|u-\mu_i|/d_{max}) \text{ for } i = 1,...,k. \tag{3}$$

where d_{max} denotes maximal distance between values of marker *C*. For *r* values we calculate the possibility vector $(R(i) \mid i = 1,...,k)$ as

$$R(i) = 1-|r-i/k| \text{ for } i = 1,...,k. \tag{4}$$

The class of marker *C* that will be suggested is

$$x_3 = arg(max\{R(i) | i = 1,...,k\}). \tag{5}$$

- The feed forward neural network FF generates a normalized value *w* of marker *C*. Based on value *w* and the limits of each class we can calculate the class $x_4 = Class(w)$ to which the output of FF network belongs. Separately we calculate the possibility vector $(F(i) \mid i = 1,...,k)$ as indirect distance to the average value of each *C* marker class

$$F(i) = 1-(|w-\mu_i|/d_{max}), \text{ for } i = 1,...,k. \tag{6}$$

We use of the Neural Network Toolbox™ of MATLAB® for designing, implementing, visualizing, and simulating the neural networks PR, FC and FF.

1) Evaluation and post processing method of rules based prediction system

Based on the calculated estimation of marker values we can establish four hypotheses x_1, x_2, x_3, x_4 for determination of classes. For each hypothesis x_1, x_2, x_3, x_4 the possibility value *P, G, R, F* is calculated too.

These hypotheses should be verified to find the maximal possible prediction. This is done by testing a couple of aggregation functions *V* on the possibility values of each hypothesis. Those aggregation functions are similar to aggregation rule in fuzzy logic decision-making systems. In this experiment we use four kinds of functions *V*:

$$Minimum: V(x_i) = min\{P(x_i)\ G(x_i)\ R(x_i)\ F(x_{i_j})\} \tag{7}$$
$$Product: V(x_i) = P(x_i)\ G(x_i)\ R(x_i)\ F(x_i) \tag{8}$$
$$Average: V(x_i) = (P(x_i) + G(x_i) + R(x_i) + F(x_i))/4 \tag{9}$$
$$Count: V(x_i) = Count\_of(x_i) \tag{10}$$

Count is the number of identical hypothesis.

The maximal value of aggregation functions determines the new predicted class. i.e.

$$x_{new} = arg\ (max\{V(x_i) \mid i = 1,..,4\}) \tag{11}$$

where $V(x_i) = V(P(x_i), G(x_i), R(x_i), F(x_i))$ is the evaluation function of arguments $P(x_i)$, $G(x_i)$, $R(x_i)$and $F(x_i)$.This kind of evaluation method leads to four decision composition rules, namely *MaxMin, MaxProd, MaxAvg* and *MaxCount* are well known in fuzzy decision systems.

Estimation of predicted marker value: Based on the determined new class the estimation of a marker value is performed. If the evaluation function *V* used in decision composition rule has a value greater than τ then the new estimated value of a marker is equal to the average value of *w* and *u*. i.e.

$$w_{new} = (w + u)/2 \quad \text{if } V(x) > \tau$$
$$w_{new} = (w + u + \mu_{xnew})/3 \quad \text{other}$$

C. Cascade Neural Networks Prediction System Structure

The general marker value estimation system contains of four neural networks (see Figure 3).

- Feed forward neural network (FF) with *p* inputs and one output: the normalized values of marker *C*
- Pattern recognition neural network (PR) with *p* inputs and *k* outputs: *k*-dimensional binary vector coding classes of marker *C*

  Combined feed forward neural network (FC) with *p* inputs and two outputs: the normalized values of marker *C* (as in network FF), and the normalized classes of marker *C* as:

$$NormClass^{j}(C) = j/k, \quad for = 1,...,k \tag{13}$$

- Feed forward neural network (Cascade FF) with 3 inputs (post processed outputs of PR, FC and FF

networks) and one output: normalized values of marker *C* (as in network FF).

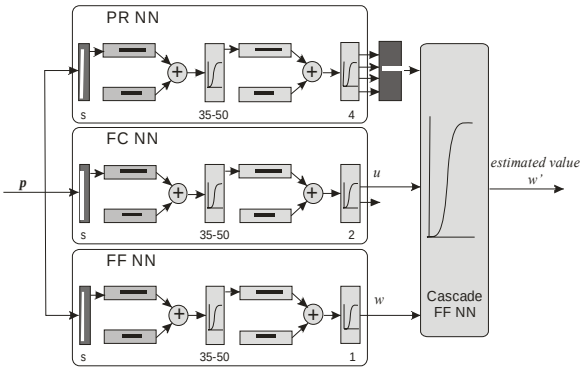

Figure 3. Cascade neural networks based tumor marker value prediction system

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

- For a given parameter vector *p* the parallel three networks calculate different output data.
- The pattern recognition neural network PR produces the *k*-value vector *(P(i) | i = 1,..,k)*, where *P(i)* describes the possibility of class *i* of marker *C* connected to the input parameter vector *p*. The cascade network needs the estimated value of a tumor marker. For this reason a post processing step of the output value of the net PR is needed. We calculate the winner class of marker *C* as $i = arg(max\{P(i)|i=1,...,k\})$ and replace the class number with the average of the marker values in this class. The post processed output of PR net is also

  $v=Avg(i)$ where $i = arg(max\{P(i)|i=1,...,k\})$ (14)
- The feed forward neural network FC uses value and class as inputs and generates two outputs: the first value *r* is interpreted as possibility of class of marker *C* and the second value *u*, representing the predicted normalized value of marker *C*. We take in consideration only value *u* as input for the cascade network.
- The feed forward neural network FF generates a normalized value *w* of marker *C*. This value is directly forwarded as input to the cascade network.
- The cascade network is a feed forward network with the input vector *(v, u, w)* and one output containing the calculated normalized values of marker *C*.

IV. CASE STUDY: C 153 TUMOR MARKER

A. Training and Test Setup

We have taken the complete data set with 20 blood parameters from patient data: The input vector *p* contains complete data of following blood parameters *p* = (HB, WBC, HKT, MCV, RBC, PLT, KREA, BUN, GT37, ALT, AST, TBIL, CRP, LD37, HS, CNEA, CMOA, CLYA, CEOA, CBAA). We use 4427 samples as Learning Pattern Set for the neural networks system and 491 independent samples as Test Set.

B. Experiments and Results

For determination of an appropriate number of neurons in the hidden layer of the feed-forward networks we performed small batch set trainings of networks using different numbers of neurons.

The empirical test shows that networks are best performing with 40-60 neurons in the hidden layer. We finally decided to use neural networks having one hidden layer with 50 neurons and tan-sigmoid activation functions. These are the neural networks settings used:

- Feed forward neural network (FF) with 20 inputs (normalized values of parameter vectors *p*) and one output (normalized values of C153 marker)
- Pattern recognition neural network (PR) with 20 inputs (normalized values of parameter vectors *p*) and four outputs (four dimensional binary vectors coding classes of C153 marker)
- Feed forward neural network (FC) with 20 inputs (normalized values of parameter vectors *p*) and two outputs (normalized values of C153 marker and normalized classes of C153 with: *Class 1* = 0.25; *Class 2* = 0.5; *Class 3* = 0.75; *Class 4* = 1.0)
- Cascade feed forward neural network (Cascade FF) with 3 inputs (output values of FF, FC and PR networks), 1 hidden layer with 20 neurons and one output (normalized values of C153 marker)

All four neural networks were trained with Levenberg-Marquardt algorithm and a validation failure factor 6.

We conduct two experiments. In the first experiment, the test outputs of all network is post processed by aggregation rules based on the decision system and finally compared with original values and classes of C153 tumor marker from test set.

1) Result of rules based marker value prediction system

The regression functions between the outputs of three networks (PR, FC, FF), post processed final estimation and test C153 values are presented in Figure 4. It can be observed that regression of the rules based final estimation of C153 value is greater (R = 0.732) than the individual estimation of separate networks (R= 0.62, R = 0.70, and R = 0.68 for PR, FC, and FF network respectively). The bounded area in the right corner indicates fatal mismatching i.e. originally large values of C153 marker and small-predicted values.

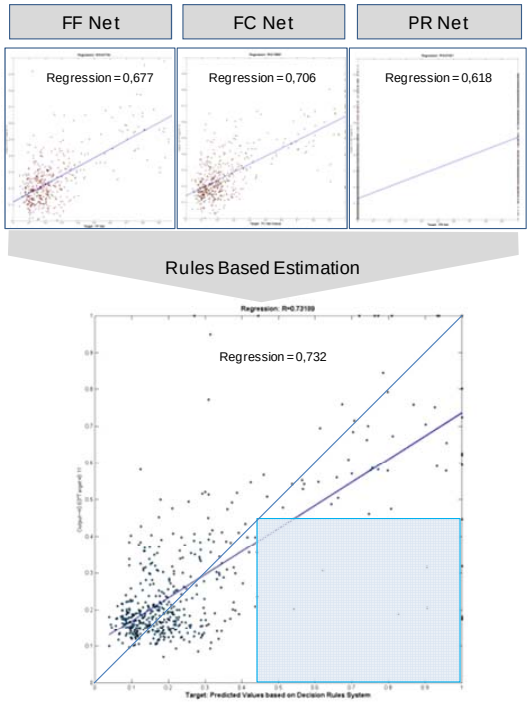

Figure 4. Results of cascade neural networks system for tumor marker C153

Based on the predicted values of tumor marker C153 we calculate the matching and mismatching ratios for each class of marker. Matching and mismatching of classes between the test data of marker C153 classes and the predicted classes of different networks are presented as confusion matrixes. After post processing, we obtain 72 % matching of four classes, whereas the original networks range between 59% and 66% matching cases. The ratio of fatal mismatching is 2.9 %.

Additionally we conducted an experiment with a reduced number of training classes of marker C153. We merge *Class 1* and *Class 2* of marker C153 into a new *Class I* and *Class 3* and *Class 4* into a new *Class II*. That means that all values of tumor marker C153 less than Extreme Normal Value determine *Class I* and values greater than Extreme Normal Value determine *Class II*. Normalization of blood parameters remains unchanged. Full training and test of networks was performed on input data modified in this way. We obtain a quite good ratio (94.7 %) of matching cases and the ratio of fatal mismatching is reduced to 2 %.

The dependency of quality of estimation on different composition rules is presented in Table II. It is shown that the *MaxAvg* rule produces the best results.

TABLE II. CONFUSION VALUES AFTER APPLICATION OF COMPOSITION RULES

Rule	PR Net	FF Net	FC Value Net	FC Class Net	New Value	Two Class (Class based)	Two Class (Value based)
MaxCount	68,2	59,5	66,1	63,5	67,6	93,7	94,5
MaxAvg	68,2	59,5	66,1	63,5	71,7	94,7	94,7
MaxMin	68,2	59,5	66,1	63,5	68,2	89,8	94,3
MaxProd	68,2	59,5	66,1	63,5	68	90,2	94,5

1) Result of marker value prediction system based on cascade neural networks

The output values of separate PR, FC, and FF networks are used to train the cascade FF network.

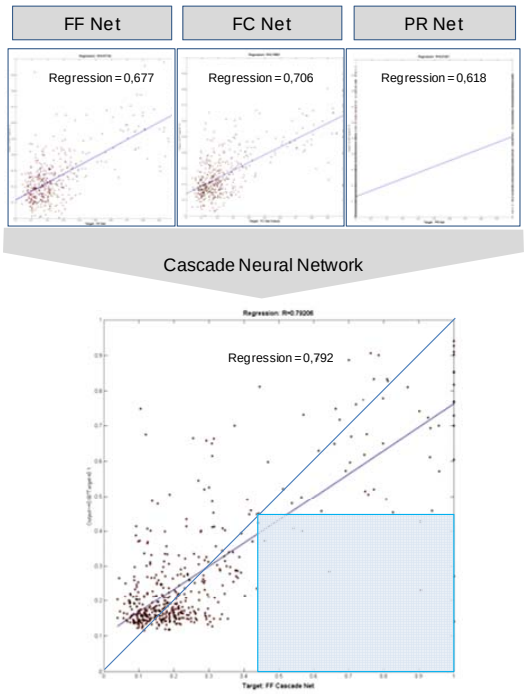

Figure 5. Results of cascade neural networks system for tumor marker C153

The regression functions between the outputs of three networks (PR, FC, FF), and the output of the cascade network, which gives the final estimation and test of C153 values are presented in Figure 5. It can be observed that the regression of the final estimation of C153 value being rules based is larger (R =0.792) than the individual estimation of the separate networks.

The prediction system which is based on cascade neural networks structure generates better estimation of marker values than the rules based system but the number of fatal mismatching i.e. originally large values of C153 marker and small-predicted values, is larger.

V. FINAL REMARKS

In this paper we focused our presentation on experiments predicting tumor marker C153. Similar results were obtained for markers CEA and C199. In the next experiments we will continue testing further marker types depending on the availability of sufficient patient data. It can be expected that not all markers can be predicted with similar performance as C153.

In a further step the system will be extended by predicting combinations of markers as output of neural networks. Short tests show that such combinations can increase the quality of prediction. Moreover it is seems to be more effective to create different classes of markers as used in the recent experiment based on fuzzy c-means clustering methods. This prediction system will be an important component to the whole data driven cancer diagnosis support system.

ACKNOWLEDGMENTS

Special thanks goes to Prim. Herbert Stekel, M.D, Head of Institute of Laboratory Medicine, General Hospital Linz, Austria for providing thousands of samples of anonymous blood examination data for our experiments.

REFERENCES

[1] 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.

[2] 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

[3] 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.

[4] Jacak, W., Proell, K., "Data Driven Tumor Marker Prediction System", unpublished

[5] 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

[6] Matsui, Y., Utsunomiya, N., Ichioka, K., Ueda, N., Yoshimura, K., Terai, A., Arai, Y., 2004, The use of artificial neural network analysis to improve the predictive accuracy of prostate biopsy in the Japanese population. Jpn J Clin Oncol. 34(10):602-7.

[7] Penny, W., Frost, D., 1996, Neural Networks in Clinical Medicine. Med Decis Making, Vol 16: 386-398

[8] Reibnegger, G., Weiss, G., Werner-Felmayer, G., Judmaier, G., Wachter, H., 1991, Neural networks as a tool for utilizing laboratory information: Comparison with linear discriminant analysis and with classification and regression trees. PNAS: vol. 88 no. 24, 11426-11430