Chapter 6
On the Identification of Virtual Tumor Markers and Tumor Diagnosis Predictors Using Evolutionary Algorithms

Stephan M. Winkler, Michael Affenzeller, Gabriel K. Kronberger,
Michael Kommenda, Stefan Wagner, Witold Jacak, and Herbert Stekel

Abstract. In this chapter we present results of empirical research work done on the data based identification of estimation models for tumor markers and cancer diagnoses: Based on patients' data records including standard blood parameters, tumor markers, and information about the diagnosis of tumors we have trained mathematical models that represent virtual tumor markers and predictors for cancer diagnoses, respectively. We have used a medical database compiled at the Central Laboratory of the General Hospital Linz, Austria, and applied several data based modeling approaches for identifying mathematical models for estimating selected tumor marker values on the basis of routinely available blood values; in detail, estimators for the tumor markers AFP, CA-125, CA15-3, CEA, CYFRA, and PSA have been identified and are discussed here. Furthermore, several data based modeling approaches implemented in HeuristicLab have been applied for identifying estimators for selected cancer diagnoses: Linear regression, k-nearest neighbor learning, artificial neural networks, and support vector machines (all optimized using evolutionary algorithms) as well as genetic programming. The investigated diagnoses of breast cancer, melanoma, and respiratory system cancer can be estimated correctly in up to 81%, 74%, and 91% of the analyzed test cases, respectively; without tumor markers up to 75%, 74%, and 87% of the test samples are correctly estimated, respectively.

Stephan M. Winkler · Michael Affenzeller · Gabriel K. Kronberger · Michael Kommenda · Stefan Wagner · Witold Jacak
Heuristic and Evolutionary Algorithms Laboratory,
University of Applied Sciences Upper Austria, School of Informatics,
Communication and Media, Softwarepark 11, 4232 Hagenberg, Austria

Herbert Stekel
Central Laboratory, General Hospital Linz, Krankenhausstraße 9, 4021 Linz, Austria

R. Klempous et al. (eds.), *Advanced Methods and Applications in Computational Intelligence*,
Topics in Intelligent Engineering and Informatics 6,
DOI: 10.1007/978-3-319-01436-4_6,

6.1 Introduction and Research Goals

In this chapter we present research results achieved within the research center *Heureka!*[1]: Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been analyzed in order to identify mathematical models for tumor markers and tumor diagnoses. We have used a medical database compiled at the blood laboratory of the General Hospital Linz, Austria, in the years 2005 – 2008: 28 routinely measured blood values of thousands of patients are available as well as several tumor markers; not all values are measured for all patients, especially tumor marker values are determined and documented if there are indications for the presence of cancer.

In Figure 6.1 the main modeling tasks addressed in this research work are illustrated: Tumor markers are modeled using standard blood parameters and tumor marker data; tumor diagnosis models are trained using standard blood values, tumor marker data, and diagnosis information, and alternatively we also train diagnosis estimation models only using standard blood parameters and diagnosis information.

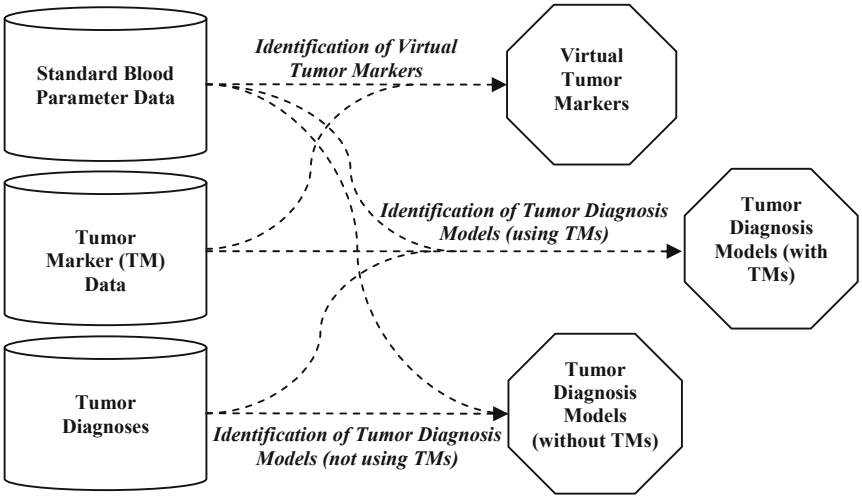

Fig. 6.1. Modeling tasks investigated in this research work: Tumor markers are modeled, and tumor diagnosis models are trained using standard blood values, diagnosis information, and optionally tumor marker data

6.1.1 Identification of Virtual Tumor Markers

In general, tumor markers are substances found in humans (especially blood and / or body tissues) that can be used as indicators for certain types of cancer. There are several different tumor markers which are used in oncology to help detect the presence of cancer; elevated tumor marker values can indicate the presence of cancer,

[1] Josef Ressel Center for Heuristic Optimization; `http://heureka.heuristiclab.com/`

but there can also be other causes. As a matter of fact, elevated tumor marker values themselves are not diagnostic, but rather only suggestive; tumor markers can be used to monitor the result of a treatment (as for example chemotherapy). Literature discussing tumor markers, their identification, their use, and the application of data mining methods for describing the relationship between markers and the diagnosis of certain cancer types can be found for example in [26] (where an overview of clinical laboratory tests is given and different kinds of such test application scenarios as well as the reason of their production are described), [10], [40], [56], and [57].

We have used data based modeling approaches (including enhanced genetic programming as well as other established data mining methods) for identifying mathematical models for estimating selected tumor marker values on the basis of routinely available blood values; in detail, estimators for the tumor markers *AFP*, *CA-125*, *CA15-3*, *CEA*, *CYFRA*, and *PSA* have been identified and are analyzed in this chapter. These tumor marker estimation models are also referred to as *virtual tumor markers*.

The documented tumor marker values are classified as "normal" (class 0), "slightly elevated" (class 1), "highly elevated" (class 2), and "beyond plausible" (class 3); this classification is done according to medical aspects using classification rules based on medical knowledge. In principle, our goal is to design classifiers for classifying samples into one of these classes. Still, in the context of the research work summarized here we have decided to produce classifiers that classify samples as "normal (belonging to class 0)" or "elevated (belonging to class 1, 2, or 3)"; i.e., we here document results for a simplified 2-class-classification problem.

6.1.2 Identification of Tumor Diagnosis Estimators

In addition, information about cancer diagnoses is also available in the AKH database: If a patient is diagnosed with any kind of cancer, then this is also stored in the database.

Our goal in the research work described here is to identify estimation models for the presence of the following types of cancer: Malignant neoplasms in the respiratory system (RSC, cancer classes C30–C39 according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)), melanoma and malignant neoplasms on the skin (Mel, C43–C44), and breast cancer (BC, C50).

Tumor markers are used optionally - on the one hand information about tumor markers values increases the accuracy of diagnosis estimations, on the other hand their acquisition is considered expensive and they are therefore not available by default.

We have applied two modeling methods for identifying estimation models for tumor markers and cancer diagnoses:

- Several machine learning methods (implemented in HeuristicLab [47]) have been used for producing classifiers, namely linear regression, k-nearest neighbor classification, neural networks, and support vector machines.

- Evolutionary algorithms have been applied for parameter optimization and feature selection. Feature selection is often considered an essential step in data based modeling; it is used to reduce the dimensionality of the datasets and often leads to better analyses. Additionally, each data based modeling method (except plain linear regression) has several parameters that have to be set before starting the modeling process. We have used evolutionary algorithms for finding optimal feature sets as well as optimal modeling parameters for models for tumor markers; details can be found in [54] and [52], e.g.
- Alternatively, we have applied genetic programming (GP, [28]) using a structure identification framework described in [4] and [50]. Genetic programming has been repeatedly used successfully for building formulas that describe the behavior of systems from measured data, see for example [4], [28], [32], or [50].

6.1.3 Organization of This Chapter

This chapter is structured in the following way: In Section 6.2 we give details about the data basis investigated in the research summarized here, in Section 6.3 we describe the modeling methods applied for identifying classification models for tumor data, and in Sections 6.4 and 6.5 we summarize empirical results achieved modeling tumor markers and tumor diagnoses using machine learning and evolutionary algorithms. This chapter is completed by a conclusion given in Section 6.6.

6.2 Data Basis

The blood data measured at the AKH in the years 2005–2008 have been compiled in a database storing each set of measurements (belonging to one patient): Each sample in this database contains an unique ID number of the respective patient, the date of the measurement series, the ID number of the measurement, and a set of parameters summarized in the Tables 6.1 and 6.2; standard blood parameters are stored as well as tumor marker values. Patients personal data (e.g. name, date of birth, etc.) where at no time available to the authors except the head of the laboratory.

In total, information about 20,819 patients is stored in 48,580 samples. Please note that of course not all values are available in all samples; there are very many missing values simply because not all blood values are measured during each examination.

Information about the blood parameters stored in this database and listed in Table 6.1 can for example be found in [6], [30], [34], [43], and [49].

In Table 6.2 we list those tumor markers that are available in the AKH database and have been used within the research work described here; in detail, we have analyzed data of the following tumor markers:

Table 6.1. Patient data collected at AKH Linz in the years 2005 – 2008: Blood parameters and general patient information

Parameter Name	*Description*	*Unit*	*Plausible Range*	*Number of Available Values*
ALT	Alanine transaminase, a transaminase enzyme; also called glutamic pyruvic transaminase (GPT)	U/l	[1; 225]	29,202
AST	Aspartate transaminase, an enzyme also called glutamic oxaloacetic transaminase (GOT)	U/l	[1; 175]	29,201
BSG1	Erythrocyte sedimentation rate; the rate at which red blood cells settle / precipitate within 1 hour	mm	[0; 50]	10,201
BUN	Blood urea nitrogen; measures the amount of nitrogen in the blood (caused by urea)	mg/dl	[1; 150]	28,995
CBAA	Basophil granulocytes; type of leukocytes	G/l	[0.0; 0.2]	21,184
CEOA	Eosinophil granulocytes; type of leukocytes	G/l	[0.0; 0.4]	21,184
CH37	Cholinesterase, an enzyme	kU/l	[2; 23]	7,266
CHOL	Cholesterol, a component of cell membranes	mg/dl	[40; 550]	14,981
CLYA	Lymphocytes; type of leukocytes	G/l	[1; 4]	21,188
CMOA	Monocytes; type of leukocytes	G/l	[0.2; 0.8]	21,184
CNEA	Neutrophils; most abundant type of leukocytes	G/l	[1.8; 7.7]	21,184
CRP	C-reactive protein, a protein; inflammations cause the rise of CRP	mg/dl	[0; 20]	22,560
FE	Iron	ug/dl	[30; 210]	6,792
FER	Ferritin, a protein that stores and transports iron in a safe form	ng/ml	[10; 550]	2,428
GT37	γ-glutamyltransferase, an enzyme	U/l	[1; 290]	29,173
HB	Hemoglobin; a protein that contains iron and transports oxygen	g/dl	[6; 18]	29,574
HDL	High-density lipoprotein; this protein enables the transport of lipids with blood	mg/dl	[25; 120]	7,998
HKT	Hematocrit; the proportion of red blood cells within the blood volume	%	[25; 65]	29,579
HS	Uric acid, also called urate	mg/dl	[1; 12]	24,330
KREA	Creatinine; a chemical by-product produced in muscles	mg/dl	[0.2; 5.0]	29,033
LD37	Lactate dehydrogenase (LDH); an enzyme that can be used as a general marker of injuries to cells	U/l	[5; 744]	28,356
MCV	Mean corpuscular / cell volume; the average size (i.e., volume) of red blood cells	fl ($=\mu m^3$)	[69; 115]	29,576
PLT	Thrombocytes, also called platelets; irregularly-shaped cells that do not have a nucleus	G/l	[25; 1,000]	29,579
RBC	Erythrocytes, red blood cells; the most abundant type of blood cells that transport oxygen	T/l	[2.2; 8.0]	29,576
TBIL	Bilirubin; the yellow product of the heme catabolism	mg/dl	[0; 5]	28,565
TF	Transferrin; a protein, delivers iron	mg/dl	[100; 500]	2,017
WBC	Leukocytes, also called white blood cells (WBCs); cells that help the body fight infections or foreign materials	G/l	[1.5; 50]	29,585
AGE	The patient's age	years	[0; 120]	48,580
SEX	The patient's sex	f/m		48,580

Table 6.2. Patient data collected at AKH Linz in the years 2005 – 2008: Selected tumor markers

Marker Name	*Unit*	*Normal Range*	*Elevated Range*	*Plausible Range*	*Number of Available Values*
AFP	IU/ml	[0.0; 5.8]	]5.8; 28]	[0.0; 90]	5,415
CA 125	U/ml	[0.0; 35.0]	]35.0; 80]	[0.0; 150]	3,661
CA 15-3	U/ml	[0.0; 25.0]	]25.0; 50.0]	[0.0; 100.0]	6,944
CEA	ng/ml	[0.0; 3.4]	]3.4; 12.0]	[0.0; 50.0]	12,981
CYFRA	ng/ml	[0.0; 3.3]	]3.3; 5.0]	[0.0; 10.0]	2,861
PSA	ng/ml	[0.0; 2.5] (age $\leq$ 50) [0.0; 2.5] (age 51–60) [0.0; 2.5] (age 61–70) [0.0; 2.5] (age $\geq$ 71)	]2.5; 10.0] (age $\leq$ 50)]2.5; 10.0] (age 51–60)]2.5; 10.0] (age 61–70)]2.5; 10.0] (age $\geq$ 71)	[0.0; 20.0]	23,130

- **AFP:** Alpha-fetoprotein (AFP, [36]) is a protein found in the blood plasma; during fetal life it is produced by the yolk sac and the liver. In humans, maximum AFP levels are seen at birth; after birth, AFP levels decrease gradually until adult levels are reached after 8 to 12 months. Adult AFP levels are detectable, but usually rather low.
  For example, AFP values of pregnant women can be used in screening tests for developmental abnormalities as increased values might for example indicate open neural tube defects, decreased values might indicate Down syndrome. AFP is also often measured and used as a marker for a set of tumors, especially endodermal sinus tumors (yolk sac carcinoma), neuroblastoma, hepatocellular carcinoma, and germ cell tumors [17]. In general, the level of AFP measured in patients often correlates with the size / volume of the tumor.
- **CA 125:** Cancer antigen 125 (CA 125) ([55]), also called carbohydrate antigen 125 or mucin 16 (MUC16), is a protein that is often used as a tumor marker that may be elevated in the presence of specific types of cancers, especially recurring ovarian cancer [39]. Still, its use in the detection of ovarian cancer is controversial, mainly because its sensitivity is rather low (as documented in [41], only 79% of all ovarian cancers are positive for CA 125) and it is not possible to detect early stages of cancer using CA 125.
  Even though CA 125 is best known as a marker for ovarian cancer, it may also be elevated in the presence of other types of cancers; for example, increased values are seen in the context of cancer in fallopian tubes, lungs, the endometrium, breast and gastrointestinal tract.
- **CA 15-3:** Mucin 1 (MUC1), also known as cancer antigen 15-3 (CA 15-3), is a protein found in humans; it is used as a tumor marker in the context of monitoring certain cancers [38], especially breast cancer. Elevated values of CA 15-3 have been reported in the context of an increased chance of early recurrence in breast cancer [25].
- **CEA:** Carcinoembryonic antigen (CEA; [22], [23]) is a protein that is in humans normally produced during fetal development. As the production of CEA

usually is stopped before birth, it is usually not present in the blood of healthy adults. Elevated levels are seen in the blood or tissues of heavy smokers; persons with pancreatic carcinoma, colorectal carcinoma, lung carcinoma, gastric carcinoma, or breast carcinoma, often have elevated CEA levels. When used as a tumor marker, CEA is mainly used to identify recurrences of cancer after surgical resections.

- **CYFRA:** Fragments of cytokeratin 19, a protein found in the cytoskeleton, are found in many places of the human body; especially in the lung and in malign lung tumors high concentrations of these fragments, which are also called CYFRA 21-1, are found. Due to elevated values in the presence of lung cancer CYFRA is often used for detecting and monitoring malign lung tumors. Elevated CYFRA values have already been reported for several different kinds of tumors, especially for example in stomach, colon, breast, and ovaries. The use of CYFRA 21-1 as a tumor marker has for example been discussed in [31].
- **PSA:** Prostate-specific antigen (PSA; [7], [45]) is a protein produced in the prostate gland; PSA blood tests are widely considered the most effective test currently available for the early detection of prostate cancer since PSA is often elevated in the presence of prostate cancer and in other prostate disorders. Still, the effectiveness of these tests has also been considered questionable since PSA is prone to both false positive and false negative indications: According to [45], 70 out of 100 men with elevated PSA values do not have prostate cancer, and 25 out of 100 men suffering from prostate cancer do not have significantly elevated PSA.

As already mentioned, information about cancer diagnoses is also available in the AKH database: If a patient is diagnosed with any kind of cancer, then this is also stored in the database. All cancer diagnoses are classified according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) system.

In this research work we concentrate on diagnoses regarding the following types of cancer: Malignant neoplasms in the respiratory system (RSC, cancer classes C30–C39 according to ICD-10), melanoma and malignant neoplasms on the skin (Mel, C43–C44), and breast cancer (BC, C50).

6.3 Modeling Approaches

In this section we describe the modeling methods applied for identifying estimation models for cancer diagnosis: On the one hand we apply hybrid modeling using machine learning algorithms (linear regression, neural networks, the k-nearest neighbor method, support vector machines) and evolutionary algorithms for parameter optimization and feature selection (as described in Section 6.3.5), on the other hand apply use genetic programming (as described in Section 6.3.6).

All these machine learning methods have been implemented using the HeuristicLab framework [47], a framework for prototyping and analyzing optimization techniques for which both generic concepts of evolutionary algorithms and many functions to evaluate and analyze them are available. HeuristicLab is developed by the Heuristic and Evolutionary Algorithm Laboratory[2] and can be downloaded from the HeuristicLab homepage[3]. HeuristicLab is licensed under the GNU General Public License[4].

6.3.1 Linear Modeling

Given a data collection including m input features storing the information about N samples, a linear model is defined by the vector of coefficients $\theta_{1...m}$. For calculating the vector of modeled values e using the given input values matrix $u_{1...m}$, these input values are multiplied with the corresponding coefficients and added: $e = u_{1...m} * \theta$. The coefficients vector can be computed by simply applying matrix division. For conducting the test series documented here we have used an implementation of the matrix division function: $\theta = InputValues \backslash TargetValues$. Additionally, a constant additive factor is also included into the model; i.e., a constant offset is added to the coefficients vector. Theoretical background of this approach can be found in [33].

6.3.2 kNN Classification

Unlike other data based modeling methods, k-nearest neighbor classification [16] works without creating any explicit models. During the training phase, the samples are simply collected; when it comes to classifying a new, unknown sample x_{new}, the sample-wise distance between x_{new} and all other training samples x_{train} is calculated and the classification is done on the basis of those k training samples (x_{NN}) showing the smallest distances from x_{new}.

In the context of classification, the numbers of instances (of the k nearest neighbors) are counted for each given class and the algorithm automatically predicts that class that is represented by the highest number of instances (included in x_{NN}). In the test series documented in this chapter we have applied weighting to kNN classification: The distance between x_{new} and x_{NN} is relevant for the classification statement, the weight of "nearer" samples is higher than that of samples that are "further away" from x_{new}. In this research work we have varied k between 1 and 10.

[2] http://heal.heuristiclab.com/

[3] http://dev.heuristiclab.com/

[4] http://www.gnu.org/licenses/gpl.txt

6.3.3 Artificial Neural Networks

For training artificial neural network (ANN) models, three-layer feed-forward neural networks with one linear output neuron were created applying backpropagation (using gradient descent optimization); theoretical background and details can for example be found in [37] (Chapter 11, "Neural Networks"). In the tests documented in this chapter the number of hidden (sigmoidal) nodes *hn* has been varied from 5 to 100; the learning rate as well as the momentum were also varied, the range of these parameters was set to [0.01 - 0.5]. We have applied ANN training algorithms that use internal validation sets, i.e., training algorithms use 30% of the given training data as validation data and eventually return those network structures that perform best on these internal validation samples.

6.3.4 Support Vector Machines

Support vector machines (SVMs) are a widely used approach in machine learning based on statistical learning theory [46]. The most important aspect of SVMs is that it is possible to give bounds on the generalization error of the models produced, and to select the corresponding best model from a set of models following the principle of structural risk minimization [46].

In this work we have used the LIBSVM implementation described in [12], which is used in the respective SVM interface implemented for HeuristicLab; here we have used Gaussian radial basis function kernels with varying values for the cost parameters c ($c \in [0,512]$) and the γ parameter of the SVM's kernel function ($\gamma \in [0,1]$).

6.3.5 Hybrid Modeling Using Machine Learning Algorithms and Evolutionary Algorithms for Parameter Optimization and Feature Selection

An essential step in data mining and machine learning is (especially when there are very many available features / variables) the selection of subsets of variables that are used for learning models. On the one hand, simpler models (i.e., models that use fewer variables and have simpler structures) are preferred over more complex ones following Occam's law of parsimony [8] that states that simpler theories are in general more favorable; on the other hand, simpler models are less likely to be prone to overfitting ([4], [29]).

So-called forward approaches iteratively add variables that are essentially important for improving the quality of the achievable models, while backward elimination methods initially use all variables and iteratively eliminate those that show the least statistical significance. Early variable selection algorithms were published several

decades ago (such as, e.g., [18]). Since then, numerous variable selection algorithms have been developed, many of them relying on the concept of mutual information ([11], [13], [14], [20], [21], [42], [44]). An overview of well-established variable selection methods can be found for example in [15], the use of variable selection methods in cancer classification in [35].

Unlike these variable selection methods, we here use an evolutionary algorithm that is able to simultaneously optimize variable selections and modeling parameters with respect to specific machine learning algorithms. The main advantages of this approach are on the one hand that variable selection is not necessary as a separate step in the data mining process, on the other hand variable selections and modeling parameter settings are automatically optimized for the modeling algorithm at hand. Parsimony pressure is realized by incorporating the size of sets of selected variables into the fitness function that is used for evaluating solution candidates.

Given a set of n features $F = \{f_1, f_2, \ldots, f_n\}$, our goal here is to find a subset $F' \subseteq F$ that is on the one hand as small as possible and on the other hand allows modeling methods to identify models that estimate given target values as well as possible. Additionally, each data based modeling method (except plain linear regression) has several parameters that have to be set before starting the modeling process.

The fitness of feature selection F' and training parameters with respect to the chosen modeling method is calculated in the following way: We use a machine learning algorithm m (with parameters p) for estimating predicted target values $est(F', m, p)$ and compare those to the original target values *orig*; the coefficient of determination (R^2) function is used for calculating the quality of the estimated values. Additionally, we also calculate the ratio of selected features $|F'|/|F|$. Finally, using a weighting factor α, we calculate the fitness of the set of features F' using m and p as

$$fitness(F', m, p) = \alpha * |F'|/|F| + (1 - \alpha) * (1 - R^2(est(F', m, p), orig)). \quad (6.1)$$

As an alternative to the coefficient of determination function we can also use a classification specific function that calculates the ratio of correctly classified samples, either in total or as the average of all classification accuracies of the given classes (as for example described in [50], Section 8.2): For all samples that are to be considered we know the original classifications *origCl*, and using (predefined or dynamically chosen) thresholds we get estimated classifications $estCl(F', m, p)$ for estimated target values $est(F', m, p)$. The total classification accuracy $ca_k(F', m, p)$ is calculated as

$$ca(F', m, p) = \frac{|\{j : estCl(F', m, p)[j] = origCl[j]\}|}{|estCl|} \quad (6.2)$$

Class-wise classification accuracies *cwca* are calculated as the average of all classification accuracies for each given class $c \in C$ separately:

$$ca(F', m, p)_c = \frac{|\{j : estCl(F', m, p)[j] = origCl[j] = c\}|}{|\{j : origCl[j] = c\}|} \quad (6.3)$$

$$cwca(F',m,p) = \frac{\sum_{c \in C} ca(F',m,p)_c}{|C|} \tag{6.4}$$

We can now define the classification specific fitness of feature selection F' using m and p as

$$fitness_{ca}(F',m,p) = \alpha * |F'|/|F| + (1-\alpha) * (1 - ca(F',m,p)) \tag{6.5}$$

or

$$fitness_{cwca}(F',m,p) = \alpha * |F'|/|F| + (1-\alpha) * (1 - cwca(F',m,p)). \tag{6.6}$$

In [5], for example, the use of evolutionary algorithms for feature selection optimization is discussed in detail in the context of gene selection in cancer classification; in [53] we have analyzed the sets of features identified as relevant in the modeling of tumor markers AFP and CA15-3.

We have now used evolutionary algorithms for finding optimal feature sets as well as optimal modeling parameters for models for tumor diagnosis; this approach is schematically shown in Figure 6.2. A solution candidate is here represented as $[s_{1,\ldots,n}\ p_{1,\ldots,q}]$ where s_i is a bit denoting whether feature F_i is selected or not and p_j is the value for parameter j of the chosen modeling method m. This rather simple definition of solution candidates enables the use of standard concepts for genetic operators for crossover and mutation of bit vectors and real valued vectors: We use uniform, single point, and 2-point crossover operators for binary vectors and bit flip mutation that flips each of the given bits with a given probability. Explanations of these operators can for example be found in [19] and [24].

In the test series described later in Section 6.5 we have used strict offspring selection [1] which means that individuals are accepted to become members of the next generation if they are evaluated better than both parents. Standard fitness evaluation as given in Equation 6.1 has been used during the execution of the evolutionary processes, and classification specific fitness evaluation as given in Equation 6.6 has been used for selecting the solution candidate eventually returned as the algorithm's result.

6.3.6 Genetic Programming

We have also applied a classification algorithm based on genetic programming (GP) [28] using a structure identification framework described in [4] and [50], in combination with an enhanced, hybrid selection scheme called offspring selection ([1], [2], [3]). In the left part of Figure 6.3 we show the overall GP workflow including offspring selection, in the right part the here used strict version of OS is depicted; we have used the GP implementation in HeuristicLab.

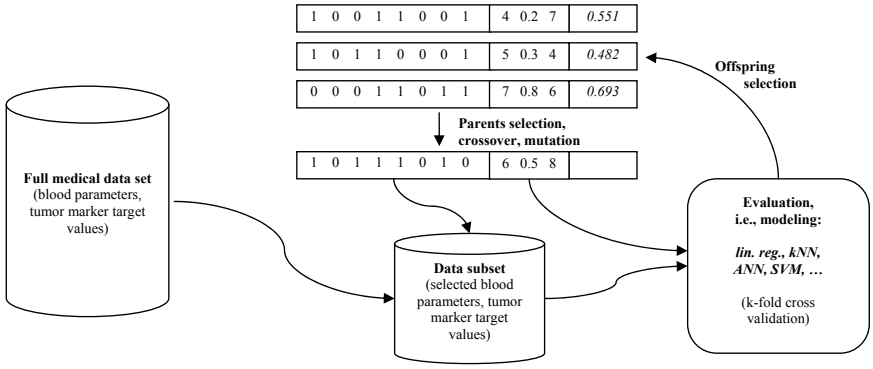

Fig. 6.2. A hybrid evolutionary algorithm for feature selection and parameter optimization in data based modeling

In addition to splitting the given data into training and test data, the GP based training algorithm implemented in HeuristicLab has been designed in such a way that a part of the given training data is not used for training models and serves as validation set; in the end, when it comes to returning classifiers, the algorithm returns those models that perform best on validation data. This approach has been chosen because it is assumed to help to cope with over-fitting; it is also applied in other GP based machine learning algorithms as for example described in [9].

We have used the following parameter settings for our GP test series: The mutation rate was set to 20%, gender specific parents selection [48] (combining random and roulette selection) was applied as well as strict offspring selection [1] (OS, with success ratio as well as comparison factor set to 1.0). The functions set described in [50] (including arithmetic as well as logical ones) was used for building composite function expressions.

The following parameter settings have been used in our GP test series:

- Single population approach; the population size was set to 700
- Mutation rate: 15%
- Varying maximum formula tree complexity
- Parents selection: Gender specific [48], random & roulette
- Offspring selection [1]: Strict offspring selection (success ratio as well as comparison factor set to 1.0)
- One-elitism
- Termination criteria:
  - Maximum number of generations: 1000; this criterion was not reached in the tests documented here, all executions were terminated via the
  - Maximum selection pressure [1]: 555

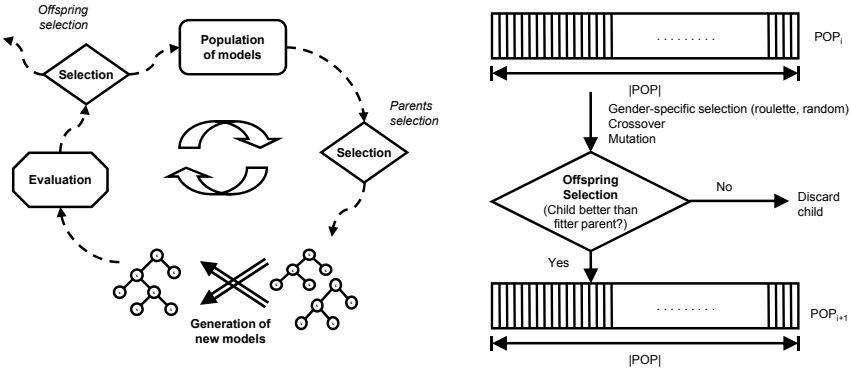

Fig. 6.3. The GP cycle [32] including offspring selection. Flowchart for embedding a simplified version of offspring selection into the GP based machine learning process.

- Function set: All functions (including arithmetic as well as logical ones) as described in [50]

In addition to splitting the given data into training and test data, the GP based training algorithm implemented in HeuristicLab has been implemented in such a way that a part of the given training data is not used for training models and serves as validation set; in the end, when it comes to returning classifiers, the algorithm returns those models that perform best on validation data. This approach has been chosen because it is assumed to help to cope with over-fitting; it is also applied in other GP based machine learning algorithms as for example described in [9].

6.4 Empirical Study: Identification of Models for Tumor Markers

In this section we summarize empirical results previously described in [51].

6.4.1 Data Preprocessing

Before analyzing the data and using them for training classifiers for tumor markers we have preprocessed the data available in the AKH data base:

- All variables have been linearly scaled to the interval [0;1]: For each variable v_i, the minimum value min_i is subtracted from all contained values and the

result divided by the difference between min_i and the maximum plausible value $maxplau_i$; all values greater than the given maximum plausible value are replaced by 1.0.
- All samples belonging to the same patient with not more than one day difference with respect to the measurement data have been merged. This has been done in order to decrease the number of missing values in the data matrix. In rare cases, more than one value might thus be available for a certain variable; in such a case, the first value is used.
- Additionally, all measurements have been sample-wise re-arranged and clustered according to the patients' IDs. This has been done in order to prevent data of certain patients being included in the training as well as in the test data.

Before modeling algorithms can be used for training classifiers, we have compiled separate data sets for each analyzed target tumor marker tm_i: First, all samples containing measured values for tm_i are extracted. Second, all variables are removed from the resulting data set that contain values in less than 80% of the remaining samples. Third, all samples are removed that still contain missing values. This procedure results in a specialized data set $dstm_i$ for each tumor marker tm_i. In Table 6.3 we summarize statistical information about all resulting data sets for the markers analyzed here[5]; the numbers of samples belonging to each of the defined classes are also given for each resulting data set.

6.4.2 Test Series and Results

All machine learning mentioned in Section 6.3 have been applied using several different parametrizations; for each modeling method we also give relevant parameter settings, namely the number of neighbors k for kNN learning ($k \in \{1,3,5,10\}$), the number of nodes n in the hidden layer of ANNs ($n \in \{10,25,50,100\}$), the γ value for SVMs ($\gamma \in \{0.001,0.01,0.05,0.1,0.5,1,8\}$), and the maximum tree size s (i.e., the number of nodes in formula trees) for GP ($s \in \{25,50,100,150\}$). Five-fold cross-validation [27] training / test series have been executed; this means that the available data are separated in five (approximately) equally sized, complementary subsets, and in each training / test cycle one data subset is chosen as used as test and the rest of the data as training samples. The classifiers learned using training data are evaluated on training as well as on test data; in the following Tables 6.4 – 6.9 we give statistics about the quality on the produced classifiers, namely the average classification accuracies (as the average of correct sample classifications) and their standard deviations on training as well as on test data.

[5] Please note that the number of total samples and the number of samples in class 0 for the PSA data set differ from the numbers stated in [51]; the here given numbers are the correct ones.

Table 6.3. Overview of the data sets compiled for selected tumor markers

Marker Name	*Input Variables*	*Total Samples*	*Distribution of Samples*			
			Class 0	*Class 1*	*Class 2*	*Class 3*
AFP	AGE, SEX, ALT, AST, BUN, CH37, GT37, HB, HKT, KREA, LD37, MCV, PLT, RBC, TBIL, WBC	2,755	2146 (77.9%)	454 (16.5%)	64 (2.32%)	91 (3.3%)
CA 125	AGE, SEX, ALT, AST, BUN, CRP, GT37, HB, HKT, HS, KREA, LD37,MCV, PLT, RBC, TBIL, WBC	1,053	532 (50.5%)	143 (13.6%)	84 (8.0%)	294 (27.9%)
CA 15-3	AGE, SEX, ALT, AST, BUN, CBAA,CEOA, CLYA, CMOA, CNEA, CRP,GT37, HB, HKT, HS, KREA, LD37, MCV, PLT, RBC, TBIL, WBC	4,918	3,159 (64.2%)	1,011 (20.6%)	353 (7.2%)	395 (8.0%)
CEA	AGE, SEX, ALT, AST, BUN, CBAA, CEOA, CLYA, CMOA, CNEA, CRP, GT37, HB, HKT, HS, KREA, LD37, MCV, PLT, RBC, TBIL, WBC	5,567	3,133 (56.3%)	1,443 (25.9%)	492 (8.8%)	499 (9.0%)
CYFRA	AGE, SEX, ALT, AST, BUN, CH37, CHOL, CRP, CYFS, GT37, HB, HKT, HS, KREA, MCV, PLT, RBC, TBIL, WBC	419	296 (70.6%)	37 (8.8%)	36 (8.6%)	50 (11.9%)
PSA	AGE, SEX, ALT, AST, BUN, CBAA, CEOA, CHOL, CLYA, CMOA, CNEA, CRP, GT37, HB, HKT, HS, KREA, LD37, MCV, PLT, RBC, TBIL, WBC	2,366	1,145 (48.4%)	779 (32.9%)	249 (10.5%)	193 (8.2%)

Table 6.4. Classification results for AFP

Modeling method		Classification accuracy ($\mu \pm \sigma$) Training	Test
LinReg		0.8356 ($\pm$ 0.009)	0.8221 ($\pm$ 0.018)
kNN	1	1.0000 ($\pm$ 0.000)	0.7659 ($\pm$ 0.026)
	3	1.0000 ($\pm$ 0.000)	0.7790 ($\pm$ 0.031)
	5	1.0000 ($\pm$ 0.000)	0.7974 ($\pm$ 0.037)
	10	1.0000 ($\pm$ 0.000)	0.8061 ($\pm$ 0.041)
ANN	10	0.8625 ($\pm$ 0.014)	0.8215 ($\pm$ 0.045)
	25	0.8608 ($\pm$ 0.029)	0.8239 ($\pm$ 0.049)
	50	0.8564 ($\pm$ 0.017)	**0.8548** ($\pm$ 0.020)
	100	0.8642 ($\pm$ 0.014)	0.8217 ($\pm$ 0.049)
SVM	0.01	0.7846 ($\pm$ 0.012)	0.7826 ($\pm$ 0.067)
	0.05	0.8301 ($\pm$ 0.015)	0.8137 ($\pm$ 0.048)
	0.1	0.8490 ($\pm$ 0.009)	0.8107 ($\pm$ 0.054)
	0.5	0.9057 ($\pm$ 0.011)	0.8075 ($\pm$ 0.043)
	1	0.9426 ($\pm$ 0.009)	0.7998 ($\pm$ 0.039)
	8	1.0000 ($\pm$ 0.000)	0.7773 ($\pm$ 0.045)
GP	25	0.771 ($\pm$ 0.014)	0.7735 ($\pm$ 0.028)
	50	0.7884 ($\pm$ 0.014)	0.7681 ($\pm$ 0.014)
	100	0.7985 ($\pm$ 0.026)	0.7739 ($\pm$ 0.037)

Table 6.5. Classification results for CA125

Modeling method		Classification accuracy ($\mu \pm \sigma$) Training	Test
LinReg		0.7243 ($\pm$ 0.022)	0.5913 ($\pm$ 0.071)
kNN	1	1.0000 ($\pm$ 0.000)	0.5611 ($\pm$ 0.070)
	3	1.0000 ($\pm$ 0.000)	0.5449 ($\pm$ 0.074)
	5	1.0000 ($\pm$ 0.000)	0.5560 ($\pm$ 0.062)
	10	1.0000 ($\pm$ 0.000)	0.5680 ($\pm$ 0.084)
ANN	10	0.7941 ($\pm$ 0.049)	0.5342 ($\pm$ 0.073)
	25	0.7707 ($\pm$ 0.184)	0.6275 ($\pm$ 0.071)
	50	0.7601 ($\pm$ 0.034)	0.5851 ($\pm$ 0.059)
	100	0.7661 ($\pm$ 0.048)	0.5993 ($\pm$ 0.047)
SVM	0.01	0.7140 ($\pm$ 0.024)	0.5594 ($\pm$ 0.081)
	0.05	0.7663 ($\pm$ 0.023)	0.5503 ($\pm$ 0.080)
	0.1	0.8054 ($\pm$ 0.015)	0.5239 ($\pm$ 0.080)
	0.5	0.9180 ($\pm$ 0.017)	0.5357 ($\pm$ 0.068)
	1	0.9905 ($\pm$ 0.003)	0.5607 ($\pm$ 0.063)
	8	1.0000 ($\pm$ 0.000)	0.5548 ($\pm$ 0.103)
GP	**25**	0.7194 ($\pm$ 0.017)	**0.6810** ($\pm$ 0.081)
	50	0.7474 ($\pm$ 0.022)	0.6677 ($\pm$ 0.067)
	100	0.7628 ($\pm$ 0.014)	0.6344 ($\pm$ 0.048)

Table 6.6. Classification results for CA 15-3

Modeling method		Classification accuracy ($\mu \pm \sigma$) Training	Test
LinReg		0.7533 ($\pm$ 0.017)	0.7069 ($\pm$ 0.020)
kNN	1	1.0000 ($\pm$ 0.000)	0.6495 ($\pm$ 0.030)
	3	1.0000 ($\pm$ 0.000)	0.6767 ($\pm$ 0.029)
	5	1.0000 ($\pm$ 0.000)	0.6905 ($\pm$ 0.032)
	10	1.0000 ($\pm$ 0.000)	0.7028 ($\pm$ 0.029)
ANN	10	0.7920 ($\pm$ 0.008)	0.7127 ($\pm$ 0.022)
	25	0.8000 ($\pm$ 0.027)	0.7007 ($\pm$ 0.028)
	50	0.7878 ($\pm$ 0.008)	0.6997 ($\pm$ 0.012)
	100	0.7866 ($\pm$ 0.040)	0.6974 ($\pm$ 0.009)
SVM	0.01	0.7252 ($\pm$ 0.017)	0.7077 ($\pm$ 0.039)
	0.05	0.7700 ($\pm$ 0.010)	0.7236 ($\pm$ 0.034)
	0.1	0.8089 ($\pm$ 0.006)	0.7181 ($\pm$ 0.034)
	0.5	0.9405 ($\pm$ 0.001)	0.6843 ($\pm$ 0.026)
	1	0.9874 ($\pm$ 0.001)	0.7135 ($\pm$ 0.078)
	8	1.0000 ($\pm$ 0.000)	0.6433 ($\pm$ 0.031)
GP	25	0.7348 ($\pm$ 0.012)	0.7167 ($\pm$ 0.036)
	50	0.7327 ($\pm$ 0.009)	0.7163 ($\pm$ 0.039)
	100	0.7504 ($\pm$ 0.011)	**0.7302** ($\pm$ 0.030)

Table 6.7. Classification results for CEA

Modeling method		Classification accuracy ($\mu \pm \sigma$) Training	Test
LinReg		0.6847 ($\pm$ 0.020)	0.6566 ($\pm$ 0.053)
kNN	1	1.0000 ($\pm$ 0.000)	0.5877 ($\pm$ 0.009)
	3	1.0000 ($\pm$ 0.000)	0.5979 ($\pm$ 0.017)
	5	1.0000 ($\pm$ 0.000)	0.6006 ($\pm$ 0.022)
	10	1.0000 ($\pm$ 0.000)	0.6115 ($\pm$ 0.034)
ANN	10	0.7526 ($\pm$ 0.025)	0.6556 ($\pm$ 0.030)
	25	0.7681 ($\pm$ 0.026)	0.6393 ($\pm$ 0.039)
	50	0.7486 ($\pm$ 0.010)	0.6555 ($\pm$ 0.044)
	100	0.7459 ($\pm$ 0.005)	0.6625 ($\pm$ 0.039)
SVM	0.001	0.7277 ($\pm$ 0.012)	0.6461 ($\pm$ 0.054)
	0.01	0.6758 ($\pm$ 0.014)	0.6470 ($\pm$ 0.061)
	0.05	0.7432 ($\pm$ 0.008)	0.6417 ($\pm$ 0.048)
	0.1	0.7904 ($\pm$ 0.003)	0.6373 ($\pm$ 0.043)
	0.5	0.9321 ($\pm$ 0.011)	0.6030 ($\pm$ 0.034)
	1	0.9788 ($\pm$ 0.008)	0.5813 ($\pm$ 0.026)
	8	1.0000 ($\pm$ 0.000)	0.5508 ($\pm$ 0.014)
GP	25	0.6791 ($\pm$ 0.034)	0.6532 ($\pm$ 0.007)
	50	0.6854 ($\pm$ 0.009)	0.6686 ($\pm$ 0.014)
	100	0.6874 ($\pm$ 0.005)	0.6772 ($\pm$ 0.008)
	150	0.6822 ($\pm$ 0.005)	**0.6828** ($\pm$ 0.011)

Table 6.8. Classification results for CYFRA

Modeling method		Classification accuracy ($\mu \pm \sigma$) Training	Test
LinReg		0.7957 ($\pm$ 0.020)	0.7061 ($\pm$ 0.076)
kNN	1	1.0000 ($\pm$ 0.000)	0.6372 ($\pm$ 0.048)
	3	1.0000 ($\pm$ 0.000)	0.6445 ($\pm$ 0.111)
	5	1.0000 ($\pm$ 0.000)	0.6586 ($\pm$ 0.119)
	10	1.0000 ($\pm$ 0.000)	0.6965 ($\pm$ 0.106)
ANN	10	0.7590 ($\pm$ 0.014)	0.7139 ($\pm$ 0.073)
	25	0.7639 ($\pm$ 0.037)	0.7134 ($\pm$ 0.080)
	50	0.8304 ($\pm$ 0.045)	0.7087 ($\pm$ 0.079)
	100	0.7768 ($\pm$ 0.035)	**0.7303** ($\pm$ 0.048)
	200	0.7969 ($\pm$ 0.051)	0.7160 ($\pm$ 0.055)
SVM	0.001	0.7063 ($\pm$ 0.039)	0.6727 ($\pm$ 0.058)
	0.01	0.7195 ($\pm$ 0.025)	0.7041 ($\pm$ 0.071)
	0.05	0.8269 ($\pm$ 0.015)	0.7206 ($\pm$ 0.073)
	0.1	0.8585 ($\pm$ 0.069)	0.7229 ($\pm$ 0.056)
	0.5	0.9988 ($\pm$ 0.002)	0.6985 ($\pm$ 0.059)
	1	1.0000 ($\pm$ 0.000)	0.6634 ($\pm$ 0.092)
	8	1.0000 ($\pm$ 0.000)	0.7088 ($\pm$ 0.086)
GP	25	0.7565 ($\pm$ 0.033)	0.6763 ($\pm$ 0.069)
	50	0.7708 ($\pm$ 0.025)	0.6520 ($\pm$ 0.056)
	100	0.7865 ($\pm$ 0.039)	0.6563 ($\pm$ 0.050)

Table 6.9. Classification results for PSA

Modeling method		Classification accuracy ($\mu \pm \sigma$) Training	Test
LinReg		0.6403 ($\pm$ 0.014)	0.5858 ($\pm$ 0.026)
kNN	1	1.0000 ($\pm$ 0.000)	0.5365 ($\pm$ 0.037)
	3	1.0000 ($\pm$ 0.000)	0.5420 ($\pm$ 0.029)
	5	1.0000 ($\pm$ 0.000)	0.5517 ($\pm$ 0.018)
	10	1.0000 ($\pm$ 0.000)	0.5622 ($\pm$ 0.021)
ANN	10	0.6507 ($\pm$ 0.043)	0.5840 ($\pm$ 0.018)
	25	0.6327 ($\pm$ 0.014)	0.5828 ($\pm$ 0.014)
	50	0.6456 ($\pm$ 0.016)	0.5718 ($\pm$ 0.013)
	100	0.6462 ($\pm$ 0.017)	0.5786 ($\pm$ 0.018)
	200	0.6661 ($\pm$ 0.016)	0.5808 ($\pm$ 0.016)
SVM	0.001	0.5882 ($\pm$ 0.027)	0.5536 ($\pm$ 0.052)
	0.01	0.6403 ($\pm$ 0.016)	0.5913 ($\pm$ 0.022)
	0.05	0.6813 ($\pm$ 0.011)	0.5672 ($\pm$ 0.018)
	0.1	0.7331 ($\pm$ 0.005)	0.5740 ($\pm$ 0.015)
	0.5	0.9286 ($\pm$ 0.012)	0.5600 ($\pm$ 0.026)
	1	0.9903 ($\pm$ 0.003)	0.5499 ($\pm$ 0.021)
	8	0.9996 ($\pm$ 0.001)	0.5351 ($\pm$ 0.024)
GP	25	0.6857 ($\pm$ 0.090)	0.6527 ($\pm$ 0.077)
	50	0.7176 ($\pm$ 0.080)	**0.6746** ($\pm$ 0.080)
	100	0.7071 ($\pm$ 0.029)	0.6730 ($\pm$ 0.039)

6.5 Empirical Study: Identification of Models for Tumor Diagnoses

In this section we summarize empirical results previously described in [52].

6.5.1 Data Preprocessing

Before analyzing the data and using them for training data-based tumor diagnosis estimators we have preprocessed the available data:

- All variables have been linearly scaled to the interval [0;1]: For each variable v_i, the minimum value min_i is subtracted from all contained values and the result divided by the difference between min_i and the maximum plausible value $maxplau_i$; all values greater than the given maximum plausible value are replaced by 1.0.
- All samples belonging to the same patient with not more than one day difference with respect to the measurement data have been merged. This has been done in order to decrease the number of missing values in the data matrix. In rare cases, more than one value might thus be available for a certain variable; in such a case, the first value is used.
- Additionally, all measurements have been sample-wise re-arranged and clustered according to the patients' IDs. This has been done in order to prevent data of certain patients being included in the training as well as in the test data.

Before starting the modeling algorithms for training classifiers we had to compile separate data sets for each analyzed target tumor t_i: First, blood parameter measurements were joined with diagnosis results; only measurements and diagnoses with a time delta less than a month were considered. Second, all samples containing measured values for t_i are extracted. Third, all samples are removed that contain less than 15 valid values. Finally, variables with less than 10% valid values are removed from the data base.

This procedure results in a specialized data set dst_i for each tumor marker t_i. In Table 6.10 we summarize statistical information about all resulting data sets for the markers analyzed here; the numbers of samples belonging to each of the defined classes are also given for each resulting data set.

6.5.2 Test Series and Results

Five-fold cross-validation [27] training / test series have been executed; this means that the available data are separated in five (approximately) equally sized,

complementary subsets, and in each training / test cycle one data subset is chosen is used as test and the rest of the data as training samples.

In this section we document test accuracies (μ, σ) for the investigated cancer types; we here summarize test results for modeling cancer diagnoses using tumor markers (TMs) as well as for modeling without using tumor markers. Linear modeling, kNN modeling, ANNs, and SVMs have been applied for identifying estimation models for the selected tumor types, genetic algorithms with strict OS have been applied for optimizing variable selections and modeling parameters; standard fitness calculation as given in Equation 6.1 has been used by the evolutionary process, the classification specific one as given in Equation 6.6 has been used for selecting the eventually returned model. The probability of selecting a variable initially was set to 30%. Additionally, we have also applied simple linear regression using all available variables. Finally, genetic programming with strict offspring selection (OSGP) has also been applied.

Table 6.10. Overview of the data sets compiled for selected cancer types

Cancer Type	*Input Variables*	*Total Samples*	*Samples in*		Missing Values
			Class 0	*Class 1*	
Breast Cancer	AGE, SEX, AFP, ALT, AST, BSG1, BUN, C125, C153, C199, C724, CBAA, CEA,	706	324 (45.9%)	382 (54.1%)	46.67%
Melanoma	CEOA, CH37, CHOL, CLYA, CMOA, CNEA, CRP, CYFS,	905	485 (53.6%)	420 (46.4%)	47.79%
Respiratory System Cancer	FE, FER, FPSA, GT37, HB, HDL, HKT, HS, KREA, LD37, MCV, NSE, PLT, PSA, PSAQ, RBC, S100, SCC, TBIL, TF, TPS, WBC	2,363	1,367 (57.9%)	996 (42.1%)	44.76%

In all test series the maximum selection pressure [1] was set to 100, i.e., the algorithms were terminated as soon as the selection pressure reached 100. The population size for genetic algorithms optimizing variable selections and modeling parameters was set to 10, for GP the population size was set to 700. In all modeling cases except kNN modeling regression models have been trained, the threshold for classification decisions was in all cases set to 0.5 (since the absence of the specific tumor is represented by 0.0 in the data and its presence by 1.0).

Table 6.11. Modeling results for breast cancer diagnosis

Modeling Method	Using TMs Test accuracies		Not using TMs Test accuracies	
	μ	σ	μ	σ
LR, full features set	79.32%	*1.06*	70.63%	*1.28*
OSGA + LR, $\alpha = 0.0$	81.78%	*0.21*	73.13%	*0.36*
OSGA + LR, $\alpha = 0.1$	81.49%	*1.18*	72.66%	*0.14*
OSGA + LR, $\alpha = 0.2$	81.44%	*0.37*	71.40%	*0.57*
OSGA + kNN, $\alpha = 0.0$	79.21%	*0.78*	74.22%	*2.98*
OSGA + kNN, $\alpha = 0.1$	78.99%	*0.57*	75.55%	*0.87*
OSGA + kNN, $\alpha = 0.2$	78.33%	*1.04*	74.50%	*0.20*
OSGA + ANN, $\alpha = 0.0$	81.41%	*1.14*	75.60%	*2.47*
OSGA + ANN, $\alpha = 0.1$	80.19%	*1.68*	72.38%	*6.08*
OSGA + ANN, $\alpha = 0.2$	79.37%	*1.17*	70.54%	*6.10*
OSGA + SVM, $\alpha = 0.0$	81.23%	*1.10*	73.90%	*2.36*
OSGA + SVM, $\alpha = 0.1$	80.46%	*1.80*	72.19%	*0.94*
OSGA + SVM, $\alpha = 0.2$	77.43%	*3.55*	71.89%	*0.70*
OSGP, $ms = 50$	79.72%	*1.80*	75.32%	*0.45*
OSGP, $ms = 100$	75.50%	*4.95*	71.63%	*2.75*
OSGP, $ms = 150$	79.20%	*6.60*	75.75%	*2.16*

Table 6.12. Modeling results for melanoma diagnosis

Modeling Method	Using TMs Test accuracies		Not using TMs Test accuracies	
	μ	σ	μ	σ
LR, full features set	73.81%	*3.39*	71.09%	*4.14*
OSGA + LR, $\alpha = 0.0$	72.45%	*4.69*	72.36%	*2.30*
OSGA + LR, $\alpha = 0.1$	74.73%	*2.35*	72.09%	*4.01*
OSGA + LR, $\alpha = 0.2$	73.85%	*2.54*	72.70%	*2.02*
OSGA + kNN, $\alpha = 0.0$	68.77%	*2.38*	71.00%	*1.97*
OSGA + kNN, $\alpha = 0.1$	71.33%	*0.27*	70.21%	*3.41*
OSGA + kNN, $\alpha = 0.2$	67.33%	*0.31*	69.65%	*3.14*
OSGA + ANN, $\alpha = 0.0$	74.78%	*1.63*	69.17%	*2.97*
OSGA + ANN, $\alpha = 0.1$	73.81%	*2.23*	71.82%	*0.61*
OSGA + ANN, $\alpha = 0.2$	74.12%	*1.03*	71.40%	*0.49*
OSGA + SVM, $\alpha = 0.0$	69.72%	*7.57*	68.87%	*4.78*
OSGA + SVM, $\alpha = 0.1$	71.75%	*4.88*	68.22%	*1.88*
OSGA + SVM, $\alpha = 0.2$	61.48%	*3.99*	63.20%	*2.09*
OSGP, $ms = 50$	71.24%	*9.54*	74.89%	*3.66*
OSGP, $ms = 100$	69.91%	*5.20*	65.16%	*13.06*
OSGP, $ms = 150$	71.79%	*4.31*	70.13%	*3.60*

Table 6.13. Modeling results for respiratory system cancer diagnosis

Modeling Method	Using TMs Test accuracies		Not using TMs Test accuracies	
	μ	σ	μ	σ
LR, full features set	91.32%	*0.37*	85.97%	*0.27*
OSGA + LR, $\alpha = 0.0$	91.57%	*0.46*	86.41%	*0.36*
OSGA + LR, $\alpha = 0.1$	91.16%	*1.18*	85.80%	*0.45*
OSGA + LR, $\alpha = 0.2$	89.45%	*0.37*	85.02%	*0.15*
OSGA + kNN, $\alpha = 0.0$	90.98%	*0.84*	87.09%	*0.46*
OSGA + kNN, $\alpha = 0.1$	90.01%	*2.63*	87.01%	*0.83*
OSGA + kNN, $\alpha = 0.2$	90.16%	*0.74*	86.92%	*0.81*
OSGA + ANN, $\alpha = 0.0$	90.28%	*1.63*	85.97%	*4.07*
OSGA + ANN, $\alpha = 0.1$	90.99%	*1.97*	85.82%	*4.52*
OSGA + ANN, $\alpha = 0.2$	88.64%	*1.87*	87.24%	*1.91*
OSGA + SVM, $\alpha = 0.0$	89.03%	*1.38*	83.12%	*3.79*
OSGA + SVM, $\alpha = 0.1$	89.91%	*1.58*	86.25%	*0.79*
OSGA + SVM, $\alpha = 0.2$	88.33%	*1.94*	84.66%	*2.06*
OSGP, $ms = 50$	89.58%	*2.75*	85.98%	*5.74*
OSGP, $ms = 100$	90.44%	*3.02*	86.54%	*6.02*
OSGP, $ms = 150$	89.58%	*3.75*	87.97%	*5.57*

In Table 6.14 we summarize the effort of the modeling approaches applied in this research work: For the combination of GAs and machine learning methods we document the number of modeling executions, and for GP we give the number of evaluated solutions (i.e., models).

For the combination of genetic algorithms with linear regression, kNN modeling, ANNs, and SVMs (with varying variable ratio (*vr*) weighting factors) as well as GP with varying maximum tree sizes *ms* we give the sizes of selected variable sets, and (where applicable) also *k*, *hn*, *c*, and γ. Obviously there are different variations in the parameters identified as optimal by the evolutionary process: The numbers of variables used as well as the neural networks' hidden nodes vary to a relatively small extent, e.g., whereas especially the SVMs' parameters (especially the *c* factors) vary very strongly.

Table 6.14. Effort in terms of executed modeling runs and evaluated model structures

Modeling Method	Modeling executions					
	$vrfw = 0.0$		$vrfw = 0.1$		$vrfw = 0.2$	
	μ	σ	μ	σ	μ	σ
LR	3260.4	*717.8*	2339.6	*222.6*	2465.2	*459.2*
kNN	2955.3	*791.8*	3046.0	*362.4*	3791.3	*775.9*
ANN	3734.0	*855.9*	3305.0	*582.6*	3297.0	*475.9*
SVM	2950.0	*794.8*	2846.0	*391.4*	3496.7	*859.8*

Modeling Method	Evaluated solutions (models)		
	$ms = 50$	$ms = 100$	$ms = 150$
	μ, σ	μ, σ	μ, σ
OSGP	1483865.0, *674026.2*	1999913.3, *198289.1*	2238496.7, *410123.6*

Table 6.15. Optimized parameters for linear regression

Problem Instance, *vr* weighting		Variables used μ	Variables used σ
BC, TM	$\alpha = 0.0$	16.6	*2.10*
	$\alpha = 0.1$	11.8	*1.50*
	$\alpha = 0.2$	6.4	*0.60*
BC, no TM	$\alpha = 0.0$	9.6	*1.15*
	$\alpha = 0.1$	8.8	*0.58*
	$\alpha = 0.2$	6.4	*1.20*
Mel, TM	$\alpha = 0.0$	16.6	*0.55*
	$\alpha = 0.1$	12.2	*0.84*
	$\alpha = 0.2$	9.2	*4.09*
Mel, no TM	$\alpha = 0.0$	10.8	*1.79*
	$\alpha = 0.1$	8.8	*2.28*
	$\alpha = 0.2$	8.2	*1.92*
RSC, TM	$\alpha = 0.0$	17.2	*2.95*
	$\alpha = 0.1$	13.4	*2.51*
	$\alpha = 0.2$	9.0	*2.55*
RSC, no TM	$\alpha = 0.0$	16.0	*4.64*
	$\alpha = 0.1$	9.6	*0.89*
	$\alpha = 0.2$	8.6	*3.21*

Table 6.16. Optimized parameters for kNN modeling

Problem Instance, *vr* weighting		Variables used μ	Variables used σ	k μ	k σ
BC, TM	$\alpha = 0.0$	18.2	*2.20*	9.8	*2.10*
	$\alpha = 0.1$	14.0	*3.60*	12.6	*4.60*
	$\alpha = 0.2$	11.0	*1.80*	11.2	*3.00*
BC, no TM	$\alpha = 0.0$	14.4	*1.67*	11.2	*1.64*
	$\alpha = 0.1$	14.0	*2.45*	13.8	*3.11*
	$\alpha = 0.2$	11.8	*0.84*	18.8	*1.10*
Mel, TM	$\alpha = 0.0$	15.6	*1.82*	17.8	*2.86*
	$\alpha = 0.1$	16.4	*1.34*	14.4	*5.90*
	$\alpha = 0.2$	13.6	*1.67*	19.4	*1.34*
Mel, no TM	$\alpha = 0.0$	15.0	*1.58*	14.2	*1.10*
	$\alpha = 0.1$	10.4	*1.52*	18.2	*1.64*
	$\alpha = 0.2$	9.6	*1.14*	16.8	*2.05*
RSC, TM	$\alpha = 0.0$	14.6	*1.67*	20.0	*0.00*
	$\alpha = 0.1$	13.6	*1.67*	16.8	*6.06*
	$\alpha = 0.2$	10.4	*1.52*	12.8	*3.90*
RSC, no TM	$\alpha = 0.0$	15.6	*2.79*	15.2	*1.64*
	$\alpha = 0.1$	12.2	*1.10*	10.6	*1.82*
	$\alpha = 0.2$	10.2	*2.95*	13.2	*2.95*

Table 6.17. Optimized parameters for ANNs

Problem Instance, *vr* weighting		Variables used μ	Variables used σ	hn μ	hn σ
BC, TM	$\alpha = 0.0$	17.0	*1.20*	75.6	*20.80*
	$\alpha = 0.1$	14.8	*3.40*	51.0	*5.80*
	$\alpha = 0.2$	11.0	*0.80*	35.8	*13.00*
BC, no TM	$\alpha = 0.0$	12.6	*1.41*	82.4	*23.46*
	$\alpha = 0.1$	12.2	*0.89*	70.8	*26.40*
	$\alpha = 0.2$	11.2	*1.10*	68.2	*14.58*
Mel, TM	$\alpha = 0.0$	19.6	*2.19*	56.8	*13.31*
	$\alpha = 0.1$	12.8	*2.28*	61.0	*2.55*
	$\alpha = 0.2$	15.6	*5.18*	51.2	*6.98*
Mel, no TM	$\alpha = 0.0$	15.4	*2.51*	68.6	*14.24*
	$\alpha = 0.1$	8.2	*1.64*	59.8	*3.83*
	$\alpha = 0.2$	8.0	*1.00*	58.6	*5.81*
RSC, TM	$\alpha = 0.0$	13.4	*3.44*	64.6	*10.97*
	$\alpha = 0.1$	11.2	*2.28*	68.2	*6.69*
	$\alpha = 0.2$	8.2	*1.64*	60.2	*13.92*
RSC, no TM	$\alpha = 0.0$	13.2	*2.28*	71.2	*12.38*
	$\alpha = 0.1$	12.2	*2.05*	70.6	*12.99*
	$\alpha = 0.2$	11.6	*2.19*	64.4	*14.24*

Table 6.18. Optimized parameters for SVMs

Problem Instance, *vr* weighting		Variables used		C		γ	
		μ	σ	μ	σ	μ	σ
BC, TM	$\alpha = 0.0$	21.6	3.50	101.50	92.30	0.05	0.06
	$\alpha = 0.1$	18.8	3.50	12.44	13.85	0.09	0.01
	$\alpha = 0.2$	16.0	2.00	64.79	67.59	0.04	0.01
BC, no TM	$\alpha = 0.0$	15.6	1.83	47.16	12.63	0.05	0.05
	$\alpha = 0.1$	15.4	1.10	22.50	25.88	0.07	0.04
	$\alpha = 0.2$	13.0	2.65	8.14	10.09	0.07	0.04
Mel, TM	$\alpha = 0.0$	13.0	4.53	166.23	236.61	0.27	0.25
	$\alpha = 0.1$	10.8	3.42	204.74	210.43	0.18	0.19
	$\alpha = 0.2$	4.2	2.95	123.08	44.14	0.26	0.20
Mel, no TM	$\alpha = 0.0$	21.4	6.95	116.21	196.73	0.41	0.30
	$\alpha = 0.1$	19.8	1.64	492.73	8.10	0.48	0.41
	$\alpha = 0.2$	14.4	3.29	310.17	208.60	0.36	0.35
RSC, TM	$\alpha = 0.0$	21.2	8.50	183.54	95.38	0.27	0.26
	$\alpha = 0.1$	14.6	1.14	74.56	67.98	0.09	0.10
	$\alpha = 0.2$	11.2	3.83	37.55	68.31	0.45	0.35
RSC, no TM	$\alpha = 0.0$	13.4	4.10	23.14	31.91	0.35	0.25
	$\alpha = 0.1$	12.4	3.21	144.73	96.79	0.19	0.08
	$\alpha = 0.2$	12.4	3.21	376.66	206.84	0.09	0.10

Table 6.19. Number of variables used by models returned by OSGP

Problem Instance, maximum tree size *ms*		Variables used by returned model	
		μ	σ
BC, TM	$ms = 50$	9.0	*2.74*
	$ms = 100$	9.6	*1.34*
	$ms = 150$	17.8	*0.45*
BC, no TM	$ms = 50$	10.5	*0.71*
	$ms = 100$	10.0	*1.41*
	$ms = 150$	11.5	*0.71*
Mel, TM	$ms = 50$	10.2	*2.05*
	$ms = 100$	10.0	*2.35*
	$ms = 150$	12.0	*2.00*
Mel, no TM	$ms = 50$	8.0	*1.58*
	$ms = 100$	8.8	*0.84*
	$ms = 150$	11.4	*3.36*
RSC, TM	$ms = 50$	7.8	*2.05*
	$ms = 100$	12.0	*2.35*
	$ms = 150$	12.0	*1.22*
RSC, no TM	$ms = 50$	9.4	*3.91*
	$ms = 100$	12.2	*2.17*
	$ms = 150$	13.6	*3.13*

6.6 Conclusion

We have described the data based identification of mathematical models for the tumor markers AFP, CA-125, CA15-3, CEA, CYFRA, and PSA as well as selected tumors, namely breast cancer, melanoma, and respiratory system cancer. Data collected at the General Hospital Linz, Austria have been used for creating models that predict tumor marker values and tumor diagnoses; several different techniques of applied statistics and computational intelligence have been applied, namely linear regression, kNN learning, artificial neural networks, support vector machines (all optimized using evolutionary algorithms), and genetic programming.

On the one hand, it seems that none of the methods used here produced the best results for all modeling tasks; in two cases (AFP and CYFRA) ANNs produced models that perform best on test data, in all remaining four cases (CA-125, CA15-3,

CEA, and PSA) extended genetic programming has produced best results. Additionally, we here see that in those modeling cases, for which GP found best results, these best results (produced by GP) are significantly better than those produced by other methods; for the other two modeling tasks, results found by linear regression were almost quite as good as the best ones (trained used ANNs). I.e., for those medical modeling tasks described here, GP performs best among those techniques that are able to identify nonlinearities (ANNs, SVMs, GP). Furthermore, we also see that GP results show less overfitting than those produced using other methods.

On the other hand, the investigated diagnoses of breast cancer, melanoma, and respiratory system cancer can be estimated correctly in up to 81%, 74%, and 91% of the analyzed test cases, respectively; without tumor markers up to 75%, 74%, and 88% of the test samples are correctly estimated, respectively. Linear modeling performs well in all modeling tasks, feature selection using genetic algorithms and nonlinear modeling yield even better results for all analyzed modeling tasks. No modeling method performs best for all diagnosis prediction tasks.

Acknowledgements. The work described in this chapter was done within the Josef Ressel-Centre *Heureka!* for Heuristic Optimization sponsored by the Austrian Research Promotion Agency (FFG).

References

1. Affenzeller, M., Wagner, S.: SASEGASA: A new generic parallel evolutionary algorithm for achieving highest quality results. Journal of Heuristics - Special Issue on New Advances on Parallel Meta-Heuristics for Complex Problems 10, 239–263 (2004)
2. Affenzeller, M., Wagner, S.: Offspring selection: A new self-adaptive selection scheme for genetic algorithms. In: Ribeiro, B., Albrecht, R.F., Dobnikar, A., Pearson, D.W., Steele, N.C. (eds.) Adaptive and Natural Computing Algorithms, Springer Computer Science, pp. 218–221. Springer (2005)
3. Affenzeller, M., Wagner, S., Winkler, S.: Goal-oriented preservation of essential genetic information by offspring selection. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), vol. 2, pp. 1595–1596. Association for Computing Machinery, ACM (2005)
4. Affenzeller, M., Winkler, S., Wagner, S., Beham, A.: Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications. Chapman & Hall / CRC (2009)
5. Alba, E., Garca-Nieto, J., Jourdan, L., Talbi, E.G.: Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In: IEEE Congress on Evolutionary Computation 2007, pp. 284–290 (2007)
6. Alberts, B.: Leukocyte functions and percentage breakdown. In: Molecular Biology of the Cell. NCBI Bookshelf (2005)
7. Andriole, G.L., Crawford, E.D., Grubband, R.L., Buys, S.S., Chia, D., Church, T.R., et al.: Mortality results from a randomized prostate-cancer screening trial. New England Journal of Medicine 360(13), 1310–1319 (2009)
8. Ariew, R.: Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony. University of Illinois, Champaign-Urbana (1976)

9. Banzhaf, W., Lasarczyk, C.: Genetic programming of an algorithmic chemistry. In: O'Reilly, U., Yu, T., Riolo, R., Worzel, B. (eds.) Genetic Programming Theory and Practice II, pp. 175–190. Ann Arbor (2004)
10. Bitterlich, N., Schneider, J.: Cut-off-independent tumour marker evaluation using ROC approximation. Anticancer Research 27, 4305–4310 (2007)
11. Brown, G.: A new perspective for information theoretic feature selection. In: International Conference on Artificial Intelligence and Statistics, pp. 49–56 (2009)
12. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
13. Cheng, H., Qin, Z., Feng, C., Wang, Y., Li, F.: Conditional mutual information-based feature selection analyzing for synergy and redundancy. Electronics and Telecommunications Research Institute (ETRI) Journal 33(2) (2011)
14. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience, New York (1991)
15. Duch, W.: Feature Extraction: Foundations and Applications. Springer (2006)
16. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley Interscience (2000)
17. Duffy, M.J., Crown, J.: A personalized approach to cancer treatment: how biomarkers can help. Clinical Chemistry 54(11), 1770–1779 (2008)
18. Efroymson, M.A.: Multiple regression analysis. Mathematical Methods for Digital Computers. Wiley (1960)
19. Eiben, A., Smith, J.: Introduction to Evolutionary Computation. Natural Computing Series. Springer, Heidelberg (2003)
20. El Akadi, A., El Ouardighi, A., Aboutajdine, D.: A powerful feature selection approach based on mutual information. International Journal of Computer Science and Network Security 8(4), 116–121 (2008)
21. Fleuret, F.: Fast binary feature selection with conditional mutual information. The Journal of Machine Learning Research 5, 1531–1555 (2004), http://dl.acm.org/citation.cfm?id=1005332.1044711
22. Gold, P., Freedman, S.O.: Demonstration of tumor-specific antigens in human colonic carcinomata by immunological tolerance and absorption techniques. The Journal of Experimental Medicine 121, 439–462 (1965)
23. Hammarstrom, S.: The carcinoembryonic antigen (cea) family: structures, suggested functions and expression in normal and malignant tissues. Seminars in Cancer Biology 9, 67–81 (1999)
24. Holland, J.H.: Adaption in Natural and Artifical Systems. University of Michigan Press (1975)
25. Keshaviah, A., Dellapasqua, S., Rotmensz, N., Lindtner, J., Crivellari, D., et al.: Ca15-3 and alkaline phosphatase as predictors for breast cancer recurrence: a combined analysis of seven international breast cancer study group trials. Annals of Oncology 18(4), 701–708 (2007)
26. Koepke, J.A.: Molecular marker test standardization. Cancer 69, 1578–1581 (1992)
27. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137–1143. Morgan Kaufmann (1995)
28. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. The MIT Press (1992)
29. Kronberger, G.K.: Symbolic regression for knowledge discovery - bloat, overfitting, and variable interaction networks. Ph.D. thesis, Institute for Formal Models and Verification, Johannes Kepler University Linz (2010)

30. LaFleur-Brooks, M.: Exploring Medical Language: A Student-Directed Approach, 7th edn. Mosby Elsevier, St. Louis (2008)
31. Lai, R.S., Chen, C.C., Lee, P.C., Lu, J.Y.: Evaluation of cytokeratin 19 fragment (cyfra 21-1) as a tumor marker in malignant pleural effusion. Japanese Journal of Clinical Oncology 29(9), 421–424 (1999)
32. Langdon, W.B., Poli, R.: Foundations of Genetic Programming. Springer, Heidelberg (2002)
33. Ljung, L.: System Identification – Theory For the User, 2nd edn. PTR Prentice Hall, Upper Saddle River (1999)
34. Maton, A., Hopkins, J., McLaughlin, C.W., Johnson, S., Warner, M.Q., LaHart, D., Wright, J.D.: Human Biology and Health. Prentice Hall, Englewood Cliffs (1993)
35. Meyer, P., Bontempi, G.: On the use of variable complementarity for feature selection in cancer classification. In: Evolutionary Computation and Machine Learning in Bioinformatics, pp. 91–102 (2006)
36. Mizejewski, G.J.: Alpha-fetoprotein structure and function: relevance to isoforms, epitopes, and conformational variants. Experimental Biology and Medicine 226(5), 377–408 (2001)
37. Nelles, O.: Nonlinear System Identification. Springer, Heidelberg (2001)
38. Niv, Y.: Muc1 and colorectal cancer pathophysiology considerations. World Journal of Gastroenterology 14(14), 2139–2141 (2008)
39. Osman, N., O'Leary, N., Mulcahy, E., Barrett, N., Wallis, F., Hickey, K., Gupta, R.: Correlation of serum ca125 with stage, grade and survival of patients with epithelial ovarian cancer at a single centre. Irish Medical Journal 101(8), 245–247 (2008)
40. Rai, A.J., Zhang, Z., Rosenzweig, J., Ming Shih, I., Pham, T., Fung, E.T., Sokoll, L.J., Chan, D.W.: Proteomic approaches to tumor marker discovery. Archives of Pathology & Laboratory Medicine 126(12), 1518–1526 (2002)
41. Rosen, D.G., Wang, L., Atkinson, J.N., Yu, Y., Lu, K.H., Diamandis, E.P., Hellstrom, I., Mok, S.C., Liu, J., Bast, R.C.: Potential markers that complement expression of ca125 in epithelial ovarian cancer. Gynecologic Oncology 99(2), 267–277 (2005)
42. Shannon, C.E.: A mathematical theory of communication. The Bell Systems Technical Journal 27, 379–423 (1948)
43. Tallitsch, R.B., Martini, F., Timmons, M.J.: Human anatomy, 5th edn. Pearson/Benjamin Cummings, San Francisco (2006)
44. Tesmer, M., Estevez, P.A.: Amifs: Adaptive feature selection by using mutual information. In: IEEE International Joint Conference on Neural Networks, vol. 1 (2004)
45. Thompson, I.M., Pauler, D.K., Goodman, P.J., Tangen, C.M., et al.: Prevalence of prostate cancer among men with a prostate-specific antigen level $<= 4.0$ ng per milliliter. New England Journal of Medicine 350(22), 2239–2246 (2004)
46. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
47. Wagner, S.: Heuristic optimization software systems – modeling of heuristic optimization algorithms in the heuristiclab software environment. Ph.D. thesis, Johannes Kepler University Linz (2009)
48. Wagner, S., Affenzeller, M.: SexualGA: Gender-specific selection for genetic algorithms. In: Callaos, N., Lesso, W., Hansen, E. (eds.) Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2005). International Institute of Informatics and Systemics, vol. 4, pp. 76–81 (2005)
49. Williams, P.W., Gray, H.D.: Gray's anatomy, 37th edn. C. Livingstone, New York (1989)
50. Winkler, S.: Evolutionary system identification - modern concepts and practical applications. Ph.D. thesis, Institute for Formal Models and Verification, Johannes Kepler University Linz (2008)

51. Winkler, S., Affenzeller, M., Jacak, W., Stekel, H.: Classification of tumor marker values using heuristic data mining methods. In: Proceedings of the GECCO 2010 Workshop on Medical Applications of Genetic and Evolutionary Computation, MedGEC 2010 (2010)
52. Winkler, S., Affenzeller, M., Jacak, W., Stekel, H.: Identification of cancer diagnosis estimation models using evolutionary algorithms - a case study for breast cancer, melanoma, and cancer in the respiratory system. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2010 (2011)
53. Winkler, S., Affenzeller, M., Kronberger, G., Kommenda, M., Wagner, S., Jacak, W., Stekel, H.: Feature selection in the analysis of tumor marker data using evolutionary algorithms. In: Proceedings of the 7th International Mediterranean and Latin American Modelling Multiconference, pp. 1–6 (2010)
54. Winkler, S., Affenzeller, M., Kronberger, G., Kommenda, M., Wagner, S., Jacak, W., Stekel, H.: On the use of estimated tumor marker classifications in tumor diagnosis prediction - a case study for breast cancer. In: Proceedings of 23rd IEEE European Modeling & Simulation Symposium, EMSS 2011 (2011)
55. Yin, B.W., Dnistrian, A., Lloyd, K.O.: Ovarian cancer antigen CA125 is encoded by the MUC16 mucin gene. International Journal of Cancer 98(5), 737–740 (2002)
56. Yonemori, K., Ando, M., Taro, T.S., Katsumata, N., Matsumoto, K., Yamanaka, Y., Kouno, T., Shimizu, C., Fujiwara, Y.: Tumor-marker analysis and verification of prognostic models in patients with cancer of unknown primary, receiving platinum-based combination chemotherapy. Journal of Cancer Research and Clinical Oncology 132(10), 635–642 (2006)
57. Zhong, L., Zhou, X., Wei, K., Yang, X., Ma, C., Zhang, C., Zhang, Z.: Application of serum tumor markers and support vector machine in the diagnosis of oral squamous cell carcinoma. Shanghai Kou Qiang Yi Xue (Shanghai Journal of Stomatology) 17(5), 457–460 (2008)