VARIABLE INTERACTION NETWORKS IN MEDICAL DATA

**Stephan M. Winkler $^{(a)}$, Michael Affenzeller $^{(b)}$, Gabriel Kronberger $^{(c)}$,
Michael Kommenda $^{(d)}$, Stefan Wagner $^{(e)}$, Witold Jacak $^{(f)}$, Herbert Stekel $^{(g)}$**

(a – f) University of Applied Sciences Upper Austria
Heuristic and Evolutionary Algorithms Laboratory
Softwarepark 11, 4232 Hagenberg, Austria

(a, b, f) University of Applied Sciences Upper Austria
Bioinformatics Research Group
Softwarepark 11, 4232 Hagenberg, Austria

(g) General Hospital Linz
Central Laboratory
Krankenhausstraße 9, 4021 Linz, Austria

(a) stephan.winkler@fh-hagenberg.at, (b) michael.affenzeller@fh-hagenberg.at, (c) gabriel.kronberger@fh-hagenberg.at,
(d)michael.kommenda@fh-hagenberg.at, (e)stefan.wagner@fh-hagenberg.at,
(f) witold.jacak@fh-hagenberg.at, (g) herbert.stekel@akh.linz.at

ABSTRACT

In this paper we describe the identification of variable interaction networks in a medical data set. The main goal is to generate mathematical models for standard blood parameters as well as tumor markers using other available parameters in this data set. For each variable we identify those variables that are most relevant for modeling it; relevance of a variable can in this context be defined via the frequency of its occurrence in models identified by evolutionary machine learning methods or via the decrease in modeling quality after removing it from the data set.

Several data based modeling approaches implemented in HeuristicLab have been applied for identifying estimators for selected tumor markers and cancer diagnoses: Linear regression and support vector machines (optimized using evolutionary algorithms) as well as genetic programming.

Keywords: medical data analysis, data mining, evolutionary algorithms, variable interaction networks

1. INTRODUCTION, RESEARCH GOALS

In this paper we present research results achieved within the Josef Ressel Centre for Heuristic Optimization *Heureka!*: Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been analyzed in order to identify mathematical models for cancer diagnoses. We have used a medical database compiled at the central laboratory of AKH in the years 2005 – 2008: A series of routinely measured blood values of thousands of patients are available as well as several tumor markers (TMs, substances found in humans that can be used as indicators for certain types of cancer). Not all values are measured for all patients; especially tumor marker values are determined and documented mainly if there are indications for the presence of cancer. The results of empirical research work done on the data based identification of estimation models for standard blood parameters as well as tumor markers are presented in this paper: The main goal is to generate mathematical models for standard blood parameters as well as tumor markers using other available parameters in this data set. For each variable we identify those variables that are most relevant for modeling it; relevance of a variable can in this context be defined via the frequency of its occurrence in models identified by evolutionary machine learning methods or via the decrease in modeling quality after removing it from the data set.

2. VARIABLE INTERACTION NETWORKS

2.1. General Approach

The main goal in the identification of variable interaction networks is to determine in how far variables interact which other variables in the given data set. We identify these interactions by modeling all given variables and determining the importance / relevance of all potential input variables; influences above some pre-defined threshold are considered relevant and are shown in respective graphical representation of this interaction network. An exemplary interaction network is shown in Figure 1.

In (Winkler et al. 2006), for example, variable selection results were shown partially resembling interaction networks; in (Kronberger 2011) the author showed variable interaction networks for a blast furnace process, an industrial chemical process, and macro-economic data dependencies using symbolic regression based on genetic programming.

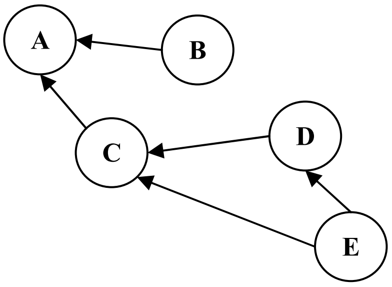

Figure 1: An exemplary variables interaction network. Variable A is influenced by variables B and C, while C is influenced by D and E (which also influences D).

2.2. Calculation of the Impact (Relevance) of Variables

In the present research work we estimate the relevance of the available variables (i.e., standard blood parameters and virtual tumor markers) with respect to models. There are several methods for calculating the relevance of variables (see for example (Winkler 2009) and (Affenzeller et al. 2009)); the following approaches are used in the context of this research work:

2.2.1. Frequency Based Relevance of Variables:

The relevance of a variable (at index *i*) with respect to a given model m can either be defined as the number of references in this model (calculated using function *freq₁*) to this variable, or simply as 1 if there is a reference to this variable and 0 if not (*freq₂*).

$$freq_1(i,m) = |\{\, v{:}\, v \in m.variables \;\&\; v.VarIndex = i\}| \tag{1}$$

$$freq_2(i,m) = \begin{cases} 1{:} & \exists\, v : (v \in m.variables \;\&\; v.VarIndex = i) \\ 0{:} & otherwise \end{cases} \tag{2}$$

2.2.2. Impact Based Relevance of Variables - Impact of Variables with Respect to Models

The estimation of a variable's impact on the evaluation of a model is done by temporarily replacing the values of exactly this variable and evaluating the model using the resulting manipulated data base. We here use replacement methods using averaging, constants, linear regression and additive Gaussian noise; a given variable (at index *i*) is replaced without changing any other part of the data basis. We transfer the original data basis *Data* consisting of *N* variables with *n* samples each to a manipulated data basis *Data$_i$*(r) with manipulated variable number *i* using a replacement function *r* as

$$\forall(i \in [1;N]){:}\, Data_{i(r)} = [Data_1, Data_2, \ldots, r(Data_i), Data_{i+1}, \ldots, Data_N] \tag{3}$$

The replacement functions used replace a particular variable using a given constant value (r_{const}), its mean value (r_{mean}), its linear trend (r_{linreg}), or the addition of Gaussian noise (r_{agn}). Now it is possible to calculate the variable's impact with respect to the model *m* by evaluating the model on the manipulated data set *Data$_i$* and measuring the resulting difference between the original output values and those calculated on the manipulated data.

This measurement can be done on the basis of the mean squared difference function *impact$_{msd}$*, e.g.:

$$impact_{msd}(i,m,r) = \frac{1}{n}\sum_{j=1}^{n}\left([eval(m,Data)]_j - [eval(m,Data_i(r))]_j\right)^2 \tag{4}$$

2.2.3. Impact Based Relevance of Variables - Impact of Variables with Respect to Algorithm Performance:

After identifying the models that are used most frequently and seem to be most important for the models created by the identification algorithms, we estimate the importance of variables (or groups of variables) by removing them from the list of available input and repeating the modeling process. Thus, again we use *Data$_i$*(r) for calculating the relevance of variable *i* and define the relevance of variable *i* with respect to a modeling method *m*, a replacement method *r* and a fitness function *f* as

$$impact_{alg}(i,m,r,f) = \frac{f(eval(m,Data))}{f(eval(m,Data_i(r)))} \tag{5}$$

3. MACHINE LEARNING METHODS APPLIED

In this section we describe the modeling methods applied for identifying estimation models for tumor markers and cancer diagnoses: On the one hand we apply hybrid modeling using machine learning algorithms and evolutionary algorithms for parameter optimization and feature selection (as described in Section 2.1), on the other hand we use genetic programming (as described in Section 2.2). In (Winkler et al. 2011), for example, these methods have also been described in detail.

3.1. Hybrid Modeling Using Machine Learning Algorithms and Evolutionary Algorithms for Parameter Optimization and Features 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 conducts to better analyses. Given a set of *n* features *F = {f$_1$, f$_2$, ..., f$_n$}*, our goal here is to find a subset of *F*, *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 by comparing the estimated values to the originally given target values:

$$fitness(F', m, p) = \alpha \cdot |F'|/|F| + \\ + (1- \alpha) \cdot (1-R^2(est(F',m,p),orig)). \qquad (5)$$

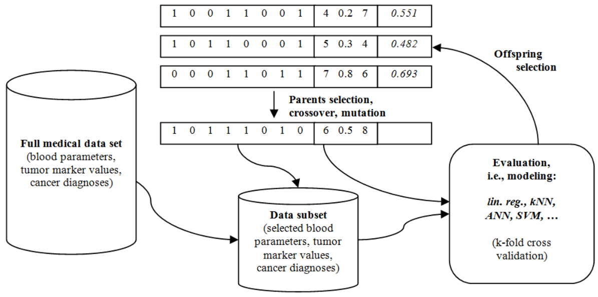
1 0 0 1 1 0 0 1 | 4 0.2 7 | 0.551
1 0 1 1 0 0 0 1 | 5 0.3 4 | 0.482
0 0 0 1 1 0 1 1 | 7 0.8 6 | 0.693
Parents selection, crossover, mutation
1 0 1 1 1 0 1 0 | 6 0.5 8
Offspring selection
Full medical data set (blood parameters, tumor marker values, cancer diagnoses)
Data subset (selected blood parameters, tumor marker values, cancer diagnoses)
Evaluation, i.e., modeling:
lin. reg., kNN, ANN, SVM, …
(k-fold cross validation)

Figure 2: A hybrid evolutionary algorithm for feature selection and parameter optimization in data based modeling. Machine learning algorithms are applied for evaluating feature sets.

In (Alba et al. 2007), for example, the use of evolutionary algorithms for feature selection optimization is discussed in detail in the context of gene selection in cancer classification. 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 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 (Holland 1975) and (Eiben 2003).

We have used strict offspring selection (Affenzeller et al. 2009): Individuals are accepted to become members of the next generation if they are evaluated better than both parents.

In (Winkler et al. 2011) we have documented classification accuracies for tumor diagnoses using this approach for optimizing feature set and modeling parameters.

The following machine learning algorithms have been applied for identifying estimators for selected tumor markers and cancer diagnoses: Linear regression and support vector machines.

3.2. Genetic Programming

As an alternative to the approach described in the previous sections we have also applied a classification algorithm based on genetic programming (GP, Koza (1992)) using a structure identification framework described in Winkler (2008) and Affenzeller et al. (2009).

We have used the following parameter settings for our GP test series: The mutation rate was set to 15%, gender specific parents selection (Wagner 2005) (combining random and roulette selection) was applied as well as strict offspring selection (Affenzeller et al. 2009) (OS, with success ratio as well as comparison factor set to 1.0). The functions set described in (Winkler 2008) (including arithmetic as well as logical functions) was used for building composite function expressions.

In addition to splitting the given data into training and test data, the GP based training algorithm used in our research project 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.

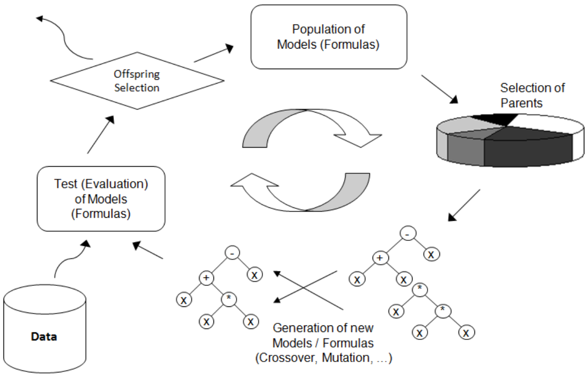
Population of Models (Formulas)
Selection of Parents
Offspring Selection
Test (Evaluation) of Models (Formulas)
Data
Generation of new Models / Formulas (Crossover, Mutation, ...)

Figure 3: Genetic programming including offspring selection.

4. DATA BASE

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 a unique ID number of the respective patient, the date of the measurement series, the ID number of the measurement, standard blood parameters, tumor marker values, and cancer diagnosis information. Patients' personal data were at no time available for the authors except for 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 many missing values simply because not all blood values are measured during each examination. Further details about the data set and necessary data preprocessing steps can for example be found in Winkler et al. (2010) and Winkler et al. (2011), e.g.

Standard blood parameters include for example the patients' sex and age, information about the amount of cholesterol and iron found in the blood, the amount of hemoglobin, and the amount of red and white blood cells; in total, 29 routinely available patient parameters are available.

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 Koepke (1992) (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) and Yonemori (2006).

As described in Winkler et al. (2010) and Winkler et al. (2011), information about the following tumor markers is stored in the AKH database: AFP, CA 125, CA 15-3, CA 19-9, CEA, CYFRA, fPSA, NSE, PSA, S-100, SCC, and TPS.

5. CASE STUDY RESEARCH RESULTS: FEATURES RELEVANT FOR PREDICTING BREAST CANCER DIAGNOSIS AND THE CORRESPONDING VARIABLES INTER-ACTION NETWORK

5.1. Data Preprocessing

From the data base mentioned in the previous section we have selected a subset of samples containing standard blood data, tumor markers, and the information whether the corresponding patient was diagnosed with breast cancer or not. All samples represent female patients; we have selected samples that contain at least 80% valid values and replaced all missing values by the median of the respective variable.

Furthermore, variables containing less than 80% valid values were removed. Finally, the samples were filtered in such a way that 50% of the remaining samples represent patients with breast cancer and 50% patients without a positive breast cancer diagnosis. This leads to a data set consisting of 636 samples; Table 1 summarizes all relevant key parameters of the so compiled data base.

Table 1: Data base used for breast cancer diagnosis variables interaction network case study

Number of samples	636
Number of samples in class 0 (no breast cancer diagnosed) class 1 (breast cancer diagnosed)	 318 318
Number of available variables (features) standard values tumor markers (AFP, C125, C153, PSA)	30 26 4

5.2. Modeling Results, Variable Impacts and Variable Interaction Network

5.2.1. Impact of Variables on Breast Cancer Diagnosis

For estimating breast cancer diagnoses we have used the data set described in 5.1 and applied tree based genetic programming using the HeuristicLab 3.3.6 framework (Wagner (2009), http://dev.heuristiclab.com); the most important GP settings are summarized in Table 2.

GP was used for learning models for the breast cancer diagnosis using all available variables; subsequently, following the strategy described in Section 2.2.3 all features were (one by one) removed from the data basis and the resulting classification accuracy compared to the accuracy achieved using all variables. Figure 4 visualizes the most important results of these tests showing the average test accuracies of five repetitions with five-fold cross-validation: Using all variables 75.47% of the given samples are correctly classified, after removing for example C125 the accuracy drops to 69.81%, and after removing C153 even to 60.85%. All not displayed features do not seem to have a relevant impact as their removal did not lead to a decrease of the classification accuracy below 75%.

Table 2: GP parameters

Population size	1000
Size limits maximum tree height maximum tree size	 8 100
Mutation rate	15%
Offspring selection max. selection pressure	strict 100
Function basis	Arithmetic and logic functions

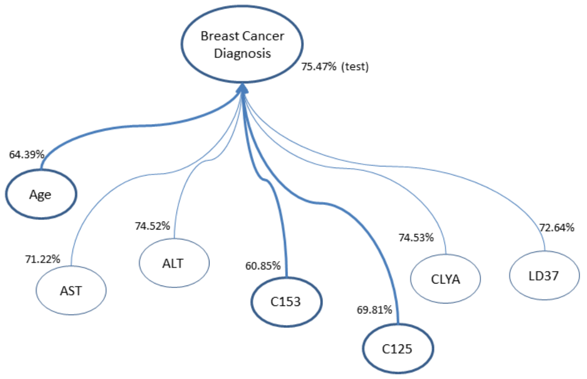

Figure 4: Classification accuracies for breast cancer diagnoses

5.2.2. Medical Variables Interaction Network

Following the modeling of breast cancer diagnoses, all variables in the data set have been modeled using the hybrid approach described in Section 3.1: A genetic algorithm with strict offspring selection was used for optimizing feature selections and algorithm parameters for linear regression and support vector machines that are used for modeling each available variable. Table 3 summarizes all relevant parameters for this modeling approach.

Again, five repetitions were executed (applying five-fold cross-validation) for each modeling scenario. All variables that were eventually selected by the evolutionary process (either for linear regression modeling or support machines) were collected; from this information we have developed a variables interaction network that represents all relevant relationships among variables. Figure 5 represents the so composed network showing relevant impacts for variables that can be modeled with test prediction squared correlation coefficient (R^2) greater than 0.4

Table 3: GA + linReg / SVM parameters

Population size	10
Mutation rates feature selection flip algorithm parameter mutation	 30% 30%
Offspring selection max. selection pressure	strict 100
Initial variable selection probability	30%

6. CONCLUSION

In this paper we have discussed a data-based approach for calculating the importance of medical variables for estimating cancer diagnoses using machine learning; furthermore, this modeling based impact analysis has also been used for constructing variable interaction networks describing the relationship among medical features.

As we have seen in the empirical section of this paper, there are several features that are of significant importance for the success of estimating breast cancer diagnoses; not surprisingly, the most important ones are tumor markers. The subsequently generated interaction network shows the most important relationships among the analyzed medical variables.

ACKNOWLEDGMENTS

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

Affenzeller, M., Winkler, S., Wagner, S., A. Beham, 2009. *Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications*. Chapman & Hall/CRC. ISBN 978-1584886297. 2009.

Alba, E., García-Nieto, J., Jourdan, L., Talbi, E.-G., 2005. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. *IEEE Congress on Evolutionary Computation 2007*, pp. 284 – 290.

Eiben, A.E. and Smith, J.E. 2003. Introduction to Evolutionary Computation. *Natural Computing Series*, Springer-Verlag Berlin Heidelberg.

Holland, J. H., 1975. *Adaption in Natural and Artifical Systems.* University of Michigan Press.

Kronberger, G., 2011. *Symbolic Regression for Knowledge Discovery*. Schriften der Johannes Kepler Universität Linz, Universitätsverlag Rudolf Trauner.

Koepke, J. A., 1992. Molecular marker test

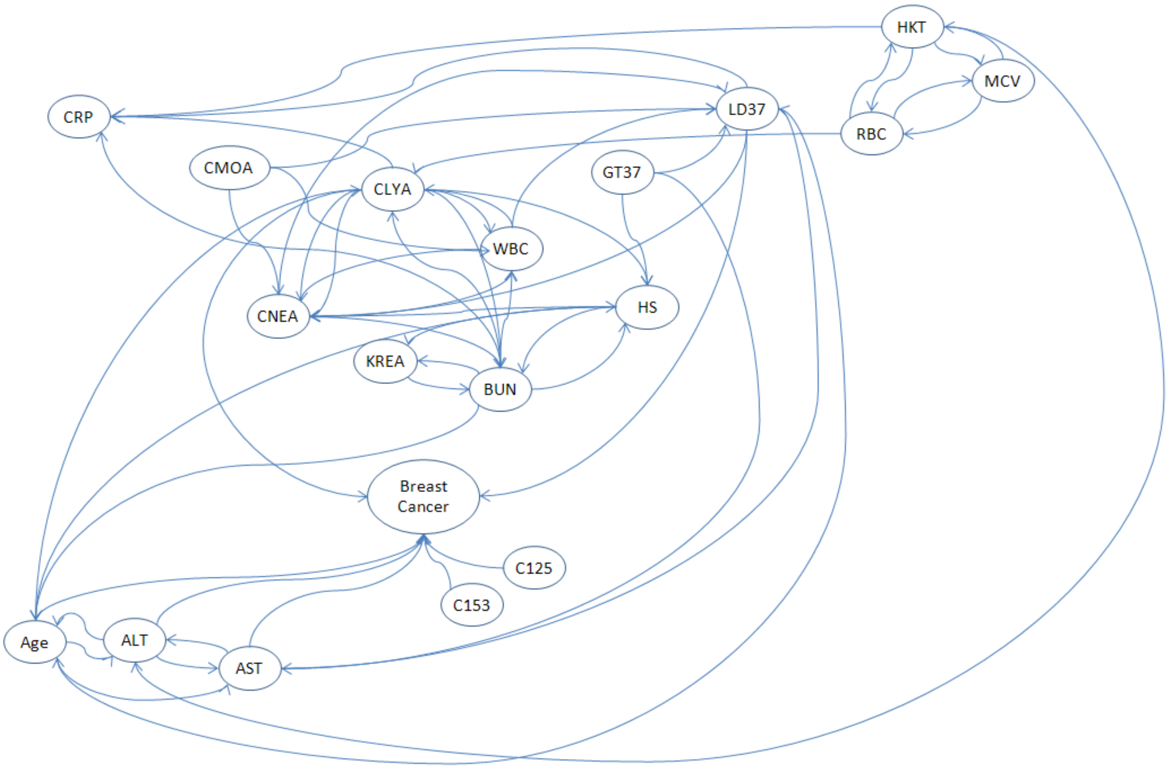

Figure 5: Interaction network of medical variables

standardization. *Cancer*, 69, pp. 1578–1581.

Wagner, S., 2009. *Heuristic Optimization Software Systems - Modeling of Heuristic Optimization Algorithms in the HeuristicLab Software Environment*. PhD Thesis, Institute for Formal Models and Verification, Johannes Kepler University Linz, Austria.

Winkler, S. M., Efendic, H., del Re, L., 2006. Quality Pre-Assesment in Steel Industry Using Data Based Estimators. In S. Cierpisz, K. Miskiewicz, A. Heyduk (ed.), *Proceedings of the IFAC Workshop MMM'2006 on Automation in Mining, Mineral and Metal Industry*, pp. 185-190. International Federation for Automatic Control.

Winkler, S., Affenzeller, M., Jacak, W., Stekel, H., 2010. Classification of Tumor Marker Values Using Heuristic Data Mining Methods. *Proceedings of Genetic and Evolutionary Computation Conference 2010, Workshop on Medical Applications of Genetic and Evolutionary Computation*, pp. 1915–1922.

Winkler, S., Affenzeller, M., Jacak, W., Stekel, H., 2011. Identification of Cancer Diagnosis Estimation Models Using Evolutionary Algorithms – A Case Study for Breast Cancer, Melanoma, and Cancer in the Respiratory System. *Proceedings of Genetic and Evolutionary Computation Conference 2011, Workshop on Medical Applications of Genetic and Evolutionary Computation*.

Winkler, S., 2009. *Evolutionary System Identification - Modern Concepts and Practical Applications.* Schriften der Johannes Kepler Universität Linz, Reihe C: Technik und Naturwissenschaften. Universitätsverlag Rudolf Trauner. ISBN 978-3-85499-569-2.

Yonemori, K., Ando, M., Taro, T. S., Katsumata, N., Matsumoto, K., Yamanaka, Y., Kouno, T., Shimizu, C., Fujiwara, Y., 2006. 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), pp. 635–642.

AUTHORS BIOGRAPHIES

STEPHAN M. WINKLER received his PhD in engineering sciences in 2008 from Johannes Kepler University (JKU) Linz, Austria. His research interests include genetic programming, nonlinear model identification and machine learning. Since 2009, Dr. Winkler is professor at the Department for Medical and Bioinformatics at the University of Applied Sciences (UAS) Upper Austria at Hagenberg Campus; since 2010, Dr. Winkler is head of the Bioinformatics Research Group at UAS, Hagenberg.

MICHAEL AFFENZELLER has published several papers, journal articles and books dealing with theoretical and practical aspects of evolutionary computation, genetic algorithms, and meta-heuristics in general. In 2001 he received his PhD in engineering sciences and in 2004 he received his habilitation in applied systems engineering, both from the Johannes Kepler University of Linz, Austria. Michael Affenzeller is professor at UAS and head of the Josef Ressel Center *Heureka!* at Hagenberg.

GABRIEL KRONBERGER received his PhD in engineering sciences in 2010 from JKU Linz, Austria, and is a research associate at the UAS Research Center Hagenberg. His research interests include genetic programming, machine learning, and data mining and knowledge discovery.

MICHAEL KOMMENDA finished his studies in bioinformatics at Upper Austria University of Applied Sciences in 2007. Currently he is a research associate at the UAS Research Center Hagenberg working on data-based modeling algorithms for complex systems within *Heureka!*.

STEFAN WAGNER received his PhD in engineering sciences in 2009 from JKU Linz, Austria; he is professor at the Upper Austrian University of Applied Sciences (Campus Hagenberg). Dr. Wagner’s research interests include evolutionary computation and heuristic optimization, theory and application of genetic algorithms, and software development.

WITOLD JACAK received his PhD in electric engineering in 1977 from the Technical University Wroclaw, Poland, where he was appointed Professor for Intelligent Systems in 1990. Since 1994 Prof. Jacak is head of the Department for Software Engineering at the Upper Austrian University of Applied Sciences (Campus Hagenberg).

HERBERT STEKEL received his MD from the University of Vienna in 1985. Since 1997 Dr. Stekel is chief physician at the General Hospital Linz, Austria, where Dr. Stekel serves as head of the central laboratory.