Appendix S1
Mathematical foundations of the RLS method of feature selection

We consider a data set that describes \( m \) patients \( O_j \), where \( j \) is the order number (index) of the patient \( O_j \) in the cohort. Each patient \( O_j \) is represented by the \( n \)-dimensional feature vector \( x_j = [x_{j1}, \ldots, x_{jn}]^T \), where \( n \) is the total number of features (i.e., clinical characteristics and results of diagnostic tests including genetic polymorphisms). The component \( x_{ji} \) of the vector \( x_j \) is the numerical result of the \( i \)-th measurement or diagnostic test taken on patient \( O_j \). Both the clinical characteristics and the results of laboratory tests performed in patient \( O_j \) can be represented by the components \( x_{ji} (x_{ji} \in \{0,1\} \text{ or } x_{ji} \in \mathbb{R}^1) \).

Linear separability of two learning sets
The feature vectors \( x_j \) can be considered as points in the feature space \( F \). We examine a possibility of linear separation of two groups of patients \( O_j \) (for example \( m^+ \) healthy patients vs. \( m^- \) ill patients) in \( F \) and for that purpose two learning sets \( G^+ = \{x_j; j \in J^+\} \) and \( G^- = \{x_j; j \in J^-\} \) where \( J^+ \) and \( J^- \) are disjoined sets of indices \( j (J^+ \cap J^- = \emptyset) \), are defined. The set \( J^+ \) contains \( m^+ \) indices, and the set \( J^- \) contains \( m^- \) indices.

**Definition 1** The sets \( G^+ \) and \( G^- \) are linearly separable in the feature space \( F \), if and only if there exist such a weight vector \( w = [w_1, \ldots, w_n]^T \in \mathbb{R}^n \), and threshold \( \theta \in \mathbb{R}^1 \) that \((\forall x_j \in G^+) : w^T x_j > \theta \) and \((\forall x_j \in G^-) : w^T x_j < \theta \).

The parameters \( w \) and \( \theta \) define the hyperplane \( H(w, \theta) \) in the feature space \( F \):

\[
H(w, \theta) = \{ x : w^T x = \theta \}
\]  

(1)

**Figure 1.** Linearly separable sets. An example of linearly separable sets \( G^+ \) (denoted by \( + \)) and \( G^- \) (denoted by \( - \)) in the two-dimensional feature space \( F \), where \( m^+ = 8 \) and \( m^- = 6 \).

If the sets \( G^+ \) and \( G^- \) are linearly separable then all the elements \( x_j \) of the set \( G^+ \) are situated on the positive side of the hyperplane \( H(w, \theta) \) \( w^T x_j > \theta \) and all the elements of the set \( G^- \) are situated on the negative side of this hyperplane \( w^T x_j < \theta \), see Figure 1.

The concept of linear separability is used in the theory of neural networks (Perceptron theory) and in pattern recognition methods [1]. Optimal linear classifiers can be designed through exploration of the linear separability of the data subsets [2].
Convex and piecewise-linear criterion functions

The positive penalty function \( \varphi^+(w, \theta; x_j) \) is defined for \( x_j \in G^+ [1,2] \):

\[
\varphi^+(w, \theta; x_j) = 1 + \theta - w^T x_j \quad \text{if} \quad w^T x_j < \theta + 1 \\
\varphi^+(w, \theta; x_j) = 0 \quad \text{otherwise}
\]  

(2)

Similarly, the negative penalty function \( \varphi^-(w, \theta; x_j) \) is defined for \( x_j \in G^- [1,2] \):

\[
\varphi^-(w, \theta; x_j) = 1 - \theta + w^T x_j \quad \text{if} \quad w^T x_j > \theta - 1 \\
\varphi^-(w, \theta; x_j) = 0 \quad \text{otherwise}
\]  

(3)

The penalty functions \( \varphi^+(w, \theta; x_j) \) and \( \varphi^-(w, \theta; x_j) \) are convex and piecewise-linear (CPL) (Figure 2).

The function \( \varphi^+(w, \theta; x_j) \) is equal to zero if the feature vector \( x_j \in G^+ \) is situated on the positive side of the hyperplane \( H(w, \theta) \) and is not too close to it. Similarly, \( \varphi^-(w, \theta; x_j) \) is equal to zero if the vector \( x_j \in G^- \) is situated on the negative side of the hyperplane \( H(w, \theta) \) and is not too close to it.

\[ w^T x_j - \theta \]

Figure 2. The penalty functions. The penalty functions \( \varphi^+(w, \theta; x_j) \), equation (2), and \( \varphi^-(w, \theta; x_j) \), equation (3).

The perceptron criterion function \( \Phi(w, \theta) \) is defined on the learning sets \( G^+ \) and \( G^- \) as the weighted sum of the penalty functions \( \varphi^+(w, \theta; x_j) \) and \( \varphi^-(w, \theta; x_j) \) [1,2]:

\[
\Phi(w, \theta) = \sum_{j \in J^+} \alpha_j \varphi^+(w, \theta; x_j) + \sum_{j \in J^-} \alpha_j \varphi^-(w, \theta; x_j)
\]  

(4)

where nonnegative parameters \( \alpha_j \) can represent prices of particular feature vectors \( x_j \). The default values of the parameters \( \alpha_j \) are equal to \( \alpha_j = 1/(2m^+) \) for \( x_j \in G^+ \) and \( \alpha_j = 1/(2m^-) \) for \( x_j \in G^- \).

The optimal parameters \( w^* \) and \( \theta^* \) are defined as the values of \( w \) and \( \theta \) at the minimum \( \Phi(w^*, \theta^*) = \Phi^* \geq 0 \) of the criterion function \( \Phi(w, \theta) \). It was proved that there is a unique set of optimal parameters \( w^* \) and \( \theta^* \) and that \( \Phi^* \) is equal to zero if and only if the sets \( G^+ \) and \( G^- \) are linearly separable [2].

The perceptron criterion function \( \Phi(w, \theta) \) can be defined also on the feature vectors \( x_j \) with reduced dimensionality, i.e. for vectors \( x_j \) that are created with \( k \) features instead of the whole set of \( n \) features \( (x_j \in F_k, k \leq n) \). The minimal value \( \Phi_k^* \) of the function \( \Phi_k(w, \theta) \) on \( F_k \) can be used as the measures of the degree of the linear separability of the learning sets \( G^+ \) and \( G^- \) in the reduced feature subspaces \( F_k \). The zero value \( \Phi_k^* \) indicates that the learning sets \( G^+ \) and \( G^- \) are linearly separable in the reduced feature subspaces \( F_k \). The reduction of features may increase the value of \( \Phi_k^* \) [2].

A modified criterion function \( \Psi_\lambda(w, \theta) \), which includes additional CPL penalty functions based on the absolute values of components \( |w_i| \) of vector \( w \) and the costs \( \gamma_i \) \( (\gamma_i > 0) \) of particular features \( x_i \),
was found to be useful for the feature selection [2]:

\[ \Psi_\lambda(w, \theta) = \Phi(w, \theta) + \lambda \sum_{i \in \{1, \ldots, n\}} \gamma_i |w_i| \]  \hspace{1cm} (5)

where \( \lambda \geq 0 \) is the cost level. The standard values of the parameters \( \gamma_i \) are equal to one.

The modified criterion function \( \Psi_\lambda(w, \theta) \) is used in the relaxed linear separability (RLS) method of feature subset selection [3, 4]. The regularization component \( \lambda \sum_{i \in \{1, \ldots, n\}} \gamma_i |w_i| \) used in the modified criterion function \( \Psi_\lambda(w, \theta) \) is similar to that used in the Lasso method developed in the framework of the regression analysis for the model selection [5]. The main difference between the Lasso and the RLS methods is in the types of the basic criterion functions. The basic criterion function typically used in the Lasso method is the residual sum of squares, whereas the perceptron criterion function \( \Phi(w, \theta) \), equation (4), is used in the RLS method. This difference effects the computational techniques applied to minimize of the criterion functions. The criterion function \( \Psi_\lambda(w, \theta) \), similarly to the function \( \Phi(w, \theta) \) is convex and piecewise-linear (CPL). The basis exchange algorithms, which are similar to linear programming, allow the identification of the minimum of the function \( \Psi_\lambda(w, \theta) \) or \( \Phi(w, \theta) \) even in the case of large, high dimensional learning sets \( G^+ \) and \( G^- \) [2].

The optimal parameters \( w^*_\lambda = [w^*_{\lambda 1}, \ldots, w^*_{\lambda n}]^T \) are used in the feature reduction rule [3]:

\[ (w^*_{\lambda i} = 0) \Rightarrow \text{(the feature } i \text{ is omitted)} \]  \hspace{1cm} (6)

The reduction of such feature \( i \) that is related to the weight \( w^*_{\lambda i} \) equal to zero \( (w^*_{\lambda i} = 0) \) does not change the value of the inner product \( (w^*_{\lambda})^T x_j \). The positive location \( (w^*_{\lambda})^T x_j > \theta^*_\lambda \) or the negative location \( (w^*_{\lambda})^T x_j < \theta^*_\lambda \) of all feature vector \( x_j \) in respect to the optimal separating hyperplane \( H(w^*_\lambda, \theta^*_\lambda) \) remains unchanged.

Relaxed linear separability (RLS) method of feature subset selection

The \( n \)-dimensional feature space \( F \) is composed of all the \( n \) features \( x_i \) from the set \( \{x_1, \ldots, x_n\} \). Feature reduction based on the rule (6) results in appearance of the reduced feature subspaces \( F_k \). The symbol \( F_k \) denotes a feature subspace that is composed of \( k \) features \( x_i \) from the set \( \{x_1, \ldots, x_n\} \). The reduced learning sets \( G_k^+ \) and \( G_k^- \) are composed from the reduced feature vectors \( x_j \) with the same \( k \) features.

Successive increase of the value of the cost level \( \lambda \) in the criterion function \( \Psi_\lambda(w, \theta) \) causes reduction of some features \( x_j \) and, as the result, allows to generate the descended sequence of feature subspaces \( F_k \) with decreased dimensionality \( (F_k \supset F_{k-1}) \) [2]:

\[ F = F_n \rightarrow F_{n-1} \rightarrow \ldots \rightarrow F_k \rightarrow \ldots \rightarrow F_1 \]  \hspace{1cm} (7)

The reduced feature subspaces \( F_k \) are formed from the feature space \( F \) as a result of the omission of certain features according to rule (6). The reduced feature vectors \( x_j \in F_k \) are obtained from the feature vectors \( x_j \) in the same manner. The reduction of features may result in the loss of linear separability of the sets \( G^+ \) and \( G^- \) for some \( k \).

In accordance with the relaxed linear separability (RLS) method, the generation of sequence (7) of the feature subspaces \( F_k \) is determinisitic [3, 4]. Each step \( F_k \rightarrow F_{k-1} \) is realized by an increase in cost \( \lambda \) values: \( \lambda_k \rightarrow \lambda_{k-1} = \lambda_k + \Delta \) \( (\Delta > 0) \) in the criterion function \( \Psi_\lambda(w, \theta) \). The further increase of the parameter \( \lambda \) allows to reduce the next feature from the reduced vectors in \( F_k \).

The feature reduction process should be stopped by fixing the cost \( \lambda \) at a specific level. The choice of the stop criterion of the sequence (7) is a crucial issue in RLS method.
**RLS stop criterion based on minimal error rate of linear classifier**

Two stop criteria for the feature reduction were examined for the RLS method. The first of these criteria was based on the error rate of the optimal linear classifiers constructed in the successive feature subspaces \( F_k \) [3]. The optimal linear classifier was constructed in each \( F_k \) by minimizing the perceptron criterion function \( \Phi_k(w, \theta) \) on the reduced feature vectors \( x_j \in F_k \). The parameters \( w_k^* \) and \( \theta_k^* \) at the minimum of the criterion function \( \Phi_k(w, \theta) \) determine the following decision rule of the optimal linear classifier \( LC_k(w_k^*, \theta_k^*) \) in the feature subspace \( F_k \):

\[
\begin{align*}
if \ (w_k^*)^T x \geq \theta_k^*, \quad &then \quad x \ is \ allocated \ to \ the \ category \ \omega^+ \\
if \ (w_k^*)^T x < \theta_k^*, \quad &then \quad x \ is \ allocated \ to \ the \ category \ \omega^-
\end{align*}
\]

where \( x \in F_k \), the category (class) \( \omega^+ \) is represented by \( m^+ \) elements of the learning set of reduced vectors \( G_k^+ \) and the category \( \omega^- \) is represented by \( m^- \) elements of the learning set of reduced vectors \( G_k^- \).

The quality of the optimal linear classifier \( LC_k(w^*, \theta^*) \), rule (8), can be evaluated using the error estimator (apparent error rate) \( e_a(w_k^*, \theta_k^*) \) as the fraction of wrongly classified elements \( x_j \) of the sets \( G_k^+ \) and \( G_k^- \) in the feature subspace \( F_k \) [1]:

\[
e_a(w_k^*, \theta_k^*) = m_a(w_k^*, \theta_k^*)/m
\]

where \( m \) is the number of all elements in the sets \( G_k^+ \) and \( G_k^- \), and \( m_a(w_k^*, \theta_k^*) \) is the number of elements from these sets wrongly allocated by the rule (8).

If the same data are used for the design of the classifier using rule (8) and for the classifier evaluation, equation (9), the evaluation is biased (too optimistic) [6]. For example, when the learning sets \( G_k^+ \) and \( G_k^- \) are linearly separable, then all the vectors are correctly classified by the rule (8), and the apparent error (9) is equal to zero. But, as it is often found in practical applications, the error rate evaluated on vectors \( x \) that do not belong to the linearly separable learning sets \( G_k^+ \) and \( G_k^- \) may be positive.

To reduce the classifier bias, the cross validation procedures can be applied [1, 7]. In accordance with the p-fold cross validation learning sets \( G_k^+ \) and \( G_k^- \) are divided into \( p \) subsets \( P_i \), where \( i = 1, ..., p \). The vectors \( x_j \) contained in \( p - 1 \) subsets \( P_i \) are used for the definition of criterion function \( \Phi_k(i)(w, \theta) \), equation (4), and for the computation of the optimal parameters \( w_k^*(i) \) and \( \theta_k^*(i) \). The remaining vectors \( x_j \) from the other subset \( P_i \) (a test set) are used for the evaluation of error rate \( e_i^*(w_k^*(i), \theta_k^*(i)) \), equation (9). This evaluation is repeated \( p \) times, and during each turn a different p-part \( P_i \) is used as the test set. After this, the mean value \( e_{CVE} \) of the error rates \( e_i^*(w_k^*(i), \theta_k^*(i)) \), equation (9), on elements of the test sets \( P_i \) is computed.

\[
e_{CVE} = \frac{\sum_{i=1,...,p} e_i^*(w_k^*(i), \theta_k^*(i))}{p}
\]

The cross validation procedure allows for the use of different vectors for designing the classifier and for its evaluation, and, as the result, to reduce the bias of the error rate estimation. The error rate \( e_{CVE} \) estimated during the cross validation procedure is called the cross-validation error (CVE). A special case of the p-fold cross validation procedure is called the leave-one out procedure, where the number \( p \) of the parts \( P_i \) is equal to the number \( m \) of elements in the sets \( G_k^+ \) and \( G_k^- \) [8].

Another type of the classifier evaluation is based on the so called confusion matrix \( T_k(w_k^*, \theta_k^*) \) [1]:

\[
T_k(w_k^*, \theta_k^*) = \begin{bmatrix} m^{++} & m^{+-} \\ m^{-+} & m^{--} \end{bmatrix}
\]

where \( m^{++} \) is the number of elements of the set \( G_k^+ \) in the feature subspace \( F_k \) correctly allocated by the optimal classifier to the category \( \omega^+ \), and \( m^{+-} \) is the number of elements of \( G_k^+ \) wrongly allocated to the
category \( \omega^- \), \( m^- \) is the number of elements of the set \( G^-_k \) correctly allocated to the category \( \omega^- \), and \( m^+ \) is the number of elements in \( G^+_k \) wrongly allocated to the category \( \omega^+ \).

The confusion matrix \( T_k(\mathbf{w}_k^*, \theta_k^*) \) allows to represent different types of errors related to the optimal classifier in the feature subspace \( F_k \). Both the error rate \( e_{CVE} \) equation (10), as well as the confusion matrix \( T_k(\mathbf{w}_k^*, \theta_k^*) \), equation (11), can be estimated using the cross validation procedures in different feature subspaces \( F_k \) of the sequence (7).

The cross-validation error rate, \( CVE \), equation (10), of the linear classifier \( LC_k(\mathbf{w}^*, \theta^*) \), rule (8), is used in the stop criterion in the present study. The error rate \( CVE \) is evaluated for each feature subspaces \( F_k \) in the sequence (7). The feature subspace \( F_{k'} \) that is linked to the linear classifier \( LC_{k'}(\mathbf{w}^*, \theta^*) \) with the lowest error rate \( e_{CVE} \) is indicated as the one selected by the classifier. It should be stressed that the estimation of the error rates \( e_{CVE} \) is used only for indication of the feature subspace \( F_k \), which is considered as having optimal number of features and the features should not be further reduced. The estimation of the error rates \( e_{CVE}(\mathbf{w}_k^*, \theta_k^*) \) does not affect in any way the sequence (7) of feature subspaces \( F_k \). In this approach to feature subspace selection, the optimal feature subspace \( F_{k'} \) is characterized by the lowest value of \( e_{CVE} \).

**RLS stop criterion based on full linear separability**

The second type of the stop criterion used in the RLS method has been based directly on checking the linear separability of the reduced learning sets \( G^+_k \) and \( G^-_k \) [4]. This approach can be useful in the case when the learning sets \( G^+ \) and \( G^- \) are linearly separable in the full feature spaces \( F \). Let us remark, that the learning sets \( G^+ \) and \( G^- \) are usually linearly separable in the case when the number of objects, \( m \), is less than the number of the features, \( n \) [1]. Such situation appears among others in the case of genetic data sets, where the number \( n \) of the features \( x_i \) can be thousand times more than the number \( m \) of the objects (for example \( n = 50000, m = 50 \)). Typically, such learning sets \( G^+ \) and \( G^- \) are linearly separable.

In accordance with the stop criterion, the feature reduction process is carried out until the linear separability is threatened. The results of the numerical experiments with this stop criterion are described in [4]. The experiments were carried out both with a large synthetic data set as well as with two benchmarking genetic data sets (Breast cancer [9] and Leukemia [10]).

The synthetic data set contained \( m = 1000 \) feature vectors with \( n = 100 \) features. The vectors \( \mathbf{x}_j \) were generated randomly in accordance with the multivariate normal distribution \( N(0, \Sigma) \) with the covariance matrix \( \Sigma \) equal to the unit matrix \( I \) (\( \Sigma = I \)). The objects \( \mathbf{x}_j \) were divided into the learning sets \( G^+_k \) (\( m^+ = 630 \) vectors) and \( G^- \) (\( m^- = 370 \) vectors) on the basis of pre-selected linear combination of 10 features \( x_i \) (the linear key). If the value of this linear key was greater than an assumed threshold, the vector \( \mathbf{x}_j \) was attributed to the set \( G^+ \). If the value of the linear key was lower than the threshold, the vector \( \mathbf{x}_j \) was attributed to the set \( G^- \). Thus, learning sets \( G^+ \) and \( G^- \) defined in this way were linearly separable. The RLS method allowed to skip 90 features \( x_i \) while preserving the linear separability of the reduced learning sets \( G^+_k \) and \( G^-_k \), \( k \geq 10 \), and the linear key of 10 features was correctly rediscovered by the RLS method [4].

The Breast cancer data set describes 97 patients tested for the presence of the breast cancer [9]. The positive learning set \( G^+ \) contains \( m^+ = 46 \) patients and the negative set \( G^- \) contains \( m^- = 51 \) patients. Each patient is characterized by expression levels of \( n = 24481 \) genes (features). The RLS method allowed to reduce the dimensionality to \( n' = 12 \) genes while preserving the linear separability of the learning sets \( G^+_k \) and \( G^-_k \). It means that the discovered linear combination of the selected 12 features allowed for full separation of the two subgroups of patients. The computational time on standard personal computer was about 25 min.

The Leukemia data set contains expression levels of \( n = 7129 \) genes in 72 patients [10]. The positive learning set \( G^+ \) contains \( m^+ = 25 \) patients and the negative set \( G^- \) contains \( m^- = 47 \) patients. The RLS method allowed to reduce the dimension to three genes while preserving the linear separability of
the learning sets $G^+_k$ and $G^-_k$. It means that the discovered linear combination of the three selected genes allowed for 100% correct separation of the two groups of 72 patients.

References

1. Duda OR, Heart PE, Stork DG (2001) Pattern Classification. John Wiley & Sons.

2. Bobrowski L (2005) Data mining based on convex and piecewise linear (CPL) criterion functions (in Polish). Bialystok: Bialystok University of Technology.

3. Bobrowski L, Lukaszuk T (2009) Feature Selection Based on Relaxed Linear Separability. Biocybernetics and Biomedical Engineering 29: 43-59.

4. Bobrowski L, Lukaszuk T (2011) Relaxed Linear Separability (RLS) Approach to Feature (Gene) Subset Selection. In: Xia X, editor, Selected Works in Bioinformatics, InTech. URL http://www.intechopen.com/articles/show/title/relaxed-linear-separability-rls-approach-to-feature-gene-subset-selection.

5. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714-721.

6. Demšar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7: 1-30.

7. Kira K, Rendell LA (1992) A practical approach to feature selection. Proceedings of the ninth international workshop on machine learning : 249-256.

8. Lachenbruch PA (1975) Discriminant Analysis. New York: Hafner Press.

9. van ’t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536.

10. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Sciences 286: 531-537.