T-ray relevant frequencies for osteosarcoma classification

W. Withayachumnankul[a,d], B. Ferguson[b,d], T. Rainsford[d],
D. Findlay[c], S. P. Mickan[d], and D. Abbott[d]

[a]Department of Information Engineering, Faculty of Engineering,
King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

[b]Tenix Systems Pty Ltd, 2nd Avenue, Mawson Lakes, SA 5095, Australia

[c]Centre for Biomedical Engineering (CBME) and
Department of Orthopaedics & Trauma, The University of Adelaide
and Hanson Institute, SA 5005, Australia

[d]Centre for Biomedical Engineering (CBME) and
Department of Electrical & Electronic Engineering,
The University of Adelaide, SA 5005, Australia

ABSTRACT

We investigate the classification of the T-ray response of normal human bone cells and human osteosarcoma cells, grown in culture. Given the magnitude and phase responses within a reliable spectral range as features for input vectors, a trained support vector machine can correctly classify the two cell types to some extent. Performance of the support vector machine is deteriorated by the curse of dimensionality, resulting from the comparatively large number of features in the input vectors. Feature subset selection methods are used to select only an optimal number of relevant features for inputs. As a result, an improvement in generalization performance is attainable, and the selected frequencies can be used for further describing different mechanisms of the cells, responding to T-rays. We demonstrate a consistent classification accuracy of 89.6%, while the only one fifth of the original features are retained in the data set.

Keywords: Terahertz time-domain spectroscopy, T-rays, support vector machines, feature selection, osteosarcoma, signal classification, cancer detection, curse of dimensionality

1. INTRODUCTION

T-rays, spanning the range from 0.1 to 10 THz in the electromagnetic spectrum, have a great potential in biomedical applications.[1,2] This results from distinctive properties of biomolecules in this frequency range. DNA and specific molecules, such as amino acids, peptides, and proteins, have resonances at T-ray frequencies.[3] T-rays are non-ionizing radiation, and represent a totally non-invasive diagnostic technique.[4] Due to strong absorption by water, T-rays produce skin-depth level contrast, in which X-rays fail. Optical and infrared frequencies suffer from Rayleigh scattering, not present with T-rays due to the longer wavelength.[5]

Biomaterial classification is one promising application of T-rays. It employs time-gated detection techniques[6] using terahertz time-domain spectroscopy (THz-TDS) to produce high SNRs and coherent signals. The signals, when passing through materials with different quantities of interstitial water, are subject to different amounts of attenuation and dispersion. This information leads to rich features useful for classification. Woodward et al.[7,8] investigated and classified basal cell carcinoma, one form of skin cancer, *in vitro* and *in vivo* with a T-ray reflection geometry. Ferguson et al.[9] distinguished two types of meats using chirped probe T-ray imaging system. Löffler et al.[10] classified tumors in sliced tissues with T-ray pulsed imaging.

W. Withayachumnankul, Email: kwwithaw@kmitl.ac.th; B. Ferguson, Email: brad.ferguson@tenix.com;
T. Rainsford, Email: tamath@eleceng.adelaide.edu.au; D. Findlay, Email: david.findlay@adelaide.edu.au;
S. P. Mickan, Email: spmickan@eleceng.adelaide.edu.au; D. Abbott, Email: dabbott@eleceng.adelaide.edu.au

Photonics: Design, Technology, and Packaging II, edited by Derek Abbott, Yuri S. Kivshar, Halina H. Rubinsztein-Dunlop, Shanhui Fan, Proc. of SPIE Vol. 6038, 60381H, (2006) · 0277-786X/06/$15 · doi: 10.1117/12.637964

In this paper, support vector machines (SVMs) are used to discriminate normal human bone (NHB) cells and human osteosarcoma (HOS) cells, which are the most common malignant primary bone tumor.[11] T-ray frequency responses of the cells provide rich features for classification over the T-ray bandwidth. In the case of biomedical applications, such as in cancer detection, the number of available observations is smaller than the number of features. When the number of features greatly exceeds the number of observations, SVMs and also other classifiers inevitably encounter the overfitting problem. Feature subset selection methods are thus required to select only an optimal subset of features as input to SVMs. It is expected that shrinkage of input dimension will reduce test error and increase generalization performance. Two further consequences from the feature selection scheme are also expected. First, a feature ranked list is produced, leading to feature exploration. The top ranked features exhibit bands of T-ray frequencies, which interact differently with the NHB and HOS cells. There is room for further investigation into cell contrast mechanisms responding to those frequencies. Second, a computational cost for training and testing SVMs is reduced to some extent, resulting from reduction of the input dimension. However, this consequence comes into importance only when real-time discrimination of massive data sets is essential. Discussion of this is beyond the scope of this work.

The article is organized as follows. The concept of SVMs and some estimates on SVM performance are described in Section 2. Section 3 is devoted to methods on the feature subset selection. The cell preparation and inspection, along with signal processing techniques, are briefly discussed in Section 4. The implementation of feature subset selection provides T-ray frequency ranked list, shown in Section 5. In Section 6, SVMs are evaluated on input vectors containing subsets of features, selected according to the list.

2. SUPPORT VECTOR MACHINES

One major application of machine learning is pattern classification. It attempts to find unknown parameters of a discriminant function with the expectation that the optimized function can correctly assign classes or labels to unseen patterns or data. In supervised learning, the two-class discriminant function is constructed, based on a given training set, which includes input vectors and their corresponding labels:

$$(\mathbf{x}_i, y_i) \in \mathbb{R}^k \times \{-1, +1\} \qquad i = 1, \ldots, l, \tag{1}$$

assuming that the pairs are drawn independently identically distributed (iid) from an unknown probability distribution $P(\mathbf{x}, y)$.

The concept of support vector machines or SVMs, laid out by Boser et al.,[12] is that it maps the input vectors to a high-dimensional space (so-called feature space), $\mathbf{x} \mapsto \Phi(\mathbf{x}) \in \mathcal{H}$, and constructs an optimal hyperplane in that space. The mapping idea allows the linear discriminant function to perform on non-linear problems. The hyperplane in the high-dimensional space is then given by

$$f(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b \,, \tag{2}$$

where $\mathbf{w} \in \mathcal{H}$ is a normal vector of the hyperplane, and b is an offset between the hyperplane and the origin. A point that lies on the hyperplane satisfies, $\mathbf{w} \cdot \Phi(\mathbf{x}) + b = 0$. In the case of perfect separation, i.e. no training error, the following condition is held:

$$y_i \left(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b \right) \geq 1 \qquad \forall i \,. \tag{3}$$

Thus a margin, or a perpendicular distance from the hyperplane to any closest point, equals $1/\|\mathbf{w}\|$. The optimal hyperplane is constructed by maximizing this distance, according to the structural risk minimization (SRM) principle.[13] This can be formulated as a quadratic optimization problem[14]

$$W^2 = \min_{\mathbf{w}, b} \; \frac{1}{2} \|\mathbf{w}\|^2 \,, \tag{4}$$

subject to Equation 3. As $\mathbf{w}$ lies in the feature space, the minimization problem cannot be solved directly. By introducing Lagrange multipliers α_i, $i = 1, \ldots, l$, corresponding to the input vector $\mathbf{x}_i$, we form the following Lagrangian with respect to the primal variables,

$$L_P = W^2(\mathbf{w}, b, \boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \min_{\mathbf{w}, b} \; \frac{1}{2} \|\mathbf{w}\|^2 \; - \sum_{i=1}^{l} \alpha_i (y_i (\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) - 1) \qquad ; \alpha_i \geq 0 \quad \forall i \,. \tag{5}$$

Saddle points of the Lagrangian with respect to $\mathbf{w}$ and b are

$$\frac{\partial L_P}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \Phi(\mathbf{x}_i) = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \Phi(\mathbf{x}_i) \; , \tag{6}$$

$$\frac{\partial L_P}{\partial b} = -\sum_{i=1}^{l} \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \; . \tag{7}$$

By substituting Equation 6 and 7 back into Equation 5 and replacing $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ with the kernel function $\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j)$, we have the Lagrangian with respect to the dual variable,

$$L_D = W^2(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) \; , \tag{8}$$

subject to

$$\alpha_i \geq 0 \quad \forall i \qquad \text{and} \qquad \sum_{i=1}^{l} \alpha_i y_i = 0 \; . \tag{9}$$

Using ordinary quadratic programming (QP) methods to solve for a set of Lagrange multipliers, $\boldsymbol{\alpha}$, is not feasible when the matrix becomes large. Particular methods designed for such problems are chunking[12] or Osuna's decomposition.[15] In this paper we implement the method of sequential minimal optimization (SMO), proposed by Platt,[16] to solve the Lagrangian.

Once the Lagrange multipliers are achieved, the non-linear discriminant function via the kernel trick is constructed from the hyperplane (Equation 2) with weights given by Equation 6,

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b \right) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i \mathrm{K}(\mathbf{x}_i, \mathbf{x}) + b \right) \; . \tag{10}$$

In the case of linear-kernel SVMs, which are used throughout this paper, the discriminant function is simplified to

$$f(\mathbf{x}) = \mathrm{sgn}\left(\mathbf{w} \cdot \mathbf{x} + b \right) ; \qquad \mathbf{w} = \sum_{i=1}^{l} y_i \alpha_i \mathbf{x}_i \; . \tag{11}$$

Radius-margin bound: For the trained function, an error on a validation set, $\{(\tilde{\mathbf{x}}_i, \tilde{y}_i)\}_{1 \leq i \leq n}$, is calculated from

$$T = \frac{1}{2n} \sum_{i=1}^{n} |y_i - f(\mathbf{x}_i)| \; . \tag{12}$$

This is equivalent to true error and can represent the performance of SVMs if and only if $n \to \infty$. However, when the number of observations is small, the performance can be instead estimated by an upper bound on error, which is calculated on the training set. Many bounds are available, but one simple and informative bound is the radius-margin bound. The expectation for probability of error for an SVM trained with data set of size $l - 1$ is given as the inequality as the following

$$E p_{\mathrm{err}}^{l-1} \leq \frac{1}{l} E\left[\frac{R^2}{M^2} \right] = \frac{2}{l} E\left[R^2 W^2 \right] \; , \tag{13}$$

where M is the margin of the classifier, and R is a radius of hypersphere enclosing all training vectors.

The radius R is computed by optimizing the quadratic programming problem,

$$R^2(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}} \sum_{i=1}^{l} \beta_i \mathrm{K}(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i=1}^{l} \sum_{j=1}^{l} \beta_i \beta_j \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) \; , \tag{14}$$

subject to

$$\beta_i \geq 0 \quad \forall i \qquad \text{and} \qquad \sum_{i=1}^{l} \beta_i = 1 \,, \tag{15}$$

in a similar way to optimizing $W^2(\boldsymbol{\alpha})$ in Equation 8.

3. FEATURE SUBSET SELECTION

When the number of features gets larger, the curse of dimensionality[17] is introduced. It is related to slow convergence of the discriminant function in a high-dimensional space. It also causes the problem of classifier overfitting. Even though SVMs handle high-dimensional input vectors very well,[18] and circumvent both forms of the curse of dimensionality via the kernel trick,[19] it cannot avoid overfitting when the number of training vectors is extremely small, compared to the vector dimension.[20] This case is commonly encountered when dealing with small data sets, as is typical when dealing with biological samples. The problem is more severe when biological samples provide rich features. Fortunately, it seems that most features in observations are irrelevant or redundant, i.e. these features are irrelevant to the classification, and we expect that removing them could increase the classification performance. Thus, which features are involved and how many of them are appropriate to classification are issues to be resolved.

Feature ranking, in general, involves determining the relevance of features in an attendant classification problem. The feature selection method selects a subset of features, which leads to the best classification performance. *SVM induction is concerned with finding supporting vectors, whereas feature selection schemes are concerned with finding supporting features.* As we will discuss later, a good feature ranked list might not yield a good subset of features.

There exist two approaches to deal with feature selection, dubbed a filter approach and a wrapper approach. Both tackle the problem from different viewpoints and aim at different goals. The filter approach is data-driven. It ranks features based solely on given vectors and corresponding class labels (in the case of supervised learning) with the goal of a good ranked list. Thus the filter approach is valid on any classifier. We select a subset from the most relevant features to train the classifier in the hope that it will perform well. However, relevancy and optimality slightly differ, when classifiers are induced from data, not an underlying distribution. Kohavi et al.[21] gives examples when relevant features are not included in an optimal subset, and when irrelevant features are in an optimal subset. These examples conflict with the filter approach.

The wrapper approach, on the other hand, is algorithm-driven. It aims at the best classification performance rather than the best feature ranking. Several subsets of features, selected heuristically, are used to evaluate the performance of specific classifiers. A set that optimizes classification accuracy is likely to be the best. Some of the wrappers also implicitly produce the ranked list along the selection process. In comparison to the filter approach, the wrapper approach requires more complex and loaded processing, resulting from repetitive evaluation, but yields superior classification performance.

The methods regarded as the filters are, for example, correlation coefficients,[22] Fisher* criterion score,[23] RELIEF,[24] and RELIEF-F algorithm.[25] For the wrapper approach, if the number of features is small, we can construct combinatorial subsets of features, and evaluate them exhaustively to find the best combination. But the combinatorial method is not attractive, when several features are involved. Other wrappers use genetic algorithms[26] or recursive feature elimination (RFE)[20] to search for the subset that optimize the classifier. In this paper three methods are employed in parallel, including correlation coefficients, SVM-RFE, and SVM-RFE with scaling. All are presented in the following subsections.

*Historical note: R.A. Fisher worked at the University of Adelaide in the later part of his career. He died in Adelaide in 1963.

3.1. Correlation Coefficients

The correlation coefficients[22] and Fisher criterion score[23] are very similar and sometimes interchangeable. They measure a mean distance between two classes of a feature, weighted by distributions of the feature in both classes. The feature having the large distance, i.e. two classes are more separable by this feature, gets a high score. Both correlation coefficients and Fisher criterion scores are limited to linear problems.

Given that μ_k^+ and σ_k^+ are the mean and the standard deviation of feature k for all training vectors in class $+1$, and μ_k^- and σ_k^- are the mean and the standard deviation of feature k for all training vectors in class -1, the correlation coefficient is calculated from[22]

$$c_k = \frac{\mu_k^+ - \mu_k^-}{\sigma_k^+ + \sigma_k^-} \; . \tag{16}$$

The large positive c_k coefficient emphasizes a corresponding feature on class $+1$, and vice versa. Therefore, these correlation coefficients imply a ranking criterion of the features. As the coefficients have both positive and negative values, a few different criteria can be constructed from them. One possibility is to select the equal number of features from the top ranked positive and negative coefficients.[22] However, we simply use the absolute values of coefficients as the criterion, and select the features ranked at the top.[27]

An obvious drawback of this feature subset selection algorithm is due to the fact that redundant features are always assigned near ranks. As a result, they are not able to be discerned, while a minimum subset of features is needed.

3.2. SVM Recursive Feature Elimination (SVM-RFE)

The sensitivity analysis of the classification problem is related to computing the change of a cost function, when a single feature is removed. The cost function can be the norm of weight vector, $\|\mathbf{w}\|^2/2$, which is minimized in the training phase to yield the large margin. In case of the linear-kernel SVMs, when feature k is removed from the vectors, the cost function changes in direct proportion to w_k^2. Therefore, the feature with small magnitude of weight contributes less to the performance of SVMs. And we can remove the feature in accord with the corresponding weight's contribution.

When several features are removed at a time, a subset of remaining features becomes very suboptimal, as the sensitivity analysis is valid if and only if a feature is removed at once. Guyon et al.[20] proposed the recursive feature elimination algorithm. It alternates between training the classifier and removing a single or small chunk of feature, which have the smallest weights. This strategy allows re-optimizing the weight vector, which is used as the criterion. The method is a backward selection, as it constructs the list from the bottom. The recursive feature elimination algorithm for SVMs, or SVM-RFE, in case of the linear kernel is given by:

1. Initialize a feature mask $\mathbf{s} = [1, \ldots, 1]_m^T$, and a feature ranked list $\mathbf{r} = [\,]$
2. Train the classifier with a training set $\{(\mathbf{s} \circ \mathbf{x}_i, y_i)\}; \quad i = 1, \ldots, l$
3. Compute the weight vector $\mathbf{w} = \mathbf{s} \circ \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$
4. Compute the ranking criterion $c_k = w_k^2; \quad k = 1, \ldots, m$
5. Find feature with the smallest ranking criterion (the least significance) $\mathfrak{f} = \arg\min(\mathbf{c})$; $s_\mathfrak{f} \neq 0$
6. Remove that feature by set feature mask $s_\mathfrak{f} = 0$, and update the feature ranked list $\mathbf{r} = [\mathfrak{f}, \mathbf{r}]$
7. Return to step 2 until all features are ranked.

Note that the '$\circ$' operator is a Hadamard product or an element-wise multiplication.

3.3. SVM-RFE with Scaling

The feature subset selection using the radius-margin bound and gradient descent (RM-bound & gradient), proposed by Weston et al.[28] and Chapelle et al.,[29] aims to minimize the generalization error along with feature removal. It introduces a vector containing scaling factors, $\boldsymbol{\sigma} \in \mathbb{R}^k$, which maps the input vector to the scalable vector, $\mathbf{x} \mapsto (\boldsymbol{\sigma} \circ \mathbf{x})$. In the normal SVMs training procedure, only W^2 is minimized, and R^2 is fixed to the constant input vectors. But as shown by the radius-margin bound, the algorithms that minimize R^2W^2 can be expected to give better generalization performance.[19] The scaling factor helps further minimize R^2W^2 to its lowest possible value via the gradient descent search. When R^2W^2 reaches its saddle point, each scaling parameter implies the relevance of its corresponding feature to the performance. One or more irrelevant features are then removed, according to $\boldsymbol{\sigma}$ criterion, and the process starts over again until only the required number of features is left or the highest generalization performance is attained. It is clear that this method also takes the concept of recursive feature elimination to remove features. In this paper we propose the different criterion, which is a combination between the linear-kernel SVM-RFE and the RM-bound & gradient, to eliminate features. The process is itemized as follows:

1. Initialize the vector containing scaling factors $\boldsymbol{\sigma} = [1, \ldots, 1]^T_m$, and a feature ranked list $\mathbf{r} = [\,]$
2. Determine $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from a training set $\{(\boldsymbol{\sigma} \circ \mathbf{x}_i, y_i)\}; \quad i = 1, \ldots, l$
3. Update $\boldsymbol{\sigma}$ by a single local downhill gradient to minimize the radius-margin bound
4. Compute the radius-margin bound using the scaled vectors (also update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$), return to step 3 if local minimum of the bound is not reached
5. Compute the weight vector $\mathbf{w} = \boldsymbol{\sigma} \circ \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$
6. Compute the ranking criterion $c_k = w_k^2; \quad k = 1, \ldots, m$
7. Find feature with the smallest ranking criterion (the least significance) $\mathfrak{f} = \arg\min(\mathbf{c}); \; \sigma_{\mathfrak{f}} \neq 0$
8. Remove that feature by set the scaling parameter $\sigma_{\mathfrak{f}} = 0$, and the feature ranked list $\mathbf{r} = [\mathfrak{f}, \mathbf{r}]$
9. Return to step 2 until all features are ranked.

At step three, σ_k is updated with

$$\sigma_k^{\text{new}} = \sigma_k^{\text{old}} - \epsilon \frac{\partial R^2(\boldsymbol{\beta}, \boldsymbol{\sigma}) W^2(\boldsymbol{\alpha}, \boldsymbol{\sigma})}{\partial \sigma_k} \qquad \forall k \; . \tag{17}$$

A derivative of the bound in the case of linear-kernel SVMs is computed by

$$\frac{\partial R^2(\boldsymbol{\beta}, \boldsymbol{\sigma}) W^2(\boldsymbol{\alpha}, \boldsymbol{\sigma})}{\partial \sigma_k} = R^2(\boldsymbol{\beta}, \boldsymbol{\sigma}) \frac{\partial W^2(\boldsymbol{\alpha}, \boldsymbol{\sigma})}{\partial \sigma_k} + W^2(\boldsymbol{\alpha}, \boldsymbol{\sigma}) \frac{\partial R^2(\boldsymbol{\beta}, \boldsymbol{\sigma})}{\partial \sigma_k} \; , \tag{18}$$

$$\begin{aligned} \frac{\partial R^2(\boldsymbol{\beta}, \boldsymbol{\sigma})}{\partial \sigma_k} &= \sum_i \beta_i \frac{\partial \mathrm{K}(\boldsymbol{\sigma} \circ \mathbf{x}_i, \boldsymbol{\sigma} \circ \mathbf{x}_i)}{\partial \sigma_k} - \sum_{i,j} \beta_i \beta_j \frac{\partial \mathrm{K}(\boldsymbol{\sigma} \circ \mathbf{x}_i, \boldsymbol{\sigma} \circ \mathbf{x}_j)}{\partial \sigma_k} \\ &= \sum_i \beta_i \frac{\partial (\sigma_1^2 x_{i1}^2 + \ldots + \sigma_m^2 x_{im}^2)}{\partial \sigma_k} - \sum_{i,j} \beta_i \beta_j \frac{\partial (\sigma_1^2 x_{i1} x_{j1} + \ldots + \sigma_m^2 x_{im} x_{jm})}{\partial \sigma_k} \\ &= 2\sigma_k \left\{ \sum_i \beta_i x_{ik}^2 - \sum_{i,j} \beta_i \beta_j x_{ik} x_{jk} \right\} \; , \end{aligned} \tag{19}$$

$$\frac{\partial W^2(\boldsymbol{\alpha}, \boldsymbol{\sigma})}{\partial \sigma_k} = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \frac{\partial \mathrm{K}(\boldsymbol{\sigma} \circ \mathbf{x}_i, \boldsymbol{\sigma} \circ \mathbf{x}_j)}{\partial \sigma_k}$$

$$= -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \frac{\partial(\sigma_1^2 x_{i1}x_{j1} + \ldots + \sigma_m^2 x_{im}x_{jm})}{\partial\sigma_k}$$

$$= -\sigma_k \sum_{i,j}\alpha_i\alpha_j y_i y_j x_{ik}x_{jk} \ . \tag{20}$$

In fact, both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ implicitly depend on $\boldsymbol{\sigma}$, as they are optimal values for W^2 and R^2. But Chapelle et al.[29] proved that it is possible to differentiate W^2 and R^2 with respect to $\boldsymbol{\sigma}$, as if $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are constant.

4. T-RAY RESPONSES OF THE CELLS

THz-TDS was performed on human cells grown in culture. The cells were grown in transparent plastic flasks, which enabled spectroscopic measurement of the live cells in a transmission geometry. The cells considered were normal human bone (NHB) cells and human osteosarcoma (HOS) cells. Three identical flasks were used. The first two contained confluent HOS cells and confluent NHB cells in cell media. The third flask was used as a reference and contained only the cell media solution. Ferguson et al.[30] elaborated details on the cell preparation.

The measurement is performed using a standard scanning THz imaging system. A lock-in amplifier time constant of 10 ms was used. The laser was a regeneratively amplified Ti:sapphire laser producing 130 fs pulses with a 1 kHz repetition rate and an average power of 0.7 W. The THz emitter was a 2 mm thick $\langle 110 \rangle$ oriented ZnTe crystal and the THz beam was detected using electro-optic sampling in a 4 mm thick $\langle 110 \rangle$ ZnTe crystal.

A T-ray image was obtained providing spectroscopic information at 48 different locations for each flask. A waveform at each location contains 200 data points sampled every 0.067 ps with a total duration of 13.33 ps. Figure 1 shows averaged T-ray signals for the NHB cells and HOS cells. Seemingly, the presence of interstitial water in HOS cells causes signal weakening and broadening.

Since the data points in the time domain have an interval of 0.067 ps, the sampling frequency of the system equals 1/0.067 ps = 15 THz. A typical THz-TDS system provides a frequency resolution of 37.5 GHz. Points within the frequency response should also be separated by 37.5 GHz to coincide with the system. The spectroscopic signals are then padded with zeros with 15 THz/37.5 GHz = 400 points prior to Fourier transformation.

Impulse responses of the cells are obtained by deconvolving in the frequency domain the measured responses of the NHB or HOS cells with respect to an average of the system responses, which are measured on the empty flask. Note that noise in the system is low and insignificant, so we simply divide the measured responses by the system response. However, to avoid the effect of low SNR in low frequencies imposed by a standard THz-TDS system, the phase responses are unwrapped using the linearized phase unwrapping scheme.[31] The phase values at 0.0 to 0.1 THz are linearly extrapolated from the phase values at 0.1 to 0.4 THz.

After the deconvolution process to eliminate system dependency, the normalized magnitude and phase responses for two types of cells are obtained, as shown in Figure 2. A typical THz-TDS system provides a frequency resolution of 37.5 GHz, and our signals have a reliable bandwidth approximately from 0.15 to 1.6 THz. Therefore, there are up to 40 usable different frequencies for each observation. The number of features for classification is twice the number of frequencies by the fact that we have magnitude and phase information at each frequency.

In brief, 48 observations for NHB cells are labeled negative class, and other 48 for HOS cells labeled positive class. Each observation, or input vector, contains 80 features, 40 from magnitude responses and other 40 from phase responses. Obviously, the number of input vectors is low, compared with the number of available features. This addresses the overfitting problem, which will be removed by feature subset selections.

At first, all 96 vectors with completed features are used to train a linear-kernel SVM, and the results show that the SVM can find the hyperplane that separates this training set without error, or our data are linearly separable. Hence linear-kernel SVMs are valid and sufficient for our task. Satisfied by the linear kernel, we then randomly select half of the vectors from both classes for the actual training phase (including feature subset selection stage) and the remaining for the test phase. The results are reported in the next section.

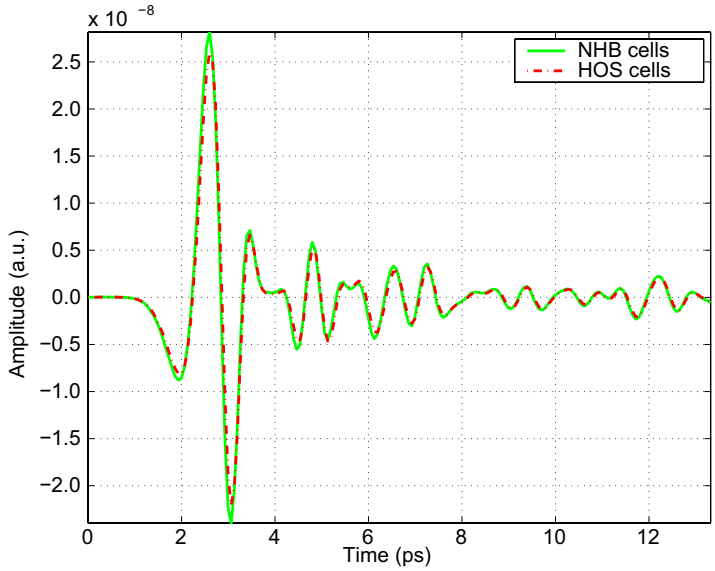

Figure 1. Averaged T-ray signals after transmission through the flasks. One flask contains cultured normal human bone cells (NHB), and the other contains cultured human osteosarcoma cells (HOS). The attenuation and broadening of the signal for the HOS cells result from increase in interstitial water content of cancerous cells, which T-rays are sensitive to. These signals are deconvolved by the reference signal prior to feature extraction in order to get rid of the system dependency.

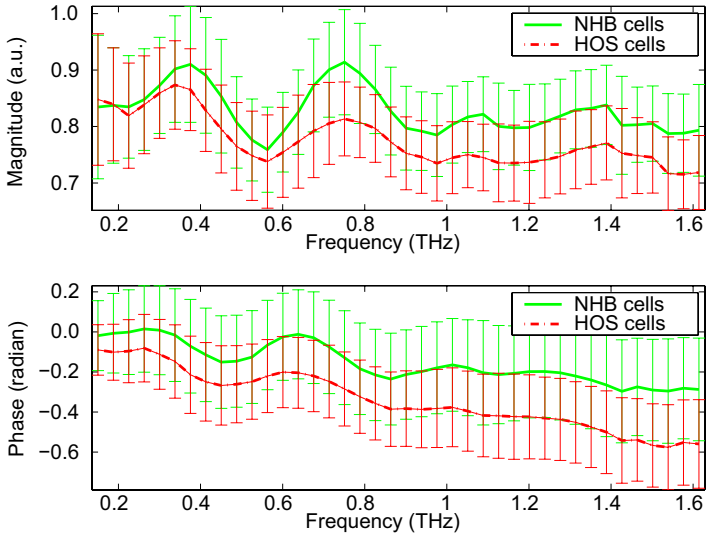

Figure 2. Averaged magnitude and phase responses of the NHB and HOS cells to T-ray impluses with bars indicating standard deviations. The magnitude and phase responses at these frequencies are used as input features for training and testing the SVM. The standard deviations show strong overlapping between the signals for NHB and HOS cells, which is hard to separate. However, in a preliminary run, a linear-kernel SVM can find weights that separate these signals without error. This means the data set is linearly separable.

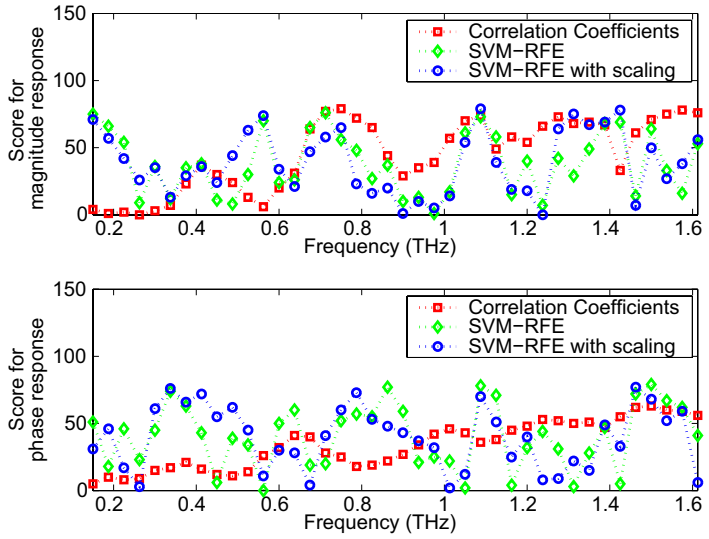

Figure 3. Scores of features extracted from T-ray magnitude and phase responses in Figure 2. The scores are obtained by running three distinct feature selection methods on the training set, containing 48 input vectors. Each score value indicates the ability of the corresponding feature to discriminate the osteosarcoma cells. The higher the score, the higher the discrimination ability. We can observe the similarity between scores obtained from SVM-RFE and SVM-RFE with scaling, but not correlation coefficients.

5. T-RAY FREQUENCY RANKED LIST

Figure 3 plots the scores of features ranked by three methods, the correlation coefficients, the SVM-RFE, and the SVM-RFE with scaling. The score is calculated for each feature in conjunction with the T-ray magnitude responses and the T-ray phase responses of the cells, and runs from zero to 79. Table 1 summarizes the first twenty features, that are most related to the osteosarcoma classification.

Most of the optimal features selected by the correlation coefficient method are from T-ray magnitude responses of the cells, while phase responses are left in lower ranks. In comparison, it can be seen that the list obtained from the correlation coefficient method is rather distinct from the list obtained from two wrappers, the SVM-RFE or SVM-RFE with scaling, as the two latter methods shuffle the features from both magnitude and phase responses equally.

The general agreement in ranking between two wrappers can be observed from Figure 3. In particular, 10 out of 20 most optimal features selected by both methods are matched, but placed in different positions (see Table 1). However, the most optimal feature of SVM-RFE is not in accordance with that of SVM-RFE with scaling.

In Ferguson et al.[30] the optimal T-ray frequencies, selected by a genetic algorithm from the same data set, are as follows: 0.22, 0.37, 1.12, 1.27, 1.34, and 1.57 THz. The paper suggested using both magnitude and phase responses at these frequencies for classification. In Table 1, the phase responses of these frequencies, 0.37, 1.12, and 1.57 THz, also appear in high positions, ranked by SVM-RFE and SVM-RFE with scaling. It is arguable that these three frequencies have high potential to discriminate the osteosarcoma cells.

Intuitively, the features extracted from adjacent frequencies are likely to be redundant. Further analysis of the score plot demonstrates ability of the methods to eliminate the redundant features. The correlation coefficient method gives a continuous score over the frequency range. But score plots from two wrappers clearly exhibit a sharp rise and fall. This implies that the wrappers are more suitable to our problem where many redundant features are present. The ability to eliminate redundant features of the wrappers is due to the recursive feature elimination strategy. How features are ranked depends on which subset of features is remaining. The filter, in contrast, determines relevance to a problem of each feature separately from the others.

Table 1. The 20 most optimal features, ranked by different methods. The normal font indicates the magnitude response at that T-ray frequency, and the boldface indicates the phase response. The magnitude responses are ranked at high positions by the correlation coefficient method, but distributed uniformly with phase responses in the case of SVM-RFE and SVM-RFE with scaling.

Rank	Correlation coefficients	SVM-RFE	SVM-RFE with scaling
1	0.7500	**1.5000**	1.0875
2	1.5750	**1.0875**	1.4250
3	0.7125	**0.8625**	**1.4625**
4	1.6125	0.7125	**0.3375**
5	1.5375	0.1500	1.3125
6	1.0875	**0.3375**	0.5625
7	1.2750	1.0875	**0.7875**
8	0.7875	**1.4625**	**0.4125**
9	1.5000	**1.1250**	0.1500
10	1.0500	0.5625	**1.0875**
11	1.3500	1.4250	1.3875
12	1.3125	1.3875	**1.5000**
13	1.3875	**1.5375**	1.3500
14	1.2375	0.1875	**0.3750**
15	0.8250	0.6750	0.7500
16	0.6750	1.5000	1.2750
17	**1.5000**	**0.3750**	0.5250
18	**1.4625**	**1.5750**	**0.4875**
19	1.4625	1.0500	**0.3000**
20	**1.5375**	**0.6375**	**0.7500**

6. OSTEOSARCOMA CLASSIFICATION RESULT

The T-ray frequency ranked list suggests a subset of the low-relevant features that should be removed from the input vectors. Following this list, we expect a better classification performance, or at least an unchanged error from SVMs. But the optimal number of features to be kept is yet in question.

At first the 48 input vectors, each of which contains the completed 80 features, 40 from the magnitude responses and the other 40 from the phase responses, are used to train the SVM. The trained SVM fails to classify 5 out of the 48 test vectors, i.e. a classification accuracy of 89.6%. A sequence of the features to be eliminated is according to the feature ranked list, given by the correlation coefficient method, the SVM-RFE, or the SVM-RFE with scaling. The irrelevant or redundant features, i.e. those with the lowest scores in the list, are removed first. Sets of the dimension-reduced vectors are used to train and test the SVM, until the two most optimal features, placed on top of the list, remain in the vectors.

As shown in Figure 4, the test error is not stable when we use the correlation coefficient scores to eliminate the features from 80 down to 27. And a rapid rise of the error apparently occurs, when the feature ranked at position 27 is removed and still 26 features remain, giving a classification accuracy of only 75% (12 misclassified). From that point the error is large and stops at an accuracy of 68.8% (15 misclassified), when only two features are used. On the other hand, SVMs trained with feature subsets selected by SVM-RFE or SVM-RFE with scaling provides very constant accuracy. We can reduce the input dimension to only 4 and 5 with accuracies of 89.6% (5 misclassified) and 87.5% (6 misclassified) in the case of SVM-RFE and SVM-RFE with scaling, respectively. At these points SVMs still give the nearly same accuracy as if all 80 features are used.

Nevertheless, the number of available test vectors is small, and thus we cannot observe a true generalization error nor a true performance of the trained SVMs from them. Restricted by the number of test vectors, we appeal to the error upper bound. In particular, the radius-margin bound, even loose, seems to be predictive,

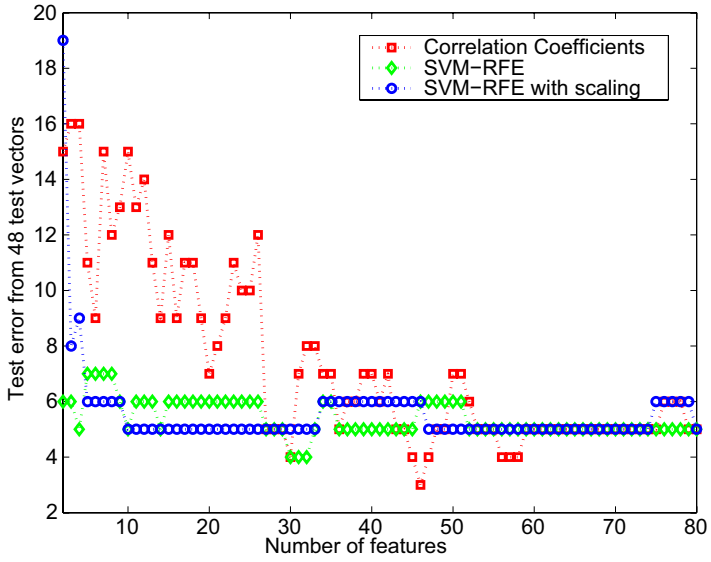

Figure 4. Number of classification errors from 48 test vectors produced by a series of trained SVMs. From the right, SVMs are trained with 80 features obtained form the magnitude and phase responses of the cells. Toward the left, each single feature is removed from the input vectors prior to SVM training and testing. The orders of features being removed are according to the features' relevant scores obtained from the correlation coefficients, the SVM-RFE, and the SVM-RFE with scaling methods. These scores are shown in Figure 3. The worst feature is repeatedly removed until only two optimal features remain in the vectors.

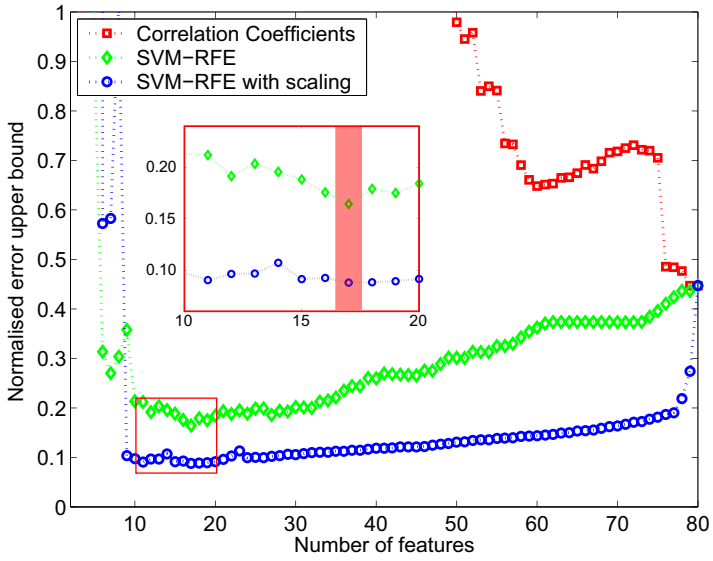

Figure 5. Error upper bound calculated according to Equation 13. As our available test set is so far from ideal that we cannot observe the true generalization error from Figure 4, the error upper bound serves to provide predictive information, when SVMs are trained and tested by vectors with reduced features. The change of the number of features in the horizontal axis is comparable to the change of that in Figure 4. The vertical axis is normalised by dividing by 4×10^3. Even though an estimated bound is loose, i.e. its magnitude is large and might not be close to the true generalization error, we can rely on its minimum point, which matches the minimum point of true generalization error. This point is useful for model selection. The band in the zoomed inset shows local minima of the bounds for the cases of feature selection according to SVM-RFE and SVM-RFE with scaling.

as it provides a local minimum, which coincides with that of the true error.[19] The parameter that minimizes this bound provides a good choice for the model.[32] The radius-margin bound is calculated at every step that a single feature is removed (see Figure 5). The similarity between the error in Figure 4 and the bound in Figure 5 substantiates the prediction capability of the bound. Note that small details on the bound are not so informative.

No local minimum appears in the bound, while subsets of features are selected using the correlation coefficient scores. And the bound is prone to increase every time that a single feature is removed. As a result, the correlation coefficient method fails to satisfy the imperative motivation to select a subset of features. It might be able to explore the relevant frequencies, and it might be able to lessen the computational cost, but it cannot improve or even retain the generalization performance for our task at all.

The bounds on error given by two wrapper methods have local minima at 17 features. They suggest that when the 63 worst features are removed from the vectors, the trained SVMs guarantee the best and improved classification performance. We can also infer that about 78 percent of the original features are irrelevant or redundant, and only the 17 optimal features ranked by SVM-RFE or SVM-RFE with scaling are sufficient for the osteosarcoma classification (see Table 1). In addition, the error upper bound given by SVM-RFE with scaling is lower than that given by normal SVM-RFE, because the scaling factor allows flexible adjustment to norms of the input vectors, which are related to the radius R in the radius-margin bound. At the optimal subset of 17 features, we trace back to the error plot in Figure 4, the SVM-RFE gives a classification accuracy on the test set of 87.5% (6 misclassified), and the SVM-RFE with scaling gives 89.6% (5 misclassified).

Another consequence apart from the accuracy improvement is observable when only 17 features are involved. Stated earlier, the curse of dimensionality causes the slow convergence of a discriminant function. Without feature reduction, the training process takes an average time of 0.52 sec[†] to converge to the optimal hyperplane. The time is reduced to 0.38 sec when the input vectors contain 17 features selected by the SVM-RFE, and reduced to 0.26 sec when the vectors contain the same number of features but selected by the SVM-RFE with scaling.

7. CONCLUSION

In this paper we use linear-kernel SVMs to discriminate between the NHB cells and HOS cells. As the number of available features from magnitude and phase responses are large, and most of the features are irrelevant or redundant, in order to avoid the overfitting problem the preprocess to select the optimal subset of features is invoked. This includes the correlation coefficient method, the SVM-RFE, and the SVM-RFE with scaling.

We show that the SVM-RFE and SVM-RFE with scaling are prefered to the correlation coefficient method, due to their ability to eliminate the redundant features, apart from the irrelevant features. The optimal subset of features selected by both SVM-RFE and SVM-RFE with scaling are comparable. In the case of SVM-RFE and SVM-RFE with scaling, we show that the appropriate number of features is indicated by plot of the radius-margin bound. When removing the irrelevant and redundant features to some points, the bound shows the lowest generalization error. Removing the features beyond this point raises the bound to its high value. Although an improvement of classification performance, when the optimal subset of features selected by SVM-RFE or SVM-RFE with scaling is employed, cannot be observed from our set of test vectors, the SVM retains its quality with no increment in the number of false classifications.

In addition to better generalization ability, the feature subset selection methods give other two advantages. One is the faster convergence of the classifier's optimization process. When only relevant features are concerned, the classifier training time is reduced as much as half of the original training time. The other merit is the exploration of the frequencies related to the osteosarcoma classification. The different reaction mechanisms of the cells to T-rays at the specific frequencies remains unresolved.

More training vectors are required to build a more precise ranked list and a more precise classifier. But, however, the biological samples both *in vitro* and *in vivo* are difficult to prepare and observe in large volumes. One possibility is to create virtual vectors.[33] By incorporating prior knowledge of vector variations, such as different thicknesses of cell layer or different geometries of measurement, the virtual vectors are reproducible

[†]The computational time is based on C code running on a Pentium 1.70 GHz.

from available signals with assistance of the transfer function model for T-ray.[34] These virtual vectors are valid for training the classifier.

SVMs offer another method, in addition to feature reduction scheme, in order to avoid overfitting. Variants of the machines, or soft-margin SVMs,[35] provide an error penalty term, C, which controls the tradeoff between complexity and the number of misclassified samples. As the parameter gets smaller, the complexity of SVMs reduces to allow more training error. This prevents the induced classifiers from overfitting to a particular training set. The parameter is freely adjustable, but algorithmic fine-tuning is required to benefit from the best generalization performance.[29,36]

Some other features are available from T-ray signals, for example, wavelet coefficients. Wavelet transforms are preferred to Fourier transforms in representing T-ray signals, as wavelets are localized in the scale-space domain, and can closely resemble[‡] the T-ray waveform.[37] An appropriate representation could probably lead to better classification performance.

One important issue that remains undiscussed is the cost of decision. Patients with osteosarcoma have survival rate of less than 20%, unless they receive proper therapy.[11] Since the decision concerns such a life-threatening disease, it is preferable to weight the decision to false positive rather than false negative. This allows medical investigation of ambiguous patients. In such cases, Morik et al.[38] and Veropoulos et al.[39] propose an unbalanced cost factor for SVMs.

ACKNOWLEDGMENTS

Useful discussions with Hong-Gunn Chew and Brian W.-H. Ng, The University of Adelaide, are gratefully acknowledged. Funding from the Australian Research Council (ARC) and the Sir Ross and Sir Keith Smith Fund is gratefully acknowledged.

REFERENCES

1. B. Ferguson, S. Wang, D. Gray, D. Abbott, and X.-C. Zhang, "Toward functional 3D T-ray imaging," *Physics in Medicine and Biology (IOP)* **47**, pp. 3735–3742, 2002.
2. S. Mickan, A. Menikh, H. Liu, C. Mannella, R. MacColl, D. Abbott, J. Munch, and X.-C. Zhang, "Label-free bioaffinity detection using terahertz technology," *Physics in Medicine and Biology (IOP)* **47**, pp. 3789–3795, 2002.
3. G. Walker, E. Berry, S. Smye, N. Zinov'ev, A. Fitzgerald, R. Miles, M. Chamberlain, and M. Smith, "Two methods for modelling the propagation of terahertz radiation in a layered structure," *Journal of Biological Physics* **29**, pp. 141–148, 2003.
4. S. Smye, J. Chamberlain, A. Fitzgerald, and E. Berry, "The interaction between terahertz radiation and biological tissue," *Physics in Medicine and Biology* **46**, pp. R101–R112, 2001.
5. D. Abbott, "Direction in terahertz technology," *IEEE GaAs Digest* , pp. 263–266, 2000.
6. Q. Wu and X.-C. Zhang, "Free-space electro-optic sampling of terahertz beams," *Applied Physics Letters* **67**(24), pp. 3523–3525, 1995.
7. R. Woodward, B. Cole, V. Wallance, R. Pye, D. Arnone, E. Linfield, and M. Pepper, "Terahertz pulse imaging in reflection geometry of human skin cancer and skin tissue," *Physics in Medicine and Biology* **47**, pp. 3853–3863, 2002.
8. R. Woodward, V. Wallance, R. Pye, B. Cole, D. Arnone, E. Linfield, and M. Pepper, "Terahertz pulse imaging of *ex vivo* basal cell carcinoma," *Journal of Investigative Dermatology* **120**(1), pp. 72–78, 2003.
9. B. Ferguson, S. Wang, D. Gray, D. Abbott, and X.-C. Zhang, "Identification of biological tissue using chirped probe THz imaging," *Microelectronics Journal* **33**(12), pp. 1043–1051, 2002.
10. T. Löffler, K. Siebert, S. Czasch, T. Bauer, and H. Roskos, "Visualization and classification in biomedical terahertz pulsed imaging," *Physics in Medicine and Biology* **47**, pp. 3847–3852, 2002.

[‡]The significance of close resemblance is that a signal can then be decomposed into a few wavelet coefficients. Thus the original signal power is spread over a few coefficients and is not overly *diluted*.

11. A. Yasko and W. Chow, "Bone sarcomas," in *Cancer Management: A Multidisciplinary Approach*, R. Pazdur, W. Hoskins, L. Coia, and L. Wagman, eds., ch. 24, pp. 573–584, F. A. Davis Company, 9th ed., 2005.
12. B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifier," in *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, pp. 144–152, 1992.
13. V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York, 1998.
14. K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithm," *IEEE Transactions on Neural Networks* **12**(2), pp. 181–202, 2001.
15. E. Osuna, R. Freund, and F. Girosi, "An improved training algorithm for support vector machines," in *Proceedings of the 1997 Neural Networks for Signal Processing*, pp. 276–285, 1997.
16. J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, eds., pp. 185–208, MIT Press, Cambridge, MA, USA, 1998.
17. R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, New Jersey, 1961.
18. E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, pp. 130–136, 1997.
19. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery* **2**, pp. 121–167, 1998.
20. I. Guyon, J. Weston, and S. Barnhill, "Gene selection for cancer classification using support vector machines," *Machine Learning* **46**, pp. 389–422, 2002.
21. R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence* **97**, pp. 273–324, 1997.
22. T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science* **286**, pp. 531–537, 1999.
23. P. Pavlidis, J. Weston, J. Cai, and W. Grundy, "Gene functional classification from heterogeneous data," in *Proceedings of the Fifth Annual International Conference on Computational Biology*, pp. 249–255, 2001.
24. K. Kira and L. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning*, pp. 249–256, 1992.
25. I. Kononenko and S. Hong, "Attribute selection for modeling," *Future Generation Computer Systems* **13**(2-3), pp. 181–195, 1997.
26. J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems and Their Applications* **13**(2), pp. 44–49, 1998.
27. T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics* **16**(10), pp. 906–914, 2000.
28. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, eds., pp. 668–674, 2000.
29. O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning* **46**, pp. 131–159, 2002.
30. B. Ferguson, H. Liu, S. Hay, D. Findlay, X.-C. Zhang, and D. Abbott, "In vitro osteosarcoma biosensing using THz time domain spectroscopy," in *Proceedings of SPIE BioMEMS and Nanotechnology*, D. V. Nicolau, ed., **5275**, pp. 304–316, 2004.
31. W. Withayachumnankul, B. Ferguson, T. Rainsford, S. P. Mickan, and D. Abbott, "Material parameter extraction for terahertz time-domain spectroscopy using fixed-point iteration," in *Proceedings of SPIE Photonic Materials, Devices, and Applications*, G. Badenes, ed., **5840**, pp. 221–231, 2005.
32. O. Chapelle and V. Vapnik, "Model selection for support vector machines," in *Advances in Neural Information Processing Systems 12*, S. Solla, T. Leen, and K.-R. Müller, eds., 1999.
33. D. Decoste and B. Schölkopf, "Traning invariant support vector machines," *Machine Learning* **46**, pp. 161–190, 2002.

34. T. Dorney, R. Baraniuk, and D. Mittleman, "Material parameter estimation with terahertz time-domain spectroscopy," *Journal of the Optical Society of America A* **18**(7), pp. 1562–1571, 2001.
35. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning* **20**, pp. 273–297, 1995.
36. K. Duan, S. Keerthi, and A.-N. Poo, "Evaluation of simple performance measures for tuning SVM hyperparameters," *Neurocomputing* **51**, pp. 41–59, 2003.
37. D. Mittleman, R. Jacobsen, R. Neelamani, R. Baraniuk, and M. Nuss, "Gas sensing using terahertz time-domain spectroscopy," *Applied Physics B: Lasers and Optics* **67**(3), pp. 379–390, 1998.
38. K. Morik, P. Brockhausen, and T. Joachims, "Combining statistical learning with a knowledge-based approach—a case study in intensive care monitoring," in *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 268–277, 1999.
39. K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI99)*, pp. 55–60, 1999.