SVM for FT-MIR prostate cancer classification: An alternative to the traditional methods

Laurinda F. S. Siqueira[1,2], Camilo L. M. Morais[3], Raimundo F. Araújo Júnior[4], Aurigena Antunes de Araújo[5], Kássio M. G. Lima[1*]

[1]Biological Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande of Norte, Natal 59072-970, Brazil

[2]Federal Institute of Education, Science and Technology of Maranhão, São Luís 65075-441, Brazil

[3]School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, United Kingdom

[4]Department of Morphology, Post graduation programme in Health Science / Post graduation programme in Structural and Functional Biology, Federal University of Rio Grande do Norte, Natal 59072-970, RN, Brazil5

[5]Department of Biophysics and Pharmacology, Post graduation programme in Public Health / Post graduation programme in Pharmaceutical Science, Federal University of Rio Grande do Norte, Natal 59072-970, RN, Brazil

***Corresponding author:** Prof. Dr. Kássio M.G. Lima, Biological Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil. Email: kassiolima@gmail.com; Tel.: +55 84 3342 2323.

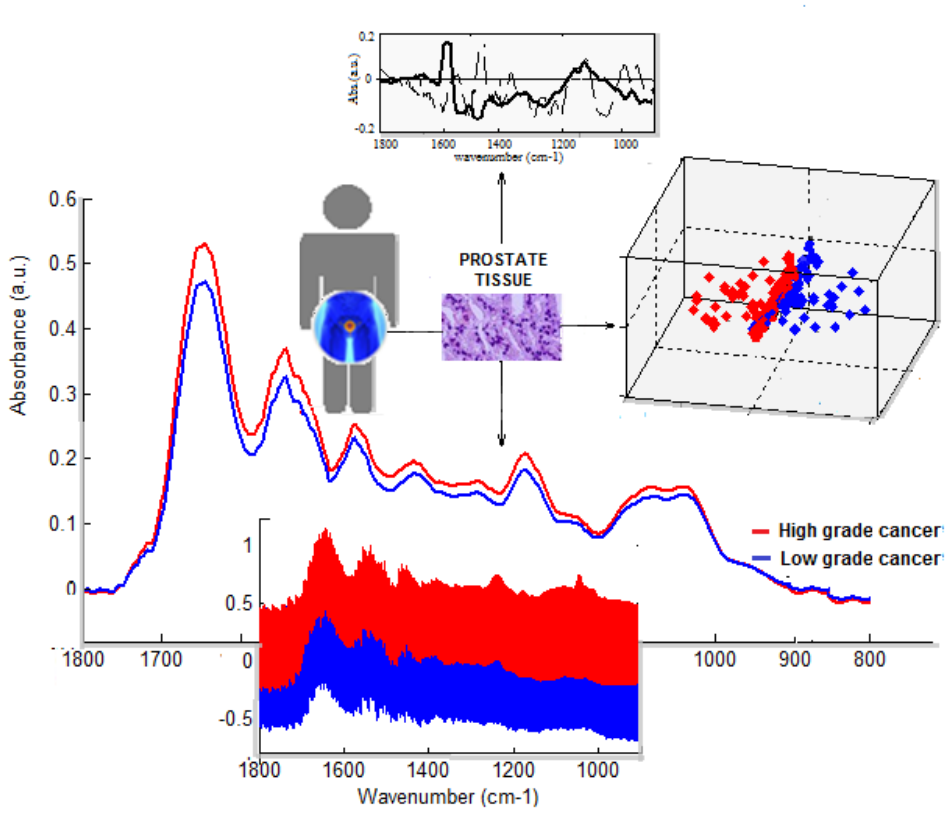
Short title: SVM for FT-MIR prostate cancer classification

Table of Content Abstract:

SVM for FT-MIR prostate cancer classification: An alternative to the traditional methods

Laurinda F. S. Siqueira[1,2], Camilo L. M. Morais[3], Kássio M. G. Lima[1*]

In this paper, principal component analysis (PCA), successive projections algorithm (SPA) and genetic algorithm (GA) followed by support vector machines (SVM), combined with Fourier-transform mid-infrared (FT-MIR) spectroscopy were presented as complementary or alternatives tools to the traditional methods for prostate cancer screening and classification.



Summary

In this paper, principal component analysis (PCA), successive projections algorithm (SPA) and genetic algorithm (GA) followed by support vector machines (SVM), combined with Fourier-transform mid-infrared (FT-MIR) spectroscopy were presented as complementary or alternatives tools to the traditional methods for prostate cancer screening and classification. These approaches were applied to analyze tissue samples; and their performances were compared within dependent SVM models and with traditional methods of diagnosis, according to class separation interpretability, time consumption and figures of merit. The results shown that variable reduction and selection methods followed by SVM can reduce drawbacks of independent SVM analysis. The potential biomarkers indicated by PCA-SVM, SPA-SVM and GA-SVM were amide I, II and III; as well as protein regions (1,400-1,585 cm^{-1}), followed by DNA/RNA (O–P–O symmetric stretch) (1,080 cm^{-1}) and DNA (O–P–O asymmetric stretch) (1,230 cm^{-1}) regions. GA-SVM was the best classification approach, with higher sensitivity (100%) and specificity (80%), particularly in early stages, being better than traditional methods of diagnosis.

Keywords: SVM; FT-MIR; tissue; prostate cancer

1. Introduction

Currently, prostate cancer recognition follows some phases. First of all, the proctologist evaluates the prostate by Digital Rectal Examination (DRE), which allows palpation of only 40-50% of the prostate, and it is mainly affected by intra and inter-observer variability, resulting in DRE sensitivity and specificity of about 21-37% and 71-91%, respectively.[1-5] Concomitantly, a measurement of serum Prostatic Specific Antigen (PSA) levels is performed which have about 21% and 64% sensitivity and specificity in earliest stages, respectively; and about 32% and 93% sensitivity and specificity in high grade prostate cancers, respectively. When combined with DRE, sensitivity and specificity increase to 51-68% and 92-94%, respectively;[1-8] however, this measurement can be affected by other factors beyond prostate cancer, such as ejaculation, bacterial prostatitis, biopsy and acute urinary retention, which may elevate PSA levels.[3,5-8]

According to the results of these exams, a biopsy coupled with an anatomopathological examination can be indicated in order to identify cancer stage. Trans-rectal ultrasound (TRUS)-guided biopsy is currently the gold standard, characterized by 39-52% sensitivity and 81-82% specificity.[9] The anatomopathological examination by Gleason score system classifies biopsy samples according to tumor aggressiveness and provides a prognostic idea. It defines Gleason 1 (best differentiation and most favorable prognosis) to Gleason 5 (least differentiation and poor prognosis).[10-12] The sum of the primary (most architectural atypia) and secondary (least architectural atypia) patterns leads to the Gleason score (e.g., Gleason 3 + Gleason 4 is equivalent to a Gleason score 3+4 = 7).[12] In 2014, a new Gleason grading system was introduced incorporating the previous Gleason score system (introduced in the 1960s)[13] in five

different categories.[14] Gleason pattern 1 includes Gleason score 6 (3+3) and represents well-formed and uniform distributed glands. Gleason pattern 2 includes Gleason score 7 (3+4) and represents predominantly well-formed glands with minor poorly-formed/fused/cribriform glands. Gleason pattern 3 includes Gleason score 7 (4+3) and represents predominantly poorly-formed/fused/cribriform glands with minor well-formed glands. Gleason pattern 4 includes Gleason score 8 (4+4, 3+5, 5+3) and represents poorly-formed/fused/cribriform glands. Gleason pattern 5 includes Gleason scores 9 and 10 (4+5, 5+4, 5+5) and represents necrosis, cords, sheets, solid nests and single cells with or without poorly-formed/fused/cribriform glands.[12] Gleason scores 2–5 are no longer assigned.[12] Figure 1 illustrates the new Gleason grading system according different types of glands.

[Insert Figure 1 here]

This system presents about 22-29% sensitivity and 81% specificity in earliest stages, respectively; and 30% sensitivity and 88-97% specificity in high grade prostate cancers, respectively.[12,15] However, biopsy-based detection as above have drawbacks, such as delays in providing the diagnostic results, sample heterogeneity, difficult preparation and time consuming procedure, harmful to the organs and the probability of spreading the cancer; in addition to visual criteria of pattern recognition which is operator-dependent and subject to intra- and inter-observer variability.[16,17]

Fourier-transform mid-infrared (FT-MIR) spectroscopy is an important technique to identify structural alterations of cellular molecules based on chemical bonds, as well as to identify spectral biomarkers.[18,19] The spectral region from 900 to 1,800 cm^{-1}, namely the ‘fingerprint region’ and the major biochemical information area, presents potential biomarkers for biochemical alterations promoted by cancer when

compared to samples of the control class. Potential biomarkers are considered as: a protein region with bands of amide I (~1,650 cm^{-1}), amide II (~1,550 cm^{-1}), methyl groups of lipids and proteins (~1,400 cm^{-1}), and amide III (~1,260 cm^{-1}); a DNA/RNA region with asymmetric phosphate stretching vibrations (~$\nu_{as}PO_2^-$; ~1,225 cm^{-1}), symmetric phosphate stretching vibrations ($\nu_sPO_2^-$; ~1,080 cm^{-1}); C–O groups of carbohydrates (~1,155 cm^{-1}); glycogen (~1,030 cm^{-1}) and protein phosphorylation (~970 cm^{-1}).[20-23] Furthermore, FT-MIR is a non-invasive and non-destructive technique, which presents objectivity, low operational cost and versatility. It allows qualitative and quantitative applications, and involves quick and easy procedures. In cancer studies, FT-MIR is an important technique for screening, being mostly performed in transmission or attenuated total reflectance (ATR) modes.[23,24]

Multivariate classification is used in several studies involving discrimination of spectral data of biological samples, particularly in cancer studies. Notably, this type of application is highlighted in the field of Discriminant Analysis, such as using linear and quadratic discriminant analysis (namely, LDA and QDA); and, in a minor scale, with algorithms considering data multidimensionality and non-linear boundaries, such as support vector machines (SVM).[23,24] Multivariate classification involves a few steps such as preprocessing, sample selection, variable reduction and/or selection methods coupled to classification techniques, and validation by statistical evaluation in terms of figures of merit and comparison to reference samples.

Variable selection techniques are used in order to reduce data size, reduce collinearity problems, and provide biomarker identification. They are often combined with classification techniques where the selected variables (e.g., wavenumbers) are used as input variables for classification. Two common techniques of variable selection are

the successive projections algorithm (SPA)[25] and genetic algorithm (GA)[26]. SPA is a progressive variable selection technique that uses the minimization of the original data multicollinearity as variable selection criterion.[27] As advantages, SPA has a deterministic nature and minor parameter optimization, however it is very time consuming. On the other hand, GA reproduces the Darwin's theory of evolution in a computational sense, where natural selection is applied for selecting the best set of variables that fits certain criterion.[27] This is performed after series of binary crossovers and mutations followed by the calculation of a fitness function, which for this paper was based on a Mahalanobis distance calculation. Although GA has a non-deterministic nature, its biggest advantages are the relatively low-computational cost compared to SPA and the reduction of the data collinearity. In addition, GA-based classification techniques often provide excellent classification performance when compared to other classification methods.[23] GA and SPA span the methodological range of variable selection methods used in multivariate classification problems which have been published in the literature.

The application of these chemometric tools aims to reduce, select and classify useful information, given the large and complex spectral information of biological samples (such as tissues, cells and biofluids) and their normal and abnormal biochemical processes. Together, multivariate classification and FT-MIR tend to be a potential method for investigating, diagnosing, categorizing and monitoring prostate cancer, as opposed to standard methods of detection and classification. In addition, these methodologies are not hindered by operator-dependence and intra- and inter-observer variability, difficult sample preparation and time-consuming procedures. For example, there are many applications using GA-based techniques for analyzing FT-MIR

biological datasets, such as characterization and classification of astrocytic glioma tissue in brain tumors;[24] diagnostic of basal cell carcinoma using blood sample analysis;[17] identification of intraepithelial lesion of cervix;[28] identification of low-grades cases of cervical cancer and possible biomarkers of disease progression;[29] diagnosis of Alzheimer's disease using blood samples;[30] and microbiological investigations, such as virus[31,32] and fungi[33,34] detections.

This paper aims to apply variable reduction and selection techniques combined with SVM in FT-MIR data from tissues samples in order to classify and detect spectral differences between early and advanced stages of prostate cancer. Combinations of principal component analysis (PCA), SPA and GA with SVM will be evaluated varying many different SVM parameters in order to study the effect of SVM optimization in the classification result.

2. Material and Methods

2.1. Tissue Collection

Prostate tissue sections were obtained from the Pathology Department of the Federal University of Rio Grande of Norte (UFRN/Brazil). Prostate tissue sections were formalin-fixed, dehydrated and paraffin-embedded (FFPE) in pathology blocks (n = 45), previously classified according to Gleason grading system by pathologists. The tissue samples were distributed according to the Gleason pattern to form 3 categories: Gleason pattern 2 (n = 23), Gleason pattern 3 (n = 15) and Gleason pattern 4 (n = 7). Tissue sections (5 μm-thick) were floated onto ZnSe slides (Bruker Optics Ltd., Coventry, UK). These were de-waxed by serial immersion in fresh xylenes baths for 5 min, and then washed and cleaned in an absolute ethanol bath for another 5 min.[23] The resulting

samples were allowed to air-dry and then placed in a desiccator until analysis. In our study, we chose to analyze Low grade (Gleason pattern 2, n = 23) and High grade (Gleason pattern 3 and Gleason pattern 4, n = 22) categorization, in order to work with concepts of early and advanced stages from a screening perspective.

2.2. FT-MIR Spectroscopy

A total of 100 FT-MIR spectra were collected per tissue in transmission mode using a Bruker Lumos FTIR microscope spectrometer (Bruker Optics Ltd., Coventry, UK). FTIR spectra represented an average of 32 scans in the mid-infrared (MIR) range of 600–4,000 cm^{-1} with a spectral resolution of 8 cm^{-1}. Spectra were acquired with a new background taken for every new sample; these were converted into absorbance by Bruker OPUS software.

2.3. Computational Analysis

Spectral data importing, preprocessing, and construction of multivariate classification models were performed within MATLAB R2012b environment (Math Works Inc., Natick, MA, USA) using PLS Toolbox 7.8 (Eigenvector Research, Inc., Wenatchee, WA, USA) and lab-made algorithms. FT-MIR spectra were cut to include wavenumbers between 800 and 1,800 cm^{-1}, the area associated with the biological spectral fingerprint. In the resulting dataset, extended multiplicative scatter correction (EMSC) was performed to correct baseline,[36,37] Savitzky-Golay smoothing (window of 15 points) to correct background noise,[38,39] and normalization to amide I peak (~1,650 cm^{-1}) to eliminate distortions. For application of each analytical model, spectral data were divided into training (60%), validation (20%) and prediction (20%) sets by applying the classic Kennard-Stone (KS) uniform sampling algorithm.[40] The training

and validation datasets were used in the modeling procedures, whereas the prediction dataset was only used for the final classification evaluation.

Application of linear, quadratic, radial basis function (RBF) and 3rd order polynomial SVMs was evaluated according to different kernel parameters (bias, correct classification of training and test sets, and *C* parameter). The same set of values for the *C* parameter was considered for each kernel, which controls the trade-off between training error and margin. *C* ranged from 0.01 to 1000. In addition, all combinations between *C* and σ parameters were trained for RBF-SVM, where σ ranged from 0.01 to 1000, totaling about 60100 RBF-SVM models for each dataset.

To deal with high dimensional data, the use of variable reduction and selection methods followed by SVM can maximize the predictive performance of the models, remove collinearity issues and reduce over-fitting, mainly when the number of samples in the training set is small. For this, the input space was transformed into feature space by means of the radial basis function (RBF) kernel in the PCA-SVM, SPA-SVM and GA-SVM models. These models were performed according to *C*-support vector classification (*C*-SVC), where *C* is the cost of misclassification. *C* regulates the balance between training errors and model complexity. In our models, both *C* and σ parameters were automatically optimized.

Variable reduction by PCA-SVM was employed after principal component (PC) number optimization based on classification rates of training and validation sets. It was applied to classify the samples according to the PCs with largest explained variance, representing as much variability as possible for classification.[41,42] Variable selection by SPA-SVM picks the best variables according to its biggest vector projection in the orthogonal space;[43] while GA-SVM selects the best variables based on the stochastic

sampling method.[17,29,44] The variable selection routine by GA-SVM was carried out utilizing 40 generations containing 80 chromosomes each. Five independent GA-SVM models with different random initial populations were performed and only the best individuals were kept.

2.4. Figures of Merit

For model validation, comparison and evaluation, sensitivity, specificity, positive and negative predictive values, Youden's index, and positive and negative likelihood ratios were analyzed. (1) Sensitivity (SENS) is the confidence that a positive result for a sample is a true positive in disease; (2) specificity (SPEC) is the confidence that a negative result for a sample is truly negative; (3) positive predictive value (PPV) is the proportion of test positives which are true positives; (4) negative predictive value (NPV) is the proportion of test negatives which are true negatives; (5) Youden's index (YOU) evaluates the classifier's ability to avoid failure, and also biomarker identification capacity; (6) positive likelihood ratio (LR+) represents the ratio between the probability of predicting a sample as truly positive, and the probability of predicting a sample as negative; and (7) the negative likelihood ratio (LR-) represents the ratio between the probability of predicting a sample as truly negative, and the probability of predicting a sample as positive.[24,45,46]

All this means that the best models must have high sensitivity and specificity to be accurate in class separation. YOU must be close to 100 to prove its classification capacity and biomarker identification. NNP and NPV must also be high for affirming or negating the group segregation. LR+ must be high, while LR- must be low, which

provides an intuitive feeling that the models rule the classification. The figures of merit can be performed by equations summarized in Table 1.

[Insert Table 1 here]

In addition, correct classification (CC%) was calculated in the training and test sets. Training set CC% involves applying the model in the same samples used to build and to optimize the model, while test set CC% is used to test the model's classification accuracy with external samples. Generally, the first tends to be higher (but closer) than the second; this means that the models are well-balanced and without presence of overfitting.[47] In addition, a comparison between single SVM models and SVM models using variable reduction and selection methods was performed by assessing these classification rates.

3. Results

3.1. Preprocessing

FT-MIR spectral dataset derived from Low and High grade categorization for prostate cancer is shown in Figure 2. Spectral raw data (Figure 2A) was cut at 1,800 - 800 cm^{-1} to emphasize the fingerprint region. EMSC, Savitzky-Golay smoothing (window of 15 points) and normalization to amide I peak (~1,650 cm^{-1}) were performed (Figure 2B).

[Insert Figure 2 here]

3.2. SVM Models

3.2.1. RBF-SVM

Bias and errors in training and test sets are shown in Figure 3, varying the values of σ and C parameters in RBF-SVM models derived from Low and High grade

categorization for prostate cancer. It is observed that bias values can be described in three ways for σ and C variation: (1) For $\sigma \leq 1$, bias remains approximately constant and had low values, independently of the C values; (2) For $1 \leq \sigma \leq 10$ and for $\sigma \geq 15$ and $C \leq 15$, bias increases; and (3) for $\sigma \geq 10$ and $C \geq 10$, bias decreases (Figure 3A). Errors in the training set were lower; however, it increased in $\sigma \geq 15$ and decreased to zero in $\sigma \leq 15$, independently of the C values. In $\sigma \geq 25$, High grade had more correct classification in training set (86%) than Low grade (78%) (Figure 3B). An elevated test error can be observed in $\sigma \geq 15$ for Low grade, while the same occurs in $\sigma \leq 1$ for the High grade, independently of the C values (Figure 3C). Maximum correct classification of test set was 60% for both Low (in $C \geq 5$) and for High grade (in $\sigma = 10$, $C \geq 1$). For most RBF-SVMs, the number of support vectors used in the classification was high, and around the same number of training samples at a minor scale for those with $\sigma \geq 15$ and $C \geq 10$ (Table 2).

[Insert Figure 3 here]

[Insert Table 2 here]

3.2.2. Linear, Quadratic and Polynomial SVMs

Bias and errors in training and test sets using linear, quadratic and 3rd order polynomial SVM models derived from the Low and High grade categorization for prostate cancer are illustrated in Figure 4. Linear SVM had larger bias values than quadratic and 3rd order polynomial SVMs, independently of the *C* values (Figure 4A). Errors in the training set were small (≤ 0.1) for Low and High grades, where 3rd order polynomial SVM had the minor correct classification of training set (90%) for both grades, independently of the *C* values (Figure 4B). Low grade had larger error in the test

set than High grade (Figure 4C). Quadratic SVM was the best model for Low grade classification, correctly classifying 40% of the test set. Linear and 3rd order polynomial SVMs correctly classified 80-100% of the High grade test set. The number of support vectors used in the classification is presented in Table 2.

[Insert Figure 4 here]

3.3. Variable Reduction and Selection Methods coupled to SVM

For PCA-SVM, the influence of the PC number in the correct classification of training and test sets from Low and High grade categorization for prostate cancer is represented in Figure 5. The best classification rates (80%) for both Low and High grades in both training and test sets were found using 10 PCs. Loadings and score plots derived from PCA-SVM are displayed in Figure 6A-B. Loadings plot identified the following distinguishing wavenumbers: 975; 1,080; 1,155; 1,230; 1,270; 1,370; 1,415; 1,465; 1,555; and 1,575 cm^{-1}. PCA-SVM classification used 10 support vectors and had bias around -0.156.

[Insert Figure 5 here]

[Insert Figure 6 here]

Loadings and score plots derived from SPA-SVM are presented in Figure 6C-D, and from GA-SVM in Figure 6E-F. SPA-SVM approach used twenty-four wavenumbers: 960; 1,000; 1,027; 1,081; 1,115; 1,134; 1,151; 1,169; 1,231; 1,296; 1,325; 1,347; 1,357; 1,376; 1,389; 1,402; 1,450; 1,468; 1,488; 1,506; 1,559; 1,595; 1,620; and 1,650 cm^{-1} (Figure 6C). SPA-SVM classification used 21 support vectors and had -0.098 bias. GA-SVM model generated the best classification (Figure 6E) using twenty selected variables. These were: 950; 1,012; 1,086; 1,226; 1,232; 1,242; 1,249;

1,268; 1,276; 1,297; 1,306; 1,330; 1,370; 1,371; 1,376; 1,399; 1,519; 1,552; 1,630 and 1,680 cm^{-1}. GA-SVM classification used 15 support vectors.

3.4. Figures of Merit

Figures of merit for RBF-SVMs, linear-SVM, quadratic-SVM, 3rd order polynomial-SVM, PCA-SVM, SPA-SVM and GA-SVM models for FT-MIR classification of prostate cancer are listed in Table 3 and Table 4.

[Insert Table 3 here]

[Insert Table 4 here]

In Table 3, SVM models had sensitivity and specificity values varying around 20-100% and 0-80% for Low and High grade, respectively. Linear-SVM had opposite sensitivity (20% *vs.* 100%) and specificity (80% *vs.* 0%) values for Low and High grades, respectively; the same occurred with sensitivity (40% *vs.* 80%) and specificity (60% *vs.* 20%) values for Low and High grades, respectively, using quadratic-SVM. 3rd order polynomial-SVM had 0% of sensitivity and specificity for both grades. RBF-SVM ($\sigma = 10$, $C \geq 5$) had 60% and 40% of sensitivity and specificity respectively for both grades.

In Table 4, variable reduction and selection methods followed by SVM presented sensitivity of 80-100% and specificity of 75-80% for the Low grade category, while it showed sensitivity of 67-80% and specificity of 71-80% for High grade. GA-SVM correctly classified around 100% of training and test sets for both grades of prostate cancer, and also had the best figures of merit in comparison to PCA-SVM and SPA-SVM. On the other hand, PCA-SVM was the second best approach, with slightly better classification rates and figures of merit than SPA-SVM.

4. Discussion

This work aimed to apply variable reduction and selection techniques followed by SVM in FT-MIR data from human tissue, in order to classify and detect spectral differences between early and advanced stages of prostate cancer. Prostate tissues taken from formalin-fixed dehydrated and paraffin-embedded (FFPE) pathology blocks were used, previously staged by pathologists in Gleason pattern 1 (n = 23, Gleason score 6), Gleason pattern 2 (n = 15, Gleason score 7 [3+4]) and Gleason pattern 4 (n = 7, Gleason score 8). The samples were stored frozen for 2 years in a biobank before spectral analysis. The biopsy samples were not gender matched and age matched. No significant changes in fixation or paraffin embedding occurred during the analysis period, and no tissue architecture degradation was observed. In addition, no diathermy effect was presented by the samples; and no contributions of paraffin vibrational modes were apparent in the low-wavenumber region of FT-MIR spectra.

FT-MIR spectra were collected by transmission mode. The necessity of a non-destructive technique which is able to map tissue area was the motivation for using this mode. Moreover, our search deals with tissues samples that have complex structures. Thus, distinct locations across the samples were considered during data collection. This fact was the purpose for taking more than one spectra per sample. The training set was selected by using the KS algorithm,[40] which selected the training samples based on a Euclidian distance calculation assigning the samples more distant to the samples' mean to the training set, and the samples closest to the samples' mean to the test set. This ensure that the classification model covers all sources of variation within the dataset. Thus, only the pure spectral information was used to form the training set, and no information regarding donor demographics was used.

Influences of scattering effect, overlapping bands, noise and some reflective loss at the substrate–sample interface appear in the spectral raw data (Figure 2A). Thus, EMSC was performed for baseline correction of scattering effects, Savitzky-Golay smoothing (15 points) was performed to remove background noise, and normalization was applied to amide I peak (~1,650 cm^{-1}) to correct distortions (Figure 2B). Although there is a very subtle spectral differentiation between Low and High grades, it was necessary to perform multivariate classification models to identify the most significant spectral markers for differentiation.

RBF, linear, quadratic and 3rd order polynomial SVM models have been used in several biomedical publications on cancer diagnosis, prognosis and genetic profile.[48-61] However, few works have applied these SVM-models on mid-infrared (MIR) spectroscopic data derived from cancer samples.[16,62-67] The number of publications is even lower when considering application of variable reduction and selection methods followed by SVM to MIR spectroscopic data from cancer sample.[62,68-72]

4.1. SVM models

4.1.1 RBF-SVM

It was noted that σ was the most important parameter in the RBF-SVM classification. The bias, the errors in training and test sets, and the number of support vectors mostly depend on σ parameter. We emphasize that bias derives from differences between the model and the true predicted behavior, being related to over-fitting problems.[73,74]

It is observed that for σ and *C* variation, the bias values had three behaviors (Figure 3A). First, a low bias region may be related to lower values of σ ($\sigma \leq 1$), where smaller or no variations in bias is observed independently of the *C* values. Secondly, a

high bias region may be related to intermediary values of σ ($1 \leq \sigma \leq 10$) when $C \leq 5$, where bias starts to increase. A third region (lower bias region) can be related to high values of σ ($\sigma \geq 15$) when $C \geq 10$, where bias suddenly decreases. In these two cases, RBF-SVMs are sensitive to low values of *C*; in other words, if *C* is too low, then bias can increase quickly. In fact, bias reached its biggest values when $\sigma \geq 15$ and $C \leq 5$.

The error in training set grows when bias increases. The error increased in $\sigma \geq 15$ and reached its biggest values when $C \leq 5$, which coincides with the largest value of bias; then, error dropped down to zero when $\sigma \leq 15$ (which coincides with some low values of bias). This is more evident for Low than High grade (Figure 3B). Hence, correct classification of Low grade training set was ~75%; and ~85% for High grade considering the high bias region, while correct classification of training set was ~100% for both Low and High grades considering the low bias region.

The errors in the test set for Low and High grades were opposite: high error can be observed in $\sigma \geq 15$ for the first, while the same occurs in $\sigma \leq 1$ for the second, independently of the *C* values (Figure 3C-D). This might mean that RBF-SVM cannot achieve good classification results in these values of σ. It is only in $5 \leq \sigma \leq 10$ that this opposite behavior does not appear and both categories were classified. In this σ range, a maximum classification rate of 60% for both Low (when $C \geq 5$) and High grades (when $\sigma = 10$ and $C \geq 1$) were found. Furthermore, the number of support vectors for most RBF-SVMs was high and close to the number of training samples. A decrease in the number of support vectors can be observed in $\sigma \geq 15$ when $C \geq 10$, which coincides with a lower bias region (Table 2).

All this information can support the hypothesis of overfitting problems with high values of σ ($\sigma \geq 15$) when *C* values are small ($C \leq 5$). This can be confirmed by the fact

that this range of σ and *C* match with high bias region, with higher values of training error and with a high number of support vectors, which was close to the number of training samples.[62,73,74] All these relationships between σ and *C* parameters, bias, errors in training and test sets, and number of support vectors were important in conceiving that the best RBF-SVM model may be considered those with $\sigma = 10$ when $C \geq 5$.

4.1.2. Linear, Quadratic and Polynomial SVMs

Linear SVM had larger bias values than quadratic and 3rd order polynomial SVMs, independently of the *C* values (Figure 4A). Indeed, 3rd polynomial SVMs had all bias values as negative. Errors in the training set were smaller (≤ 0.1) for Low and High grades, where the second had slightly larger error than the first one. All models had higher correct classification for both grades. 3rd order polynomial SVM had the correct classification training set of 90% and the others had 100% for both grades, independently of the *C* values (Figure 4B).

In Figure 4C, it is clear that High grade was better classified than Low grade considering the test set, and independently of *C* values. In fact, the same opposite behavior in the test set error found in RBF-SVM classification occurred in these approaches. Low grade had larger error in test set, and its better classification rate was 40% by quadratic SVM independently of *C* values; and by 3rd polynomial SVM, but only when $C = 1$. 3rd polynomial SVM did not rate Low grade test set at any *C* values (except when $C = 1$). Linear-SVM correctly classified only 20% of Low grade test set. For High grade, linear SVM correctly classified 100% of test set, independently of *C* values. The same occurred in 3rd polynomial SVM (except when $C = 1$), which correctly classified 80% of the test set. Quadratic SVM correctly classified 80% of the High grade test set, independently of *C* values.

These SVM approaches used less number of support vectors in the classification than RBF-SVM models. Quadratic SVM classification used the larger number of support vectors (20), while 3rd polynomial SVM used 15 support vectors, and linear SVM used 18 support vectors (Table 2). The fact that linear SVM had larger bias values, zero training error, higher error in test set for Low grade and a relatively high number of support vectors may indicate overfitting problems. Despite 3rd polynomial SVM had smaller bias values and lower number of support vectors, it also presented relatively high training error (compared to the others) and higher error in the test set for Low grade. On the other hand, although quadratic SVM classification used a larger number of support vectors, it showed low bias values, low error in training set and succeeded to classify both Low and High grades test sets.

4.1.3. Kernels Comparison

In Table 3, the best RBF-SVM model and the linear, quadratic and polynomial SVMs were compared based on figures of merit such as sensitivity and specificity. It was clear that SVM models had lower sensitivity and specificity, with values varying around 20-100% and 0-80% for Low and High grade, respectively. Regarding linear and quadratic SVMs, they presented opposite values of sensitivity and specificity for Low and High grades, meaning that while one value is higher, the other is lower, and vice-versa for both grades. 3rd order polynomial SVM and RBF-SVM presented sensitivity and specificity values close to each other for Low and High grades; nevertheless, these values were lower considering a classification perspective. Moreover, the number of support vectors used in almost all classifications was very close to the total number of samples in the training set, which is an indication of possible overfitting.[62] Based on all these facts, this may point to inefficiency of these SVM models to classify our data.

4.2. Variable Reduction and Selection coupled to SVM

The best classification rates (80%) for both Low and High grades and both training and test sets using PCA-SVM were found using 10 PCs (Figure 5), which provided 98% of the explained variance. The search for the optimum number of PCs aimed to avoid overfitting problems, arbitrary separation, too much noise and degradation of the loadings interpretation. Scores plot identified significant spectral similarity/dissimilarity ($p<0.001$) between the Low and High grades also showed visual representation and interpretation of both (Figure 6B). In addition, loadings identified the most important segregating variables (wavenumbers) responsible for Low and High grade classification (Figure 6A).

Using SPA-SVM, the scores plot identified significant spectral separation ($p<0.005$) between the Low and High grade (Figure 6D). This approach selected the most relevant variables (wavenumber) responsible for Low and High classification. Loadings plot (Figure 6C) provided a visualization of the variable selected by SPA-SVM. For GA-SVM, the scores plot identified significant spectral segregation ($p<0.005$) between the Low and High grades (Figure 6F). GA-SVM generated the best classification using twenty variables (Figure 6E).

4.2.4. Comparing Variable Reduction and Selection Methods coupled to SVM

From the figures of merit for variable reduction and selection methods followed by SVM shown in Table 4, it is noted that Low grade category had larger values for most quality parameters in comparison to High grade considering all approaches, highlighting sensitivity (80-100%) and specificity (75-90%), for example; where GA-SVM had the best performance. All models had lower bias and used a relatively low number of support vectors in the classification.

GA-SVM correctly classified 100% of both training and test sets for Low grade, while it correctly classified 100% and 90% of the training and test sets for High grade. Its sensitivity and specificity values were 100% and 90% for Low grade, respectively; and both were equal to 80% for High grade. This trend of higher values for Low grade was extended to the other figures of merit. These high values (mostly close to 1 or 100%) confirmed the effectiveness of GA-SVM classification using a smaller number of support vectors compared to the other methods.

Regarding SPA-SVM, it correctly classified 84% and 80% of the training and test sets for Low grade, respectively; while it correctly classified 67% and 60% of training and test sets for High grade, respectively. This approach was the worst in comparison to the other methods, solely considering the classification rates. SPA-SVM classification used a high number of support vectors, which was close to the number of training samples.

On other hand, the PCA-SVM reduction method was slightly better than SPA-SVM. PCA-SVM correctly classified 100% of the training set and 80% of the test set for Low and High grades. It had sensitivity and specificity of 80% and 75% for Low grade; and 67% and 80% for High grades. PCA-SVM classification used the smallest number of support vectors compared to the other methods; and it was the second best approach considering classification rates and figures of merit in comparison to the other methods.

4.3. Potential Biomarkers and Spectral Differences

Discriminating wavenumbers identified from Low and High grade categorizations for prostate cancer are represented in Table 5 using variable reduction and selection methods coupled to SVM. The most intense variation between High grade

versus Low grade spectra are highlighted. In addition, spectral differences identified based on absorbance ratio between High and Low grade spectra are displayed in Figure 6. The selected variables can be related to functional groups composing structures of proteins and nucleic acids (Figure 6D-6E, Table 4). Similar results were found by many other studies.[18,23,62,65,66,68-72,75-78]

[Insert Table 5 here]

In Figure 2B, absorbance values are clearly larger in High grade spectra than Low grade spectra. Furthermore, spectral differences are mostly apparent in bands attributed to amide I, II and III and protein regions (1,400-1,585 cm^{-1}); followed by DNA/RNA (O–P–O symmetric stretch) (1,080 cm^{-1}) and DNA (O–P–O asymmetric stretch) (1,230 cm^{-1}) regions; RNA Ribose and DNA (C–O stretching) regions (1,120-1,180 cm^{-1}); glycogen (C–O–H bend) (1,030 cm^{-1}) region; and protein phosphorylation region (970 cm^{-1}) (Table 5 and Fig. 6). In fact, phenotypic alterations can be firstly evidenced by spectral differences.[18]

The spectral bands localized at 1,250-1,680–cm^{-1} can be attributed to deformation, stretching and bend modes of C–N, C=O, C-O, C-H and N-H of fatty acids, amino acids, amides I, II, III and proteins. Changes in amino acid conformation and reduced intermolecular aggregation of the tissue proteins promoted by cancer transformation tend to increase these spectral regions for High grade. Moreover, post-translational modifications related to DNA/RNA changes and alterations in phase I/II metabolizing enzymes expression can also explain high occurrence of protein spectral variation between Low and High grades.[18,23,62,65,66,72,75-77]

In addition, the spectral band localized in 1,120-1,180 cm^{-1} can be attributed to RNA Ribose and to C–O stretching of DNA; the band in 1,030 cm^{-1} can be attributed to

C–O–H bend of glycogen; the band in 1,080 cm^{-1} can be attributed to symmetric phosphate stretching vibrations ($\nu_s PO_2^-$) of DNA/RNA; and the band in 1,230 cm^{-1} can be attributed to asymmetric phosphate stretching vibrations ($\nu_{as} PO_2^-$) of DNA. Glycogen, ribose, deoxyribose and phosphate groupings are widely associated to nucleic acid conformation and metabolism. In fact, the stronger intermolecular interactions between nucleic acids resulting from intermolecular differentiations and changes in RNA/DNA conformation promoted by cancer might cause an increase in the related High grade spectral regions. Indeed, the key of prostate cancer grade discrimination has been associated to DNA vibrational modes.[18,23,62,65,72,75-78]

The band localized at 970 cm^{-1} can be attributed to symmetric stretching of phosphorylated protein monoester di-anionic phosphate bonds and to vibrations of nucleic acid phosphate groupings. Cell protein phosphorylation is responsible for the protein regulating metabolism, which includes biochemical processes of cell proliferation, differentiation and growth. Post-translational protein modification, an increase in the proliferative and differentiation processes, and cell cycle progression possibly caused by advanced cancer may represent an increase in the High grade spectral region related to protein phosphorylation.[18,29,62,66,72,77]

4.4. Comparing Multivariate Classification and Traditional Methods Applied to Prostate Cancer Screening

In Figure 7, sensitivity and specificity values of the GA-SVM and traditional methods applied to prostate cancer screening and categorization are shown. Sensitivity (Figure 7A) and specificity (Figure 7B) values derived from GA-SVM, Digital Rectal Examination (DRE), Gleason grading system (GS) and Trans-rectal Ultrasound (TRUS)-guided biopsy for Low and High grade classification are presented.

[Insert Figure 7 here]

As shown in Figure 7A, sensitivity values for Low grade by GA-SVM classification were higher in comparison to High grade and compared to the traditional methods for both grades, while the specificity values (Figure 7B) for Low and High grades by GA-SVM classification were the same. However, specificity values were only higher for the Low grade by GA-SVM classification compared to the traditional methods which had higher specificity values for the High Grade. These results of GA-SVM classification can corroborate the initial idea of this work to classify and detect spectral differences between early and advanced stages of prostate cancer, particularly from a screening perspective.

Some research has also pointed to high performance of variable reduction and selection coupled to SVM in the classification of cancer.[62,68-72] In fact, a study by Baker *et al.*[72] confirmed the success of cancer classification by GA-SVM, suggesting the to include this algorithm on a standard list of options since this method often provides optimum classification. However, this pilot study has as limitation the fact that all samples were prepared in the same lot, therefore it does not reflect completely the variability encountered by analyzing different samples over a long-time period, which would happen in the real clinical theater. Furthermore, the model constructed only distinguish Low grade (Gleason pattern 2) from High grade (Gleason pattern 3 and Gleason pattern 4) samples, thus the identification of single Gleason grades was not evaluated due to the small number of samples used. Therefore, further studies using more samples are necessary before drawing stronger conclusions about implementing this methodology in the field.

5. Conclusions

The results showed that the combination of variable selection methods with support vector machines analysis and FT-MIR can be successfully used for detection and differentiation of Low and High grades of prostate cancer, generating high sensitivity and specificity values. This paper demonstrates that the use of variable selection methods followed by a support vector machine can reduce drawbacks of independent SVM analysis, such as high time consumption in pre-processing and parameter optimization. The optimization of parameters such as bias and *C* were evaluated in the classification performance, producing more reliable classification models and reducing the probability of overfitting. The models applied to the MIR spectral data derived from prostate cancer samples selected bands which are responsible for the separation of Low and High grade spectral data sets. The High grade spectral data showed more intensity than the Low grade. The potential biomarkers were amide I, II and III and protein regions (1,400-1,585 cm^{-1}); followed by DNA/RNA (O–P–O symmetric stretch) (1,080 cm^{-1}) and DNA (O–P–O asymmetric stretch) (1,230 cm^{-1}) regions; RNA Ribose and DNA (C–O stretching) regions (1,120-1,180 cm^{-1}); glycogen (C–O–H bend) (1,030 cm^{-1}) region; and protein phosphorylation region (970 cm^{-1}).

It was demonstrated that the combination of FT-MIR data collected from tissue samples from Low and High grades and GA-SVM may work as a complementary or alternative tool for prostate cancer screening and classification, with higher sensitivity (100%) and specificity (80%), particularly in early stages compared to traditional methods of diagnosis. Thus, the potential diagnostic tools proposed in this paper describe a less time-consuming and not observer-dependent methodology. This can imply in early detection, less aggressive and cheaper treatments, better prognosis and consequently decreased mortality rates. However, more extensive research in this field

is required to assess the ability of vibrational spectroscopy for screening and classification approaches.

Acknowledgments

Laurinda F. S. Siqueira would like to acknowledge the financial support from the PPGQ/UFRN/CAPES and IFMA. Camilo L. M. Morais would like to thank CAPES-Brazil (Doutorado Pleno no Exterior, grant 88881.128982/2016-01) for financial support. K.M.G. Lima acknowledges the CNPq (grant 305962/2014-4) for financial support. In addition, the authors would like to acknowledge Mr. Godoy from Bruker Inc. for his collaboration.

References

1. Schröder FH, van der Maas P, Beemsterboer P, et al. Evaluation of the digital rectal examination as a screening test for prostate cancer. Rotterdam section of the European Randomized Study of Screening for Prostate Cancer. *J Natl Cancer Inst*. 1998;**90**(23):1817-1823.

2. Wilbur J. Prostate cancer screening: the continuing controversy. *Am Fam Physician*. 2008;**78**(12):1377-1384.

3. Kaffenberger SD, Penson DF. The politics of prostate cancer screening. *Urol Clin North Am*. 2014;**41**(2):249-255.

4. Brawley OW, Ankerst DP, Thompson IM. Screening for prostate cancer. *CA Cancer J Clin*. 2009;**59**(4):264-273.

5. Hoffman RM. Screening for Prostate Cancer. *N Engl J Med*. 2011;**365**:2013-2019.

6. Wolf AMD, Wender RC, Etzioni RB, et al. American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA Cancer J Clin*. 2010;**60**(2):70-98.

7. Misra-Hebert AD, Kattan MW. Prostate Cancer Screening: A Brief Tool to Incorporate Patient Preferences in a Clinical Encounter. *Front Oncol*. 2016;**6**:235.

8. Alberts AR, Schoots IG, Roobol MJ. Prostate-specific antigen-based prostate cancer screening: Past and future. *Int Urol*. 2015;**22**(6):524-532.

9. Sudarshan VK, Mookiah MR, Acharya UR, et al. Application of wavelet techniques for cancer diagnosis using ultrasound images: A Review. *Comput Biol Med*. 2016;**69**:97-111.

10. Iczkowski KA, Lucia MS. Current perspectives on Gleason grading of prostate cancer. *Curr Urol Rep*. 2011;**12**(3):216-222.

11. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. *Mod Pathol*. 2004;**17**(3):292-306.

12. Chen N, Zhou Q. The evolving Gleason grading system. *Chinese J Cancer Res*. 2016;**28**(1):58-64.

13. Gleason DF, Mellinger GT. Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging. *J Urol.* 1974;**111**(1):58-64.

14. Epstein JI, Zelefsky MJ, Sjoberg DD, et. al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. *Eur Urol.* 2016;**69**(3):428-435.

15. Lattouf JB, Saad F. Gleason score on biopsy: is it reliable for predicting the final grade on pathology? *BJU Int*. 2002;**90**(7):694-698.

16. Sattlecker M, Baker R, Sone N, Bessant C. Support vector machine ensembles for breast cancer type prediction from mid-FTIR micro-calcification spectra. *Chemometr Intell Lab Syst*. 2011;**107**(2):363-370.

17. Khanmohammadi M, Ghasemi K, Garmarudi AB. Genetic algorithm spectral feature selection coupled with quadratic discriminant analysis for ATR-FTIR spectrometric diagnosis of basal cell carcinoma via blood sample analysis. *RSC Adv*. 2014;**4**:41484-41490.

18. Theophilou G, Lima KMG, Briggs M, Martin-Hirsch PL, Stringfellow HF, Martin FL. A biospectroscopic analysis of human prostate tissue obtained from different time periods points to a trans-generational alteration in spectral phenotype. *Sci Rep*. 2015;**5**:13465.

19. Theophilou G, Lima KMG, Martin-Hirsch PL, Stringfellow HF, Martin FL. ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue: classifying subtypes of human cancer. *Analyst*. 2016;**141**(2):585-594.

20. Kelly JG, Trevisan J, Scott AD, et al. Biospectroscopy to metabolically profile biomolecular structure: a multistage approach linking computational analysis with biomarkers. *J Proteome Res*. 2011;**10**(4):1437-1448.

21. Zaera F. New advances in the use of infrared absorption spectroscopy for the characterization of heterogeneous catalytic reactions. *Chem Soc Rev*. 2014;**43**:7624-7663.

22. Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*. 2007;**8**(9):1243-1266.

23. Siqueira LFS, Lima KMG. A decade (2004 – 2014) of FTIR prostate cancer spectroscopy studies: An overview of recent advancements. *Trends Anal Chem*. 2016;**82**:208-221.

24. Siqueira LFS, Lima KMG. MIR-biospectroscopy coupled with chemometrics in cancer studies. *Analyst*. 2016;**141**:4833-4847.

25. Soares SFC, Gomes AA, Araujo MCU, Galvão Filho AR, Galvão RKH. The successive projections algorithm. *Trends Analyt Chem.* 2013;**42**:84-98.

26. McCall J. Genetic algorithms for modelling and optimisation. *J Comput Appl Math.* 2005;**184**(1):205-222.

27. Santos MCD, Morais CLM, Nascimento YM, Araujo JMG, Lima KMG. Spectroscopy with computational analysis in virological studies: A decade (2006–2016). *Trends Anal Chem*. 2017;**97**:244-256.

28. Neves ACO, Silva PP, Morais CLM, Miranda CG, Crispim JCO, Lima KMG. ATR-FTIR and multivariate analysis as a screening tool for cervical cancer in women from northeast Brazil: a biospectroscopic approach. *RSC Adv.* 2016;**6**:99648-99655.

29. Purandare NC, Patel II, Lima KMG, et al. Infrared spectroscopy with multivariate analysis segregates low-grade cervical cytology based on likelihood to regress, remain static or progress. *Anal Methods*. 2014;**6**:4576-4584.

30. Paraskevaidi M, Morais CLM, Lima KMG, et al. Differential diagnosis of Alzheimer's disease using spectrochemical analysis of blood. *Proc Natl Acad Sci USA.* 2017; 201701517.

31. Santos MCD, Nascimento YM, Araújo JMG, Lima KMG. ATR-FTIR spectroscopy coupled with multivariate analysis techniques for the identification of DENV-3 in

different concentrations in blood and serum: a new approach. *RSC Adv.* 2017;**7**:25640-25649.

32. Santos MCD, Nascimento YM, Monteiro JD, et. al. ATR-FTIR spectroscopy with chemometric algorithms of multivariate classification in the discrimination between healthy vs. dengue vs. chikungunya vs. zika clinical samples. *Anal Methods*. 2018;**10**:1280-1285.

33. Costa FSL, Silva PP, Morais CLM, et al. Attenuated total reflection Fourier transforminfrared (ATR-FTIR) spectroscopy as a new technology for discrimination between Cryptococcus neoformans and Cryptococcus gattii. *Anal Methods*. 2016;**8**:7107-7115.

34. Morais CLM, Costa FSL, Lima KMG. Variable selection with a support vector machine for discriminating Cryptococcus fungal species based on ATR-FTIR spectroscopy. *Anal Methods*. 2017;**9**:2964-2970.

35. Hughes C, Gaunt L, Brown M, Clarke NM, Gardner P. Assessment of paraffin removal from prostate FFPE sections using transmission mode FTIR-FPA imaging. *Anal Methods*. 2014;**6**:1028-1035.

36. Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometr Intell Lab Syst*. 2012;**117**:92-99.

37. Martens H, Stark E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *J Pharm Biomed Anal*. 1991;**9**(8):625-635.

38. Zimmermann B, Kohler A. Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Appl Spectrosc*. 2013;**67**(8):892-902.

39. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem*. 1964;**36**(8):1627-1639.

40. Kennard R, Stone L. Computer Aided Design of Experiments. *Technometrics*. 1969;**11**(1):137-148.

41. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;**2**(4):433-459.

42. Bro R, Smilde AK. Principal component analysis. *Anal Methods*. 2014;**6**:2812-2831.

43. Soares SFC, Galvão RKH, Araújo MCU, et al. A modification of the successive projections algorithm for spectral variable selection in the presence of unknown interferents. *Anal Chim Acta*. 2011;**689**(1):22-28.

44. Padilha CAA, Barone DAC, Neto ADD. A multi-level approach using genetic algorithms in an ensemble of Least Squares Support Vector Machines. *Knowledge-Based Systems*. 2016;**106**:85-95.

45. Devos O, Downey G, Duponchel L. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. *Food Chem*. 2014;**148**:124-130.

46. Baia TC, Gama RA, de Lima LAS, Lima KMG. FTIR microspectroscopy coupled with variable selection methods for the identification of flunitrazepam in necrophagous flies. *Anal Methods*. 2016;**8**:968-972.

47. Fisher SE, Harris AT, Khanna N, Sule-Suso J. Vibrational Spectroscopy: What Does the Clinician Need? In: Moss D, ed. *Biomedical Applications of Synchrotron Infrared Microspectroscopy: A Practical Approach*. Cambridge: Royal Soceity of Chemistry; 2010:1-28.

48. Mohapatra P, Chakravarty S, Dash PK. Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. *Swarm Evol Comput*. 2016;**28**:144-160.

49. Peng S, Xu Q, Ling XB, Peng X, Du W, Chen L. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Lett*. 2003;**555**(2):358-362.

50. Chen L, Xuan J, Riggins RB, Clarke R, Wang Y. Identifying cancer biomarkers by network-constrained support vector machines. *BMC Syst Biol*. 2011;**5**:161.

51. Nayyeri M, Noghabi HS. Cancer classification by correntropy-based sparse compact incremental learning machine. *Gene Reports*. 2016;**3**:31-38.

52. Vanitha CDA, Devaraj D, Venkatesulu M. Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. *Procedia Comput Sci*. 2014;**47**:13-21.

53. Ali S, Veltri R, Epstein JI, Christudass C, Madabhushi A. Selective invocation of shape priors for deformable segmentation and morphologic classification of prostate cancer tissue microarrays. *Comput Med Imaging Graph*. 2015;**41**:3-13.

54. Sun T, Wang J, Li X, et al. Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. *Comput Methods Programs Biomed*. 2013;**111**(2):519-524.

55. Ford W, Land W. A Latent Space Support Vector Machine (LSSVM) Model for Cancer Prognosis. *Procedia Comput Sci*. 2014;**36**:470-475.

56. Cao J, Zhang L, Wang B, Li F, Yang J. A fast gene selection method for multi-cancer classification using multiple support vector data description. *J Biomed Inform*. 2015;**53**:381-389.

57. Çınar M, Engin M, Engin EZ, Ateşçi YZ. Early prostate cancer diagnosis by using artificial neural networks and support vector machines. *Expert Syst Appl*. 2009;**36**(3):6357-6361.

58. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst Appl*. 2014;**41**(4):1476-1482.

59. Gertych A, Ing N, Ma Z, et al. Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Comput Med Imaging Graph*. 2015;**46**:197-208.

60. Wang H, Huang G. Application of support vector machine in cancer diagnosis. *Med Oncol*. 2011;**28**(Suppl 1):613-618.

61. Chen AH, Lin CH. A novel support vector sampling technique to improve classification accuracy and to identify key genes of leukaemia and prostate cancers. *Expert Syst Appl*. 2011;**38**(4):3209-3219.

62. Kelly JG, Angelov PP, Trevisan J, et al. Robust classification of low-grade cervical cytology following analysis with ATR-FTIR spectroscopy and subsequent application of self-learning classifier eClass. *Anal Bioanal Chem*. 2010;**389**(5):2191-2201.

63. Hughes C, Iqbal-Wahid J, Brown M, et al. FTIR microspectroscopy of selected rare diverse sub-variants of carcinoma of the urinary bladder. *J Biophotonics*. 2013;**6**(1):73-87.

64. Cheng CG, Tian YM, Jin WY. A study on the early detection of colon cancer using the methods of wavelet feature extraction and SVM classifications of FTIR. *Spectroscopy*. 2008;**22**(5):397-404.

65. Hands JR, Dorling KM, Abel P, et al. Attenuated total reflection fourier transform infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples. *J Biophotonics*. 2014;**7**(3-4):189-199.

66. Tian P, Zhang W, Zhao H, et al. Intraoperative diagnosis of benign and malignant breast tissues by fourier transform infrared spectroscopy and support vector machine classification. *Int J Clin Exp Med*. 2015;**8**(1):972-981.

67. Sattlecker M, Stone N, Bessant C. Current trends in machine-learning methods applied to spectroscopic cancer diagnosis. *Trends Anal Chem*. 2014;**59**:17-25.

68. Bergner N, Romeike BFM, Reichart R, Kalff R, Krafft C, Popp J. Tumor margin identification and prediction of the primary tumor from brain metastases using FTIR. *Analyst*. 2013;**138**(14):3983-3990.

69. Benerjee S, Pai M, Chakrabarty J, et al. Fourier-transform-infrared-spectroscopy based spectral-biomarker selection towards optimum diagnostic differentiation of oral leukoplakia and cancer. *Anal Bioanal Chem*. 2015;**407**(26):7935-7943.

70. Lee S, Kim K, Lee H, Jun CH, Chung H, Park JJ. Improving the classification accuracy for IR spectroscopic diagnosis of stomach and colon malignancy using non-linear spectral feature extraction methods. *Analyst*. 2013;**138**(14):4076-4082.

71. Zhang X, Thiéfin G, Gobinet C, et al. Profiling serologic biomarkers in cirrhotic patients via high-throughput Fourier transform infrared spectroscopy: toward a new diagnostic tool of hepatocellular carcinoma. *Transl Res*. 2013;**162**(5):279-286.

72. Baker MJ, Clarke C, Démoulin D, et al. An investigation of the RWPE prostate derived family of cell lines using FTIR spectroscopy. *Analyst*. 2010;**135**(5):887-894.

73. Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY: Springer; 2006.

74. Valentini G, Dietterich TG. Bias-Variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods. *J March Learn Res*. 2004;**5**:725-775.

75. Patel II, Martin FL. Discrimination of zone-specific spectral signatures in normal human prostate using Raman spectroscopy. *Analyst*. 2010;**135**(12):3060-3069.

76. Patel II, Trevisan J, Singh PB, et al. Segregation of human prostate tissues classified high-risk (UK) versus low-risk (India) for adenocarcinoma using Fourier-transform infrared or Raman microspectroscopy coupled with discriminant analysis. *Anal Bioanal Chem*. 2011;**401**(3):969-982.

77. Baker MJ, Gazi E, Brown MD, Shanks JH, Gardner P, Clarke NW. FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer. *Br J Cancer*. 2008;**99**(11):1859-1866.

78. Gazi E, Baker M, Dwyer J, et al. A Correlation of FTIR Spectra Derived from Prostate Cancer Biopsies with Gleason Grade and Tumour Stage. *Eur Urol*. 2006;**50**(4):750-761.

Captions for Figures

Figure 1. Illustration of Gleason grading system.

Figure 2. FT-MIR spectral dataset derived from Low and High grade categorization for prostate cancer. (A) Non-preprocessed spectral dataset and (B) average preprocessed spectral dataset by cut in the fingerprint region (800-1800 cm^{-1}), EMSC, Savitzky-Golay smoothing (window of 15 points) and normalization applied to amide I peak (~1,650 cm^{-1}).

Figure 3. Low and High grade FT-MIR classification for prostate cancer by RBF-SVM. (A) Bias; (B) error in training set; and (C) error in test set, varying σ and *C* parameters.

Figure 4. Low and High grade FT-MIR classification for prostate cancer by linear, quadratic and 3rd order polynomial SVMs. (A) Bias, (B) error in training set, and (C) error in test set, varying *C* parameter.

Figure 5. Low and High grade FT-MIR classification for prostate cancer by PCA-SVM. Influence of the number of principal components on the correct classification rates of training and test sets.

Figure 6. Low and High grade FT-MIR classification for prostate cancer by variable reduction and selection methods coupled to SVM. (A) Loadings plot derived from PCA-SVM. (B) Scores plot calculated by PCA-SVM. (C) Twenty-four wavenumbers selected

by SPA-SVM. (D) Scores plot calculated by SPA-SVM. (E) Twenty wavenumbers selected by GA-SVM. (F) Scores plot calculated by GA-SVM. (Where LowCal: Low grade calibration set; LowVal: Low grade validation set; LowPred: Low grade test set; HighCal: High grade calibration set; HighVal: High grade validation set; HighPred: High grade test set).

Figure 7. Multivariate classification and traditional methods applied to prostate cancer screening. (A) Sensitivity (SENS) (%) and (B) specificity (SPEC) (%) values derived from Digital Rectal Examination (DRE), Trans-rectal Ultrasound (TRUS)-guided biopsy, Gleason grading system (GS) and GA-SVM for Low and High grade.

Table 1. Equations for calculating the figures of merit. TP stands for true positive, TN for true negative, FP for false positive and FN for false negative.

Figure of merit	Equation	
Sensitivity (SENS)	$\left(\frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}\right)\times 100$	(1)
Specificity (SPEC)	$\left(\frac{\mathbf{TN}}{\mathbf{TN}+\mathbf{FP}}\right)\times 100$	(2)
Positive Predictive Value (PPV)	$\left(\frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}\right)\times 100$	(3)
Negative Predictive Value (NPV)	$\left(\frac{\mathbf{TN}}{\mathbf{TN}+\mathbf{FN}}\right)\times 100$	(4)
Youden's index (YOU)	$\mathbf{SENS}-(1-\mathbf{SPEC})$	(5)
Likelihood ratio positive (LR+)	$\left(\frac{\mathbf{SENS}}{1-\mathbf{SPEC}}\right)$	(6)
Likelihood ratio negative (LR-)	$\left(\frac{\mathbf{SPEC}}{1-\mathbf{SENS}}\right)$	(7)

Table 2. Low and High grade FT-MIR classification for prostate cancer by SVM models. Number of support vectors used by SVM models.

	C / σ	0.01	0.1	1	5	10	15	25	50	100	1000
RBF-SVM	**0.01**	25	25	25	25	25	25	25	25	25	25
	0.1	25	25	25	25	25	25	25	25	25	25
	1	25	25	25	25	25	25	25	25	25	25
	5	25	25	25	25	25	25	25	25	25	25
	10	25	25	25	25	25	25	25	25	25	24
	15	25	25	25	25	25	24	24	24	24	23
	25	25	25	25	25	24	24	23	23	23	22
	50	25	25	25	25	24	24	24	23	22	21
	100	25	25	25	25	24	23	23	22	22	21
	1000	25	25	25	25	24	23	23	22	22	21
Linear-SVM	-	21	21	18	18	18	18	18	18	18	18
Quadratic-SVM	-	20	20	20	20	20	20	20	20	20	20
Polynomial-SVM	-	15	15	15	15	15	15	15	15	15	15

Table 3. Low and High grade FT-MIR classification for prostate cancer by SVM models. Figures of merit for RBF, linear, quadratic and 3rd polynomial SVM. (Where, SENS: Sensitivity; SPEC: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; YOU: Youden's Index; LR+: Positive Likelihood; LR-: Negative Likelihood).

	RBF-SVM ($\sigma = 10, C \geq 5$)		**Linear-SVM**		**Quadratic-SVM**		**Polynomial-SVM**	
	Low Grade	High Grade	Low Grade	High Grade	Low Grade	High Grade	Low grade	High grade
SENS (%)	60	60	20	100	40	80	50	50
SPEC (%)	40	40	80	0	60	20	50	50
PPV (%)	50	50	50	50	50	50	50	50
NPV (%)	50	50	50	50	50	50	50	50
YOU	0	0	0	0	0	0	0	0
LR+	1	1	1	1	1	1	1	1
LR-	1	1	1	1	1	1	1	1

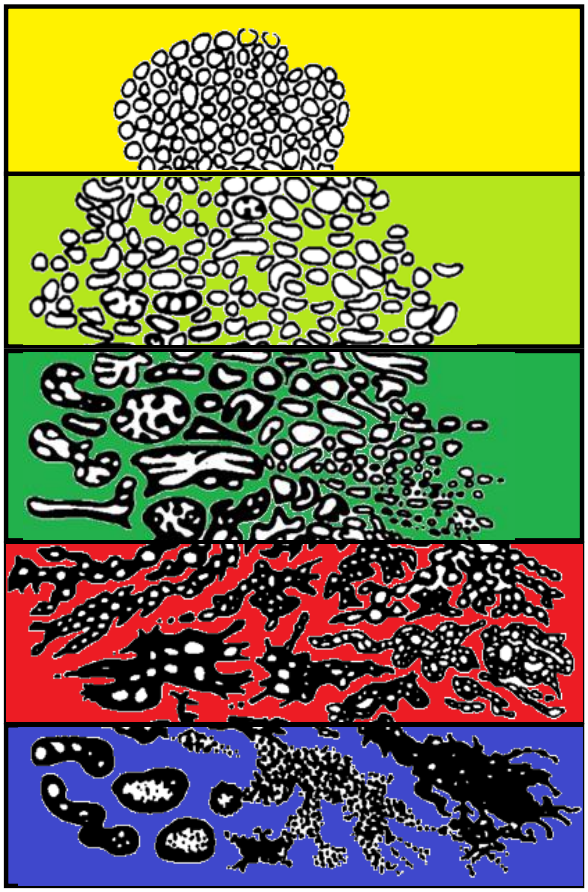
Table 4. Low and High grade FT-MIR classification for prostate cancer by variable reduction and selection methods coupled to SVM. Figures of merit for PCA-SVM, SPA-SVM and GA-SVM. (Where, SENS: Sensitivity; SPEC: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; YOU: Youden's Index; LR+: Positive Likelihood; LR-: Negative Likelihood).

	PCA-SVM		**SPA-SVM**		**GA-SVM**	
	Low Grade	High Grade	Low Grade	High Grade	Low grade	High grade
Training set CC (%)	100	100	84.62	66.67	100	100
Test set CC (%)	80	80	80	60	100	90
SENS (%)	80	66.67	80	80	100	80
SPEC (%)	75	80	75	71.43	80	80
PPV (%)	80	80	80	60	100	75
NPV (%)	80	60	80	60	100	60
YOU	60	41.67	60	60	75	60
LR+	4	2.67	4	3.5	4	1.5
LR-	0.25	0.45	0.25	0	0.64	0.40

Table 5. Low and High grade FT-MIR classification for prostate cancer by variable reduction and selection coupled to SVM. The most intense variation between High grade *versus* Low grade spectra is marked (Where: LHS: Left-hand shoulder; RHS: Right-hand side). Absorbance ratio: High grade/Low grade.

Tentative wavenumber assignments (cm^{-1})		**Absorbance ratio**
1,680	LHS Amide I (C=O stretch; C–N stretch)	1.05±0.001
1,650	Amide I (C=O stretch; C–N stretch)	1.10±0.001
1,620	RHS Amide I (C=O stretch; C–N stretch)	1.06±0.001
1,585	Amide I/II trough	1.19±0.001
1,570	LHS Amide II (N–H bend and C–N stretch)	1.17±0.001
1,550	Amide II (N–H bend and C–N stretch)	1.19±0.001
1,520	RHS Amide II (N–H bend and C–N stretch)	1.10±0.001
1,455	Protein (C–H and N–H deformation modes)	1.08±0.001
1,400	Fatty acids and amino acids (C=O stretching of COO-groups)	1.18±0.001
1,250-1,360	Amide III (C–N stretching)	1.06±0.001
1,230	DNA (O–P–O asymmetric stretch)	1.10±0.001
1,120-1,180	RNA Ribose and DNA (C–O stretching)	1.08±0.001
1,080	DNA/RNA (O–P–O symmetric stretch)	1.10±0.001
1,030	Glycogen (C–O–H bend)	1.08±0.001
970	Protein phosphorylation	1.00±0.001

Figure 1



Gleason Pattern

1. Well-formed and uniform distributed glands
(Gleason score 6 [3+3])

2. Predominantly well-formed glands with minor poorly-formed/fused/cribriform glands
(Gleason score 7 [3+4])

3. Predominantly poorly-formed/fused/cribriform glands with minor well-formed glands
(Gleason score 7 [4+3])

4. Poorly-formed/fused/cribriform glands
(Gleason score 8 [4+4, 3+5, 5+3])

5. Necrosis, cords, sheets, solid nests, single cells with or without poorly-formed/fused/cribriform glands
(Gleason score 9 [4+5, 5+4] and 10 [5+5])

Figure 2

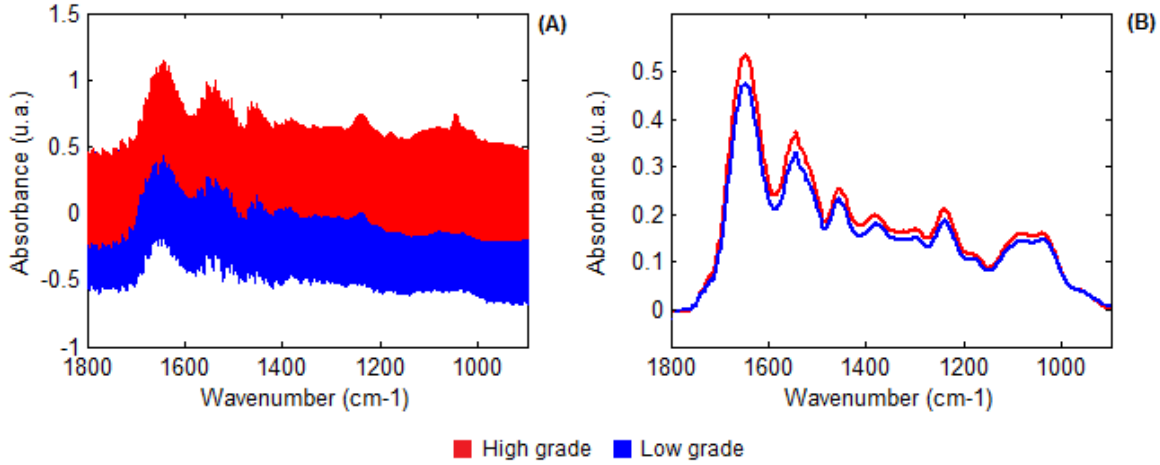

(A)
(B)
Absorbance (u.a.)
Wavenumber (cm-1)
1.5
1
0.5
0
-0.5
-1
1800
1600
1400
1200
1000
0.5
0.4
0.3
0.2
0.1
0
High grade
Low grade

Figure 3

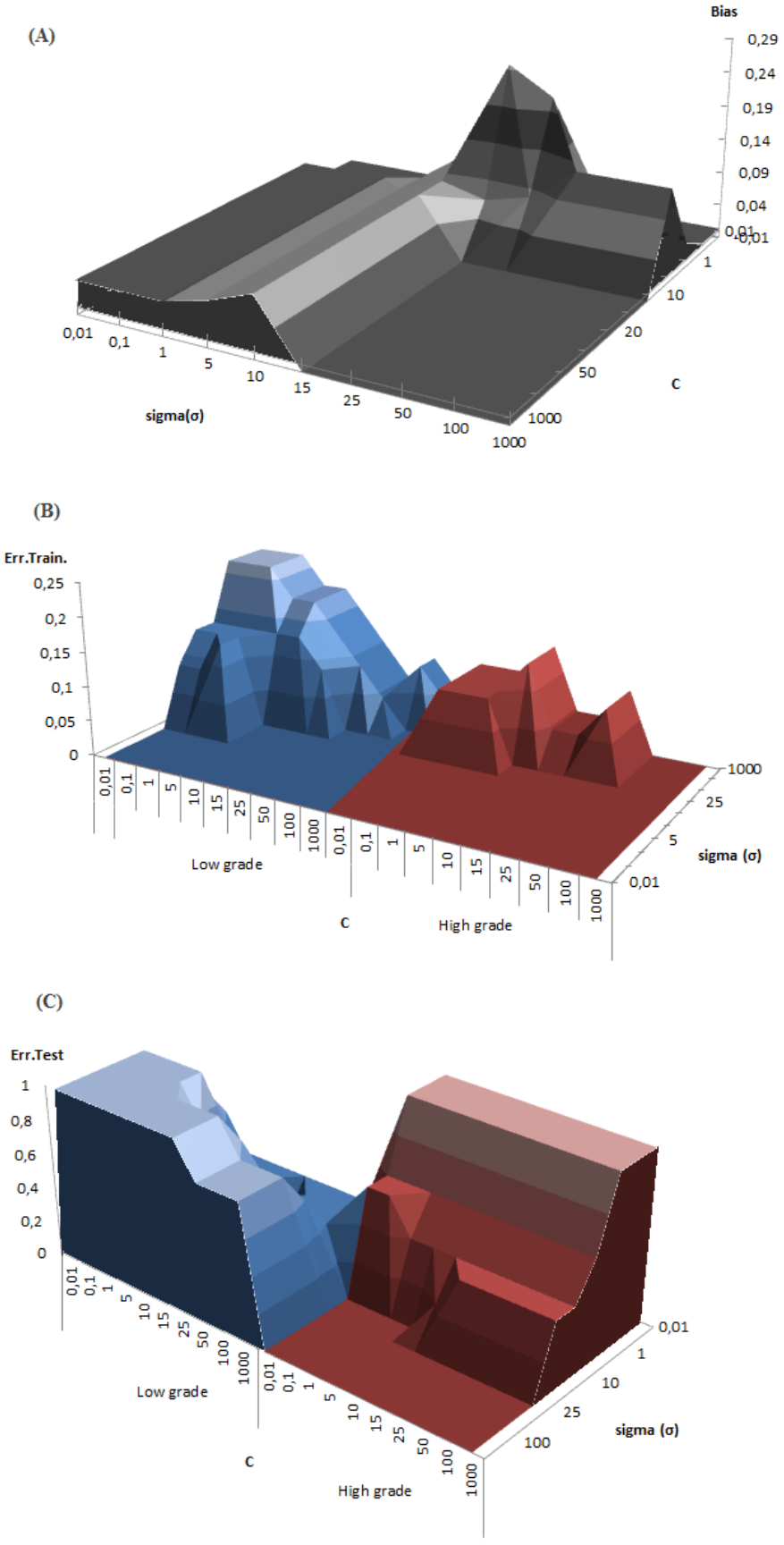

(A)
Bias
0,29
0,24
0,19
0,14
0,09
0,04
-0,01
0,01
0,1
1
5
10
15
25
50
100
1000
sigma(σ)
1
10
20
50
1000
C
(B)
Err.Train.
0,25
0,2
0,15
0,1
0,05
0
Low grade
High grade
C
sigma (σ)
(C)
Err.Test
1
0,8
0,6
0,4
0,2
0
Low grade
High grade
C
sigma (σ)

Figure 4

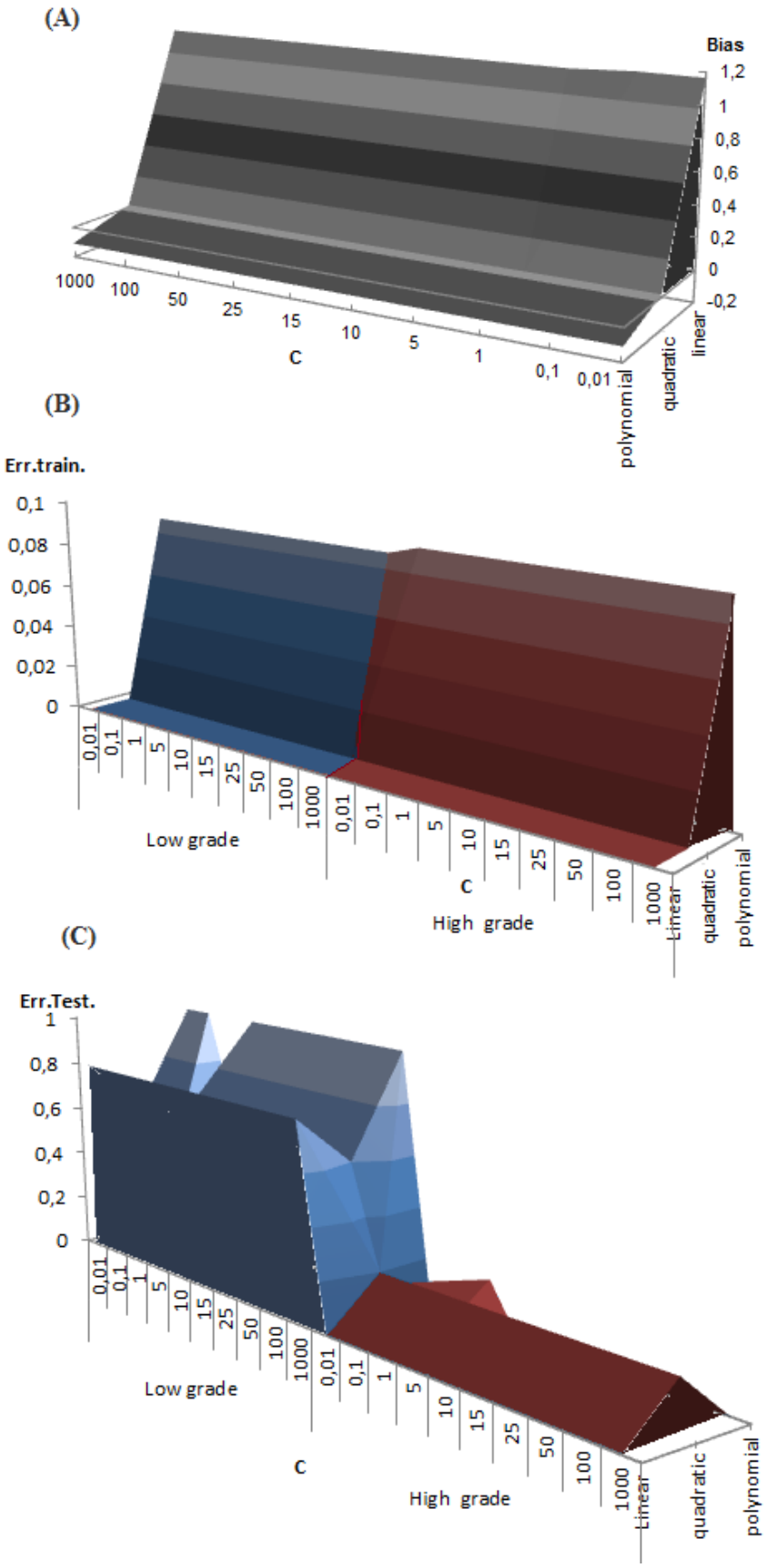

(A)
Bias
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
1000
100
50
25
15
10
5
1
0,1
0,01
C
linear
quadratic
polynomial
(B)
Err.train.
0,1
0,08
0,06
0,04
0,02
0
Low grade
High grade
C
linear
quadratic
polynomial
(C)
Err.Test.
1
0,8
0,6
0,4
0,2
0
Low grade
High grade
C
linear
quadratic
polynomial

Figure 5

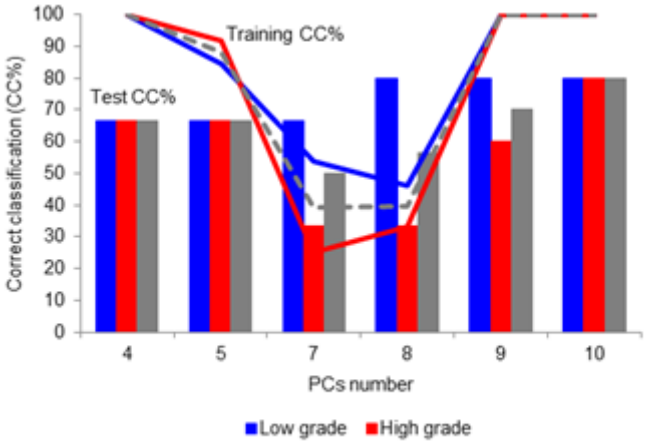

Training CC%
Test CC%
Correct classification (CC%)
100
90
80
70
60
50
40
30
20
10
0
4
5
7
8
9
10
PCs number
Low grade
High grade

Figure 6

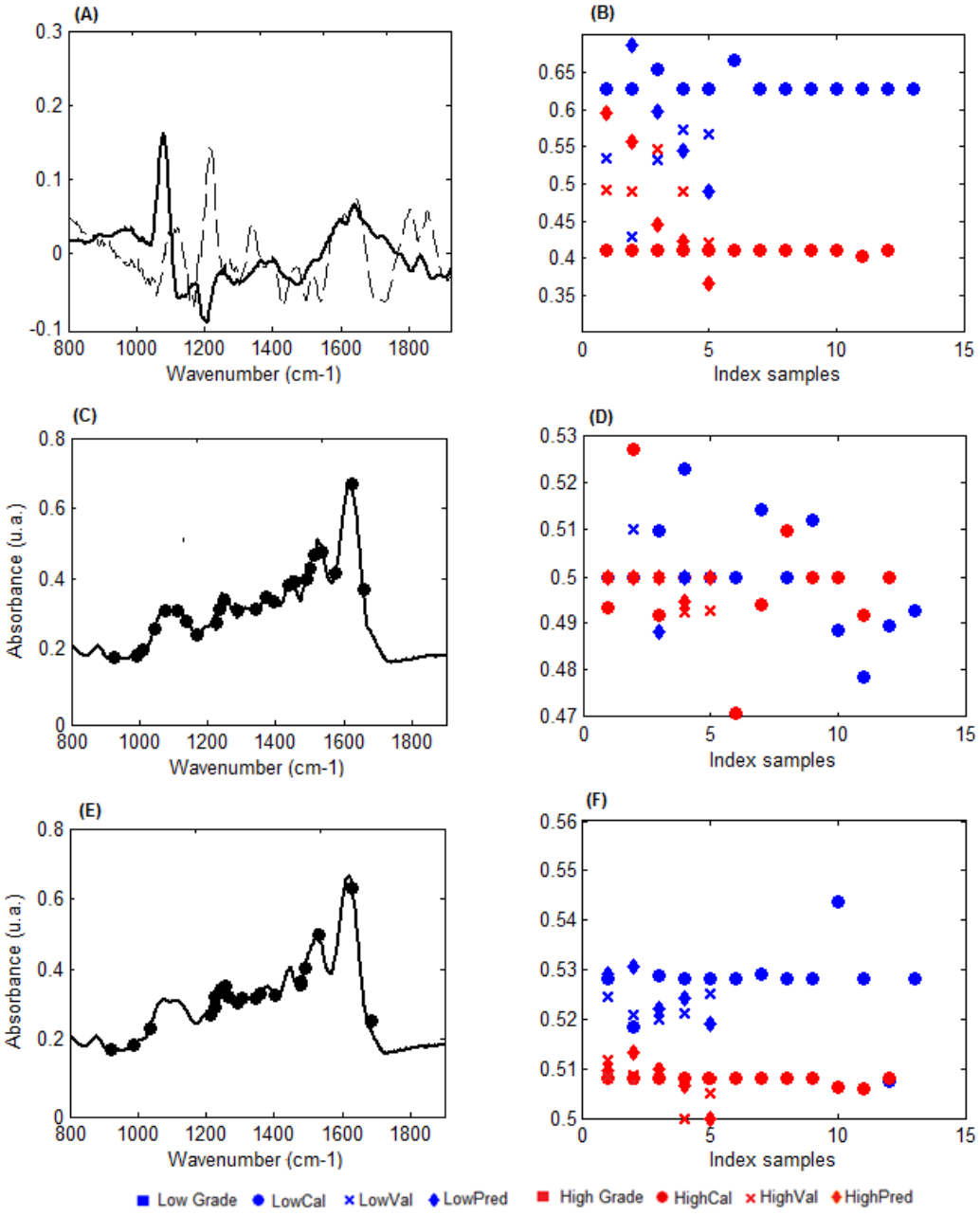
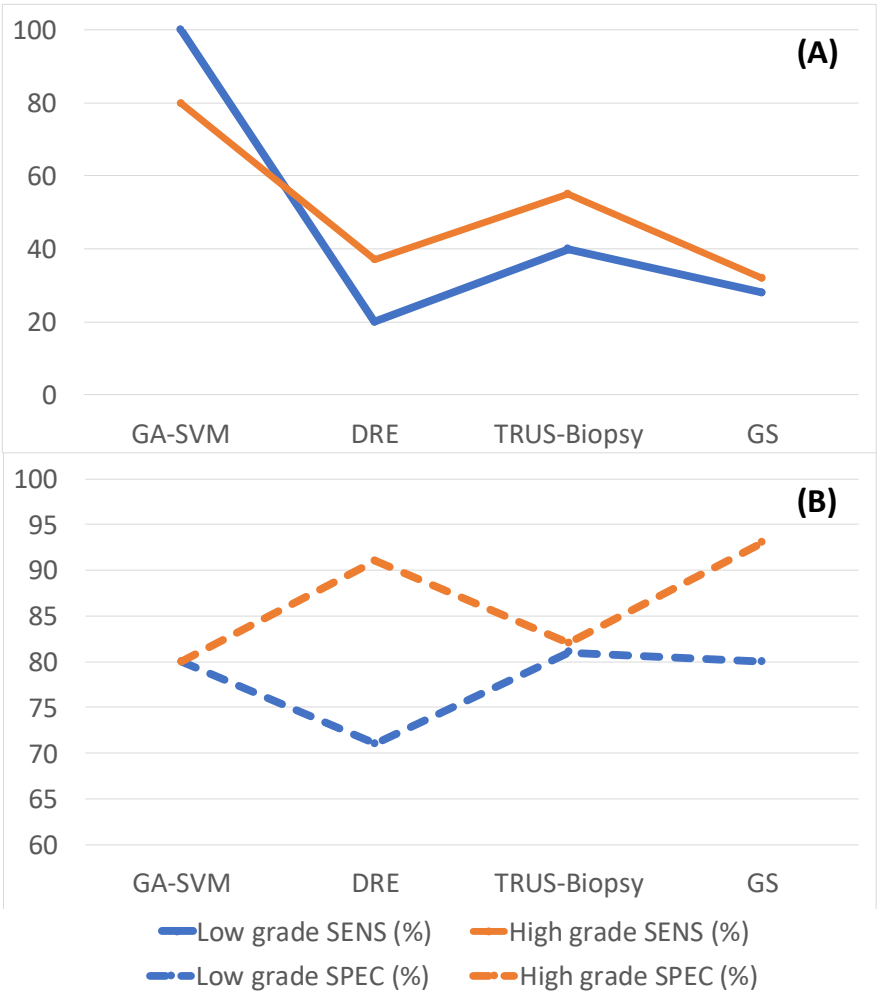

(A)
Wavenumber (cm-1)
(B)
Index samples
(C)
Absorbance (u.a.)
Wavenumber (cm-1)
(D)
Index samples
(E)
Absorbance (u.a.)
Wavenumber (cm-1)
(F)
Index samples
Low Grade
LowCal
LowVal
LowPred
High Grade
HighCal
HighVal
HighPred

Figure 7


100
80
60
40
20
0
(A)
GA-SVM
DRE
TRUS-Biopsy
GS
100
95
90
85
80
75
70
65
60
(B)
GA-SVM
DRE
TRUS-Biopsy
GS
Low grade SENS (%)
High grade SENS (%)
Low grade SPEC (%)
High grade SPEC (%)