Discriminative Spectral Pattern Analysis for Positive Margin Detection of Prostate Cancer Specimens using Light Reflectance Spectroscopy

Rahilsadat Hosseini[a], Henry Chan[b], Payal Kapur[c], Jeffrey Cadeddu[d], Hani Liu[b] and Shouyi Wang[a]

[a] *Department of Industrial Engineering,* [b] *Department of Bioengineering , at University of Texas at Arlington, Arlington, TX, USA,*
[c] *Department of Pathology,* [b] *Department of Urology, at University of Texas Southwestern Medical Center, Dallas, TX, USA*

Abstract

For localized prostate cancer, one treatment is prostatectomy which surgically removes the prostate gland. However, some undetectable cancer cells may be left as positive surgical margins, leading to a high risk of cancer recurrence. It is highly desirable to develop a portable and accurate classification methodology that detects positive margins on human prostate specimens immediately after their removal during surgery. This study applied data mining techniques on the Light Reflectance Spectroscopy (LRS) data taken from ex vivo human specimens and developed a novel classification algorithm that could enable real-time, positive-margin identification during surgery.

Specifically, a total of 184 LRS measurements taken from human prostate specimens ex vivo were classified to normal or cancerous tissue with support vector machines for binary classes and were also classified to normal, cancerous and transition-to-cancer class with an ensemble of trees for three classes. The 184 spectral data in this study were highly overlapped and imbalanced among classes. We solved the overlapping issue by first using expert knowledge to define a middle class (i.e., transition-to-cancer) between cancerous and normal tissue, and by second generating a moving spectral window through the range of LRS to find the best discriminative wavelength range. To solve the imbalanced problem, we removed irregular tissue measurements, followed by application of random undersampling from the majority class. We achieved sensitivity and specificity of 100% and 82% for binary classification, and accuracy of 0.61, 0.62, and 0.60 for the respective three classes with a spectral window length of 200 nm.

Keywords: Light Reflectance Spectroscopy (LRS), data mining, positive surgical margin, prostate cancer , support vector machines (SVM) , ensemble , random undersampling (RUS)

1. Introduction

Prostate cancer is the second cause of cancer-related deaths among men in the United States after lung cancer. For localized prostate cancer, one treatment is prostatectomy which surgically removes the cancer-containing prostate gland. During the surgery, the prostate gland with surrounding tissue is excised, with the best hope that all the cancer cells are completely removed while maximally preserving healthy surrounding tissue. However, due to limited time, technology and analysis currently available, any prostate cancer cells already spread on the capsule or/and surrounding tissue are too small to be seen/detected by the surgeons naked eyes and thus may be left behind as positive surgical margins. As stated in study [1], there is no globally accepted approach for positive margin (PM) detection of prostate cancer, however partial sampling can be used to measure this feature of prostate cancer. Although partial sampling can easily miss about 13 to 21 % of PMs and even slightly more missings in PMs of patients with low-risk to intermediate-risk prostate cancer. Harvard health publications, section of prostate knowledge, [2], discuss the ways to minimize the likelihood of a positive margin, one way is using the Gleason score (a clinical grading). The tissue sample is painted on the external surface with different colors of ink to designate to left and right sides prior to slicing. The tissue slices are used to evaluate the margins; one possible clinical feature is Gleason score as the criteria for the positive surgical margin, a high GS (greater than 7) is a sign of correlation. In a more recent study [3], a video-rate structured illumination microscopy (VR-SIM) of a removed tumor is established as an alternative to intra-operative frozen section pathology to reduce additional treatment and minimize tumor recurrence. They generate gigapixel panorama images of the surface that can be interpreted by pathologists. In the previous studies of our own research team, [4, 5], it is suggested that intraoperative frozen section analysis is time-consuming and inefficient. Therefore, it is proposed to apply light reflectance spectroscopy as a more viable, less expensive and quicker approach to differentiate malignant from benign tissue.

It is clinically important and highly desirable to develop a technique/methodology that can be used during prostatectomy to provide the surgeon on site with a portable, objective tool for accurate identification of positive margins. In this way, improved accuracy of surgical decision-making will lead to better efficacy of the treatment, elimination of further chemo- or radiation- therapy, and higher quality of patients lives. The main focus of this study was to classify positive surgical margins on human prostate specimens that were measured by a hand-held light reflectance spectroscopy (LRS) device. To our

knowledge, considering the traditional methods like frozen section, there is not such a fast computational model that learns the scattered light to detect the positive margin efficiently. Our algorithm development was based on spectral measurements of LRS with the application of data mining techniques. In particular, we overcame two issues that we encountered in this classification problem: the data taken from cancer versus normal classes were highly imbalanced, and the samples' data were overlapping. Our novel solution consisted of (1) defining a moving window through the wavelength range, (2) removal of some measurement data taken from irregular specimen tissues, (3) definition of a third class as an intermediate stage of tissue class between cancer and normal, and (4) novel approach using random undersampling (RUS) and tuning the parameters of the most promising machine learning algorithms via N-fold cross validation. In this way, our approach yielded an exhaustive search through the wavelength range and found an optimized spectral window location and window length.

Light Reflectance Spectroscopy (LRS) is a noninvasive measurement/imaging modality that can be used to quantify or classify biological tissues in vitro or in vivo [6, 7]. The LRS rests on the principle that a thin beam of white light illuminated on a piece of biological tissue through a fiber undergoes forward light scattering within the tissue, and a portion of the scattered light is randomly back-scattered near the surface and collected by a detection fiber placed a few hundred microns away from the delivery fiber. The detected optical signals are converted to an optical spectrum by a spectrometer, containing characteristic features of the measured biological tissue. Thus, LRS has been used to measure the degree of light scattering and light absorption that results from a variety of chromophores at the cellular level. In particular, light scattering intensity is highly associated with the morphology of the tissue, so its spectral signatures can facilitate differentiation between malignant and benign tissue, as demonstrated by our previous studies and others [8, 9] .

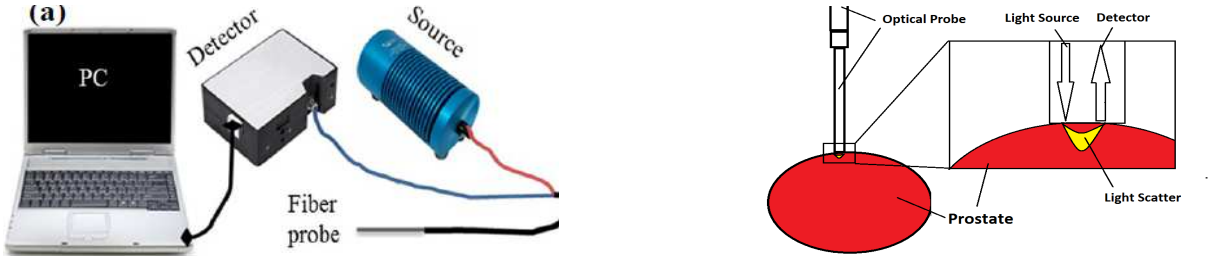
The most important step in a data mining approach is to find the best machine learning algorithm. In recent studies [4, 5], a linear fitting model was applied to extract features from the morphological properties of the LRS data and thus to classify cancer from normal. It used 5 extracted features from the spectra and fitted a logistic model, which gave rise to 0.86 and 0.85, respectively, for sensitivity and specificity. In an earlier study [9], a classification tree was fitted on the normalized raw LRS data to classify cancer from normal. With an application of clustering for label definition, they achieved 0.94 and 0.64 for sensitivity and specificity after selection of specific wavelengths. In this study, we defined the best machine learning algorithms to be the ones that yielded the highest performance in the most informative wavelength range. We applied support vector machines with a radial basis function [i.e., SVM(RBF)] and an ensemble method (boosting) of trees with a random undersampling(RUS) technique. With this approach, we classified cancer versus normal tissue as a binary problem and achieved a sensitivity and specificity of about 1.00 and 0.82, respectively. The overall results presented in this paper show much improvement with respect to the past studies [4, 9]. Furthermore, we applied the ensemble method with RUS (RUSBoost) to explore the feasibility of identifying three types of tissues: cancer, normal, and transition-toward-cancer. Our multi-class classification algorithm resulted in the accuracy of 0.72, 0.60 and 0.63 with a spectral window length of 10 nm and 0.61, 0.62 and 0.60 with a spectral window length of 200 nm, for the respective three classes. Our results demonstrated that the RUSBoost method outperforms SVM on multi-class classification.

2. Apparatus & Data

The apparatus of LRS for the detection of positive surgical margins for prostate cancer utilized a needle-like optical probe (so called needle probe hereafter) with a diameter of 1 mm holding two 100-m fibers for the source and detector. The source-detector separation was about 370 m. The source fiber was connected to a tungsten-halogen light source (HL2000HP, Ocean Optics, Inc., Dunedin, FL, USA), and the detector fiber was connected to a hand-held spectrometer (USB 2000+, Ocean Optics, Dunedin, FL USA) with a spectral range of 350 nm to 1000 nm, as shown in Figure (1).

The measurements were performed at the University of Texas Southwestern (UTSW) Medical Center, Dallas, TX. Fresh ex vivo prostate specimens were obtained from patients undergoing robotic-assisted laparoscopic radical prostatectomy [10, 11]. Immediately after removal from the patient, the prostate specimen was then transferred to the pathology room and placed on a stage for LRS measurements. The needle probe was placed normal to the surface of a selected area of the prostate. The tip of the probe was in good contact with the tissue surface without pressing it. The LRS measurements were taken from several spots on each specimens surface. The spots were selected to be either (1) potential cancer tissues or (2) normal tissues on the prostate surface. The measured locations were color marked for later pathology confirmation.

The Gleason score (GS) is a clinical grading scale to mark or classify the grade of the prostate cancer and to rank the severity of cancer. In practice, pathologists are responsible for grading the prostate cancer and assign a grade from 1 to 5 to the tissue they observe using a microscope. Grade 1 is assigned to the prostate tissue that looks very similar to the normal prostate tissue. Grade 5 is assigned to the prostate tissue that has highest abnormal morphological patterns (i.e. the highest grade of prostate cancer). Grades 2 to 4 are assigned to the prostate tissue regions that have intermediate characteristics between Grades 1 and 5. Grades 1 and 2 are not labeled for biopsies due to their very low-risk factor [12]. Figure (2) demonstrates morphological distributions of prostate normal and cancerous tissue corresponding to 5 different Gleason scores [13]. Moreover, prostate cancer often has two grades within one region. In clinical practice, the two grades are both recorded to express the severity of the cancer. The first value represents the most common grade seen in the region, while the second value represents the less common grade in the cancer by volume. As an example, a Gleason score of 3 + 4 = 7 means that the majority of the cancer is at Grade 3, and the minority of the cancer is Grade 4. These scores are added to produce a Gleason score of 7 [12].



(a) Apparatus (b) Measurement

Figure 1: LRS apparatus and measurement on prostate sample

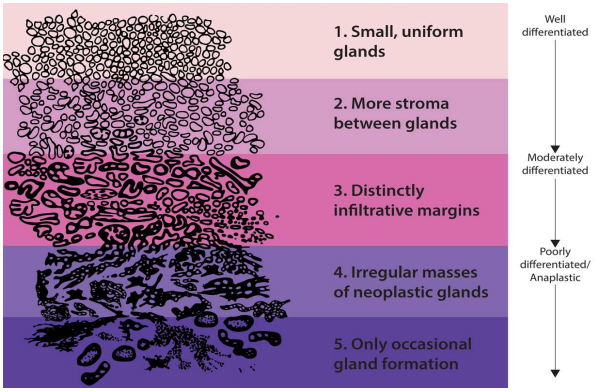


Figure 2: Gleason score corresponding to pattern

3. Method

The main purpose of this study was to develop an efficient data mining algorithm that can discriminate between normal and cancerous tissue based on tissue spectra taken by LRS. However, the experimental data obtained from normal and cancer prostate tissue were highly imbalanced and overlapped. In this study, we overcame the issue of overlapping and imbalanced classes in two steps; First, we added a new group by relabeling the cancer samples with GS(3+4) to the transition-to-cancer. Second, we defined a moving spectral window with length and location as parameters. The choice of spectral window affects the classification results significantly and needed to be optimized. These two steps contributed to more discrimination between classes. We repeated the analysis after excluding all samples with irregular/fatty tissue for comparison.

The following sub-sections include (1) explanation on how to group the data for more discriminative classes in Improvement Procedure, (2) the process of optimized window definition and spectrum selection in Feature Extraction and Selection, (3) formulation of the machine learning algorithms in Classification, and (4) validation of the results in Evaluation and Validation sub-section.

3.1. Improvement Procedure

As a part of pre-processing and noise removal, we defined the third class as transition-to-cancer between cancer and normal and removed samples with irregular/fatty tissues by applying the expert knowledge from pathology and previous experiences. In this way, we improved the result of initial classification based on the entire data set (i.e., all samples and entire wavelength range).

We had a total of 184 spectral measurements from 465-1140 nm consisting of three GS categories: 3+4, which is considered to be less aggressive cancer margins, 4+3, and 4+4, which are considered highly aggressive cancer margins and need to be clearly identified. We utilized GS initially to define cancer and normal classes and found: 29 aggressive cancer samples (labeled by +1 as cancer) and 155 non-aggressive and/or normal tissue samples (labeled by -1 as normal). Then, we relabeled cancer samples with GS= 3 + 4 to define the third class (transition-to-cancer). This intermediate class would be used to define a multi-class problem which might be useful to provide precautious lesions to the surgeons during the prostate cancer surgery. Table (1) lists the number of samples in each class (i.e., cancer, normal and transition-to-cancer) before and after noise (irregular or fatty tissue measurements) removal for binary and multi-class cases.

Table 1: Count of samples in each class: cancer, normal and Gleason Score (GS) 3 + 4, before and after removing irregular tissue measurements.

Class names	Count of the samples in class	
	Before removal of irregular tissues	After removal of irregular tissues
Cancer	9	9
Normal	155	134
GS (3 + 4) in cancer	20	15
Total count	184	158

3.2. Feature Selection

We realized that two obstacles prevented us from having a promising performance in classification: (1) imbalanced classes and (2) overlapping samples when using the entire wavelength range of the data. To overcome these issues, we decided not to use the entire wavelength range in the analysis. The spectral range of LRS data was 465.82 nm to 1140.97 nm, consisting of 2048 measurement points. The spectral interval was not a fixed value and varied from $0.27, 0.29, 0.30, \ldots, 0.38, 0.39(nm)$.

It was not wise to use a single wavelength since it would give us only one dimension for differentiating the classes. While a

range of wavelengths could offer more freedom to find discriminative patterns, it was crucial to determine an optimal spectral range. We searched over combinations of length and start point of the spectral window over the wavelength range.

3.3. Classification

We selected the support vector machines (SVM) and the ensemble of trees as the best-supervised learning approaches to construct a predictive function from a set of input-output pairs (i.e. training set). These two classifier models showed promising performance when evaluated on test set under N-fold cross validation.

3.3.1. Support Vector Machines (SVMs)

SVM is a supervised machine learning algorithm frequently used in classification [14], [15] with various applications in real-life problems including handwritten digit recognition [16], object recognition [17], speaker identification [18], face detection in images [19], text categorization [20], pattern recognition [21], biomedical applications [22, 23, 24, 25], and financial applications [26, 27].

SVM classifier uses a training set in $\Re^n$, with corresponding labels for each sample to find the decision boundary that separates classes with the highest margin (distance between closest samples of each class, known as support vectors) and the least error for misclassification. The hyperplane is represented with a normal vector $\mathbf{w}$ and a bias term b, both of which can be multiplied by any scalar $\lambda \neq 0$ without changing the hyperplane. The distance of an instance $\mathbf{x} \in \Re^n$ from the hyperplane is $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$, where $\langle \cdot \rangle$ is the dot product. This distance is called the *functional* distance, whereas the (actual) *geometrical* distance is measured as $f(\mathbf{x})/\|\mathbf{w}\|$.

Given m pairs of training data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, m$, where $y_i \in \{1, -1\}$ are the labels, the hyperplane that maximizes the margin in case of a linearly inseparable training data, can be found by solving the following problem. $C \sum_{i=1}^{n} \xi_i$ is the slack term added to each constraint and penalized in the objective as shown in Equation 1 with the loss function known as *hinge-loss* or *linear* penalty.

$$\min_{\mathbf{w},b} \{\|w\|^2/2 + C \sum_{i=1}^{n} \xi_i : y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1, \quad \xi_i \geq 0,\ i = 1, \ldots, m\}. \tag{1}$$

SVM classifier implicitly map the data into a higher dimensional space. In such a mapping $\mathbf{x} \mapsto \phi(\mathbf{x})$, $K(\mathbf{x_i}, \mathbf{x_j}) = \langle \phi(\mathbf{x_i}) \cdot \phi(\mathbf{x_j}) \rangle$, K is the *kernel* and can take many linear and nonlinear forms as long as it satisfies certain geometrical properties in the mapped space [14]. Given the new nonlinear distances through the kernel function, the rest is to find a linear separation in the mapped higher dimensional space. The calculation of these distances is done implicitly by the kernel function. Some of the most frequently used kernel functions are linear, polynomial, Gaussian radial basis and sigmoid. To solve a large problem quickly, chunking and decomposition methods have been proposed to make SVM training practical [28, 29]. Sequential minimal optimization (SMO) is an extreme decomposition method that iteratively solves the QP problem two variables at a time, whose solution can be found analytically [30, 31]. SVMs are extended to multi-class classification with one class against one class or one class against all classes training [32, 33].

In this study we applied Gaussian radial basis, $\exp(-\|\mathbf{x} - i - \mathbf{x}_j\|^2/2\sigma^2)$ for the kernel mapping, with the application of SMO, therefore, σ, the parameter in kernel mapping and C, the box constraint in objective, were tuned in N-fold cross validation.

3.3.2. Ensemble Methods of Decision Trees

There are two kinds of ensemble methods, bagging and boosting. Bootstrap aggregating (Bagging) [34] is an aggregated predictor that determine the final value by averaging for regression, and by outnumbered vote for the class. Bagging trees with bootstrap replicates outperform single tree because aggregation change good predictors to optimal ones. Boosting converts weak learners to strong ones and its focus is on error minimization. Boosting algorithm iteratively redistributes misclassified samples with higher probability [35]. And at the end combines weak learners from different subsets of data into a single prediction rule. Boosting outperforms bagging with respect to the minimization of the test error [36]. In this study, we selected to use the RUSBoost [37] boosting method that combines Adaboost.M1 for binary class or Adaboost.M2 for multi-class as ensemble techniques with random undersampling to overcome skewed classes. Adaptive boosting (AdaBoost) [38] is a popular strong ensemble method, with adaptive learning and no random selection. RUSBoost assigns equal weights to each sample in training data, then K weak learners are iteratively trained after random undersampling is applied on the majority class and a temporary training data is created K times. Sample weights are updated based on the pseudo-loss function of the temporary data in a way that more weight goes for misclassified samples and weak learners get boosted.

In this study, the weak learner in the boosting process is a decision tree with CART algorithm. Decision trees classify instances based on attributes which are represented in nodes and a sequence of rules through branches. Each attribute is selected such that impurity is reduced as much as possible in a recursive algorithm. There are several popular algorithms that have been developed in decision tree regression, such as ID3, C4.5, and classification and regression tree (CART). The most well-know algorithm to generate decision trees is known as C4.5 [39]. It builds decision trees from a set of training data by using the concept of Shannon entropy [40], a measure of uncertainty associated with a random variable. Based on the fact that each attribute of data can be used to split the data set into smaller subsets, C4.5 examines the relative entropy for each attribute. The attribute with the highest normalized information gain is used to make a decision. Large trees over-fit and are very complex, therefore they can be pruned by cross validation for cost-complexity-parameter taken from the CART algorithm

which is the combination of training error and penalty for model complexity. Another way to control the depth of the tree is to merge the leaves that come from the same parent while using the estimates of the optimal sequence of pruned sub-trees without pruning. When pruning and merging are both applied, the algorithm merges leaves with the highest vote of same class per leaf. In each step of the application of (RUSBoost) algorithm, there are some parameters that are tuned in the N-fold cross validation. The parameters in boosting are: number of ensembles in learning process, learning rate and cost matrix that sets the penalty for miss-classification of classes. Parameters in the decision tree are: minimum leaf size, split criteria, maximum number of splits, decision on pruning and merging. A parameter in random undersampling is the ratio of the larger to smallest class.

3.4. Evaluation and Validation

The classifiers were trained and evaluated with N-fold cross validation. For the binary problem (i.e., cancer vs. normal), receiver operating characteristic (ROC), the area under the ROC curve (AUC), average of sensitivity and specificity were reported. For the multi-class problem, the accuracy of each class was reported. Accuracy is defined as the ratio of correctly classified samples to all available samples. In this study, we relied on the average of sensitivity and specificity, instead of accuracy, because the data was highly imbalanced and the value of accuracy could be misleading. Sensitivity (or true positive rate) and specificity (or true negative rate) formulations are shown in Equation (2). ROC illustrates the performance of a binary classifier as decision boundary is varied. The ROC curve can be created by plotting the true positive rate (TPR) as sensitivity against the false positive rate (FPR) as (1 - specificity) at various threshold settings. In Equations (2), T means true, F means false, P means positive (cancerous), and N means negative (normal).

$$Sensitivity = \frac{TP}{TP + FN} \quad , Specificity = \frac{TN}{TN + FP} \qquad (2)$$

4. Results

In this section, we present the classification results for the binary class and multi-class prostate cancer problems. All of the classification results reported in this section are the best that we could achieve after tuning the parameters in the classifiers.

In Table (2), the classification performance of SVM on the original LRS data is shown, using the entire wavelength range and all of the samples. Out of 184 measurements consisted of three GS categories, 155 samples including GS(3+4) are considered normal, and 29 samples with GS(4+3 and 4+4) are considered cancerous. Rows one and two of the Table (2) report classification results without and with normalization, respectively. Although normalization over the wavelength range improved classification in study [9], in this study it reduced the specificity of normal class in all classification models. Therefore, we did not further report the results with normalization in the rest of the paper.

4.1. Spectral window's parameter optimization

As explained in section 3.2, we defined a moving window with parametric length through the available wavelength range. In Figure (3), we demonstrated the best performance of the binary SVM (RBF) classifier on each of the spectral windows after removal of GS (3 + 4) from the cancer class and irregular (fatty) tissues from both classes. The contours in Figure (3) summarizes the search process based on color-coded values for the average of sensitivity and specificity for different combinations of start-point of the window and its length. The brightest contours indicate that high performance of the classifier can be achieved with different lengths as long as the window includes the spectral range 634-644 (nm). It also reveals that with start-point less than 580 (nm), the average performance is about 53 %, however, after 580 nm, performance increases to the average of 83 %. Moreover, with spectral window length less than 200 (nm) and start-point greater than 580 (nm), the performance is more than 91 %.

The performance value along with the corresponding optimized window are listed in Table (3). The difference between the lowest and highest performance is about 3 %, indicating that with different optimized models and window length less than 200 nm, we can achieve classification in high accuracy. The highest, however, pertains to length 30 nm (from 634.65 to 664.84).

In Figure 4, the best performance of the RUSBoost classifier for different spectral window lengths are shown. Y-axis represents the accuracy of each class for the most promising starting point of the spectral window.

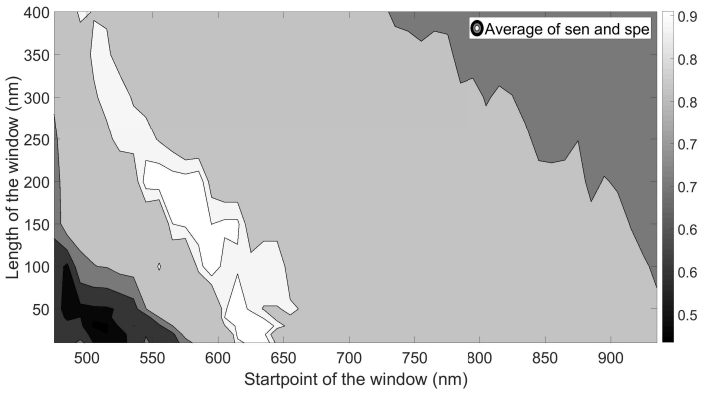


Figure 3: Best performance of the SVM model for different combinations of the length and start point of the window through the wavelength range. The range and the corresponding performance for each contour is provided in Table (3)

4.2. Binary Classification

Observing the white (brightest) contour in Figure (3), it was realized that with the length of spectral window less than 200 (nm), the performance of the best model was about 91-92 % and wider lengths decreased the accuracy. Therefore we chose the widest possible spectral window (i.e. 200 nm) for comparison in Table (4), showing the binary SVM classification performance. Table (4) clarifies improvement after defining the spectral window on wavelength range and removal of GS (3 + 4)

Table 2: SVM binary classification performance on raw data with full wavelength range, having all measurements with N-fold cross-validation

Preprocess on raw data consisting: (29 cancer, 155 normal)	SVM (RBF) binary classification performance with 10 fold cross validation				
	AUC	Accuracy %	Average of sen and spe	Sensitivity (sen)	Specificity (spe)
Without normalization	0.6293	72.0350	0.641	0.48333	0.7753

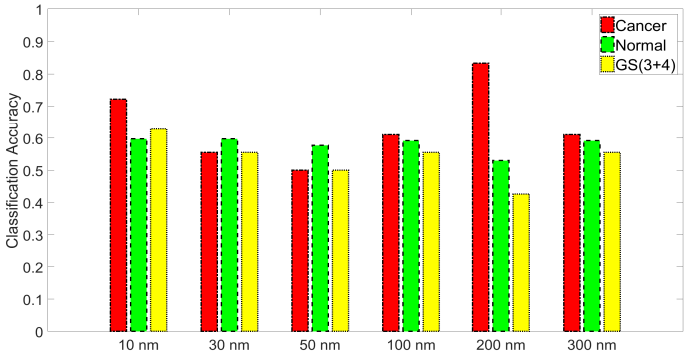


Figure 4: Best possible uniform performance of the RUSBoost method for various spectral window lengths. The start point of windows with lengths 10, 30, 50 and 100 is 624.86 nm, and for lengths 200 and 300, the start points are 874.86 and 664.84 respectively

Table 3: Performance of best model (SVM) for various spectral window lengths at different start points on wavelength after removal of GS(3 + 4) and irregular tissue measurements. Contour visualization of the values in the table is shown in Figure (3).

Window Range (nm)	Range Length	Avg (sen, spe)
634.65 to 644.75	10	0.9190
624.86 to 645.11	20	0.9120
634.65 to 664.84	30	0.9262
624.86 to 665.19	40	0.9225
614.67 to 664.84	50	0.9225
594.88 to 694.99	100	0.9190
584.93 to 735.14	150	0.9120
554.83 to 755.11	200	0.9157
524.75 to 774.89	250	0.9016
514.89 to 815.22	300	0.9016
505 to 855.07	350	0.8979
494.68 to 894.75	400	0.8979

Sen: sensitivity, spe: specificity

from the cancer samples and shows higher value for specificity. The values in Table (4) represent the best possible performance after tuning the parameters with 9 fold cross-validation. Data was divided to 9 folds since we had the maximum of 9 samples in cancer class. In each iteration, 1 fold, consisting of 1 cancer sample and 17 normal samples, was assumed to be the test set and the remaining 8 folds were used as the training set. In Figure (5a), the corresponding promising windows for the 91 % binary SVM performance are shown along with the average of samples from the three classes before removal of irregular measurements.

4.3. Multi-label Classification

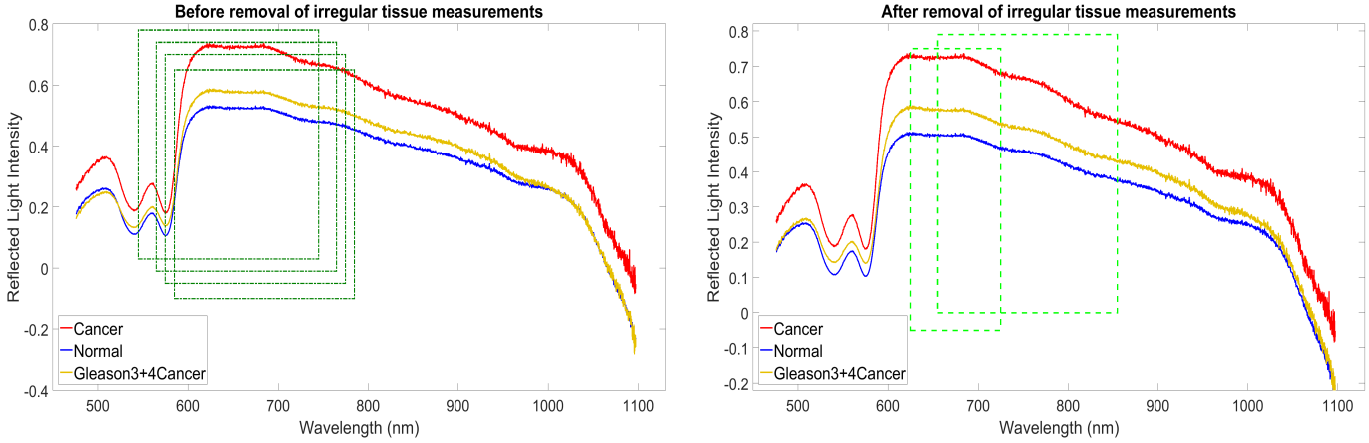
Removing samples with GS(3+4) from the cancer class provided the alternative of defining the transition-to-cancer class for a multi-class problem. However, by comparing the average values in each class in Figures (5a) and (5b), we realized that low-grade cancer samples were very similar to the class of normal and it would be very challenging to obtain high accuracy in three classes simultaneously.

Analyzing the results of the multi-class problem with RUSBoost classification, shown in Table (5), confirmed the existence of the challenge because of similarity between low-grade cancer and normal. We could not obtain high accuracy in three classes simultaneously,. While getting high accuracy in cancer and normal was easy, there was very low accuracy for the transition-to-cancer class. Nevertheless, we achieved a possible uniform accuracy (more than 50 %) among three classes with solving the multi-class problem by the RUSBoost algorithm. The most promising spectral windows after irregular tissues removal, were any length $\leq$ 100 nm in the wavelength range 625 to 725 nm, shown in Figure (4). And the most promising spectral window before irregular tissues removal, was wavelength 925-1125 nm, shown in Table (5). In Figure (4), the accuracy of each class for different lengths of window ranged from 0.5 to 0.6. However, shorter lengths yielded more promising results with respect to achieving more uniform accuracy among three classes.

5. Discussion

In this section, we discuss, first, the available options for the definition of the binary class problem, their results and possible application. Second, the effect of tuning parameters and their interpretation.

The main focus of this study was to identify high-risk margin meaning GS(4 + 3) and (4 + 4) which were considered aggressive from normal. However, there were tissues of grade 3 in cancer samples i.e. GS(3 + 4) sharing similar spectra with normal; we relabeled them as transition-to-cancer. This relabeling provided us with the alternatives on how to define the binary-class problem. Referring to the data in Table (1), which is indicating the size of the classes, we defined binary classification problems in 4 ways, which could be balanced or imbalanced. The best performance of the SVM(RBF) classifier for each problem is shown in Table (6). By comparing the results of models 1 and 3, it was confirmed that removal of GS(3 + 4) from cancer class improves the classifier performance significantly; average performance increased from 76 to 91 %. Comparison of model 3 and 4 yielded that combining normal with



(a) Average of 9 cancer samples (red) vs. average of 155 normal samples (blue) vs. average of 20 GS(3+4) samples (yellow), before removal of irregular tissue measurements, with promising windows (green rectangles) with length 200 nm shown for binary SVM performance.

(b) Average of 9 cancer samples (red) vs. average of 134 normal samples (blue) vs. average of 15 GS(3+4) samples (yellow) , after removal of irregular tissue measurements, with promising windows (green rectangles) of length 100 and 200 nm for RUSBoost performance

Figure 5: Average of the samples in each of the three classes, along with the optimized spectral windows for binary and multi label classification

Table 4: SVM binary classification performance on moving window with length 200 nm through wavelength range, classes are cancer measurements without GS 3 + 4 vs. normal measurements in 9 fold cross validation.

Wavelength (WL) Range (unit nm)		SVM RBF classification							
		Before removal of irregular tissue				After removal of irregular tissue			
WL start	WL end	Accuracy %	Avg(Sen, Spe)	Sensitivity (sen)	Specificity (spe)	Accuracy	Avg(Sen, Spe)	Sensitivity (sen)	Specificity (spe)
545.09	745.14	76.6875	0.8708	1	0.7417	84.3046	0.9120	1	0.8240
555.21	755.11	76.6875	0.8708	1	0.7417	84.9582	0.9157	1	0.8314
565.28	765.02	80.0208	0.8893	1	0.7787	84.3046	0.9120	1	0.8240
575.31	775.23	79.9916	0.8893	1	0.7787	84.2274	0.9118	1	0.8236
585.3	785.04	79.4360	0.8615	0.9444	0.7787	84.8810	0.9155	1	0.8310

Table 5: RUSBoost multi-label classification performance on moving window length 200 nm through wavelength range, classes are cancer vs. normal vs. cancer measurements with GS 3 + 4 with 9 fold cross validation

Irregualr tissue removal	Cost matrix among classes	WL start	WL end	Total Accuracy	Average of sensitivities among classes	Sen class Cancer	Sen of class Normal	Sen of class GS3+4C
No	Equal	925.13	1125.06	61.9587	0.6130	0.6111	0.6261	0.6019
No	Unequal	925.13	1125.06	58.7475	0.6255	0.6667	0.5708	0.6389
Yes	Unequal	495.07	695.34	49.7540	0.5222	0.5556	0.4926	0.5185
Yes	Unequal	875.18	1075.05	55.4358	0.5258	0.5000	0.5773	0.5000
Yes	Equal	655.18	855.07	71.7470	0.6207	0.8889	0.7880	0.1852

WL : Wavelength, Sen: sensitivity, GS3+4C: GleasonScore (3 + 4) in cancer

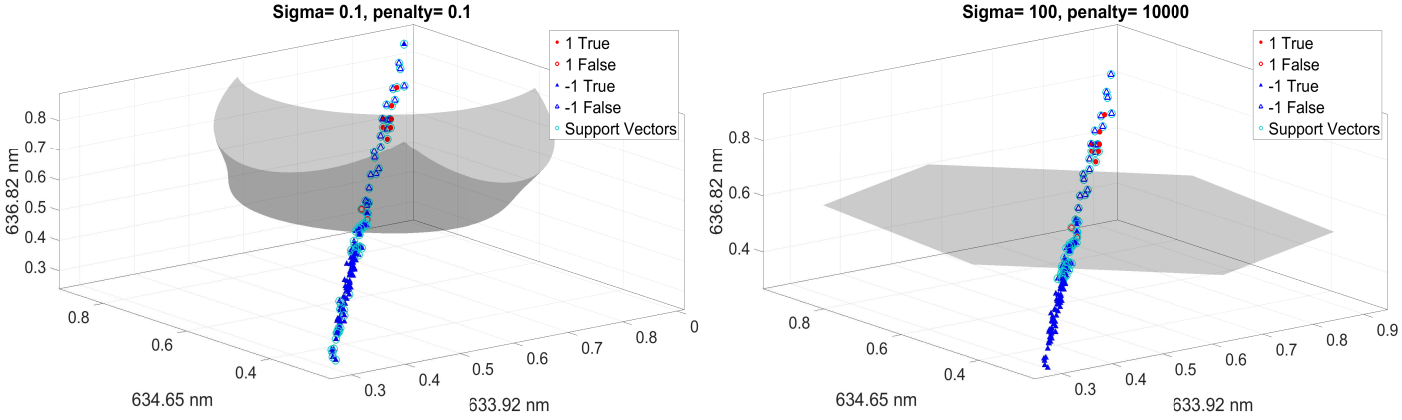
low-grade cancer samples, decreased specificity (about 2 %) with the same sensitivity. When we categorized GS(3 + 4) as the normal class, it could either be considered as an individual or in combination with normal class (refer to model 2 and 4). The former case, model 2, yielded 95 % average performance, while the latter case, model 4, 90 %. Model 2 indicated the result of a balanced problem with 9 aggressive cancer samples and 15 low-grade, while model 4 indicated the results for a highly imbalanced problem with 9 aggressive samples and 149 non-aggressive. Another practical application of these binary problems can be a 2-step classification process that would help the surgeon to predict the transition-to-cancer. Step one, is to use model 1 to discover normal pattern from cancer sample with $GS \geq 7$ with accuracy of 76 %, then if the sample is predicted to be cancer, in step two, model 2 is run to differentiate between high and low-risk factor and predict GS (4 + 4) and (4 + 3) (aggressive) from (3 + 4) (low-grade) with 95 % accuracy.

Since this data was very imbalanced, with the ratio of 1 to 17

Table 6: Four different ways of defining binary problem and the corresponding best performance of the SVM classifier for a fixed window length of 10 nm (minimum defined range) at the best performing start point on the wavelength range

Model No	Description on binary classification	Best achieved SVM performance with 9 fold cross validation						
		Wavelegn Range (nm)		AUC	Accuracy	average	sensitivity	specificity
		start	end			(sen, spe)	(cancer)	(non cancer)
1	cancer 44,43,34 vs. normal	584.93	595.25	0.7625	76.9038	0.7625	0.7500	0.7750
2	cancer 44, 43 vs. cancer 34	624.86	635.01	0.9537	95.5556	0.9537	0.9444	0.9630
3	cancer 44, 43 vs. normal	635.01	645.11	0.9190	85.5392	0.9190	1.0000	0.8380
4	cancer 44, 43 vs. normal and cancer 34	635.01	645.11	0.9099	83.6647	0.9099	1.0000	0.8199

44 : GS(4 + 4), 43 : GS(4 + 3), 34 : GS(3 + 4), sen: sensitivity, spe : specificity
All the irregular tissue measurements were removed for the anlysis of this table



(a) Low value for sigma and penalty

(b) High value for sigma and penalty

Figure 6: Three-dimensional plot of separating hyperplane in Support Vector Machines, different values for sigma and penalty results in the different shape of hyperplane and number of support vectors. High values for sigma results in high bias and low variance in the model, and high value in penalty results in low bias and high variance, and vice-versa. Circles show cancer and, triangles show normal samples. If they are correctly classified, they are filled, otherwise, they are hollow. The samples close to hyperplane that were taken as support vectors have an aqua circle around.

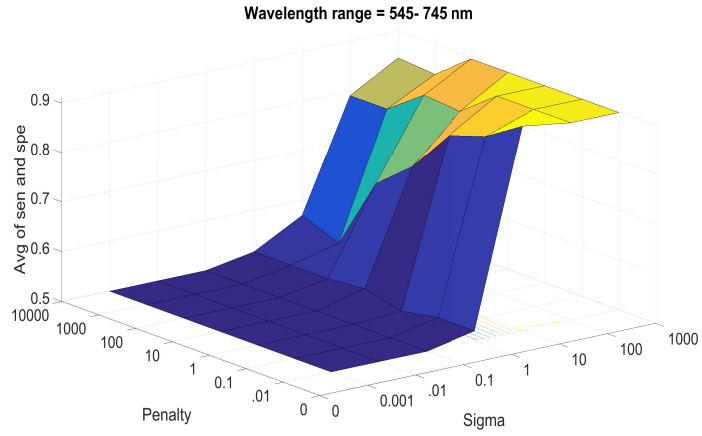


Figure 7: Grid search for optimized values of sigma and penalty (trade off between bias and variability of the model). Penalty (c) is the parameter of soft margin (cost of classification) and sigma is the parameter of Gaussian kernel to handle non-linearity. Binary classifier performance is defined as average of sensitivity and specificity on Z-axis.

for the class of cancer to normal, most of the classifiers tended to predict the test set as the majority class. The initial sensitivity and specificity with various machine learning classifiers were about less than 50 % and more than 88 % respectively. This problem was solved by tuning sigma and penalty in SVM and application of random undersampling (RUS) in boosting method. In Figure (6), the 3-dimensional visualization of binary classification and performance of SVM (RBF) for three wavelength values of 633.92, 634.65 and 636.82 nm is shown. These three wavelength values were selected based on the estimated predictor importance by permutation of out of bag observations in boosting trees from the promising spectral windows shown in Table (3 and 4). The value of sigma decided the shape of the separating hyperplane, and value for penalty decided the number of support vectors (SV). In the left sub-figure of Figure (6), sigma and penalty are 0.1, therefore, the hyperplane is curvy and many SVs are selected because of low penalty. In the right sub-figure, sigma is 100 and penalty is 10000 therefore, there is a linear hyperplane with less number of SVs. In Figure (6) We can see how the cancer class (red circles) are overlapping with normal class (blue triangles) at the upright corner of each 3-D plot. Hence, we demonstrate the vital role of proper parameter-tuning in Figure (7).

In Figure (7) performance of the binary-SVM on Z-axis is shown for different combination of sigma and penalty in a grid-search for wavelength range of (545.09 nm to 745.14 nm). The

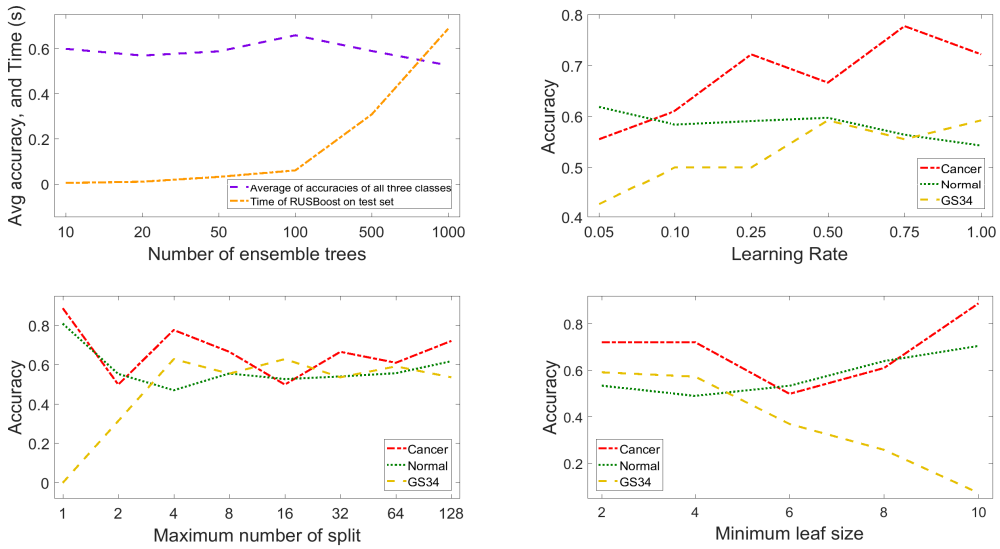


Figure 8: Grid search for parameters of ensemble of trees (boosting), values are shown for wavelength range of 624.86 – 635.01 (nm), for multi-class problem and after removal of irregular tissues

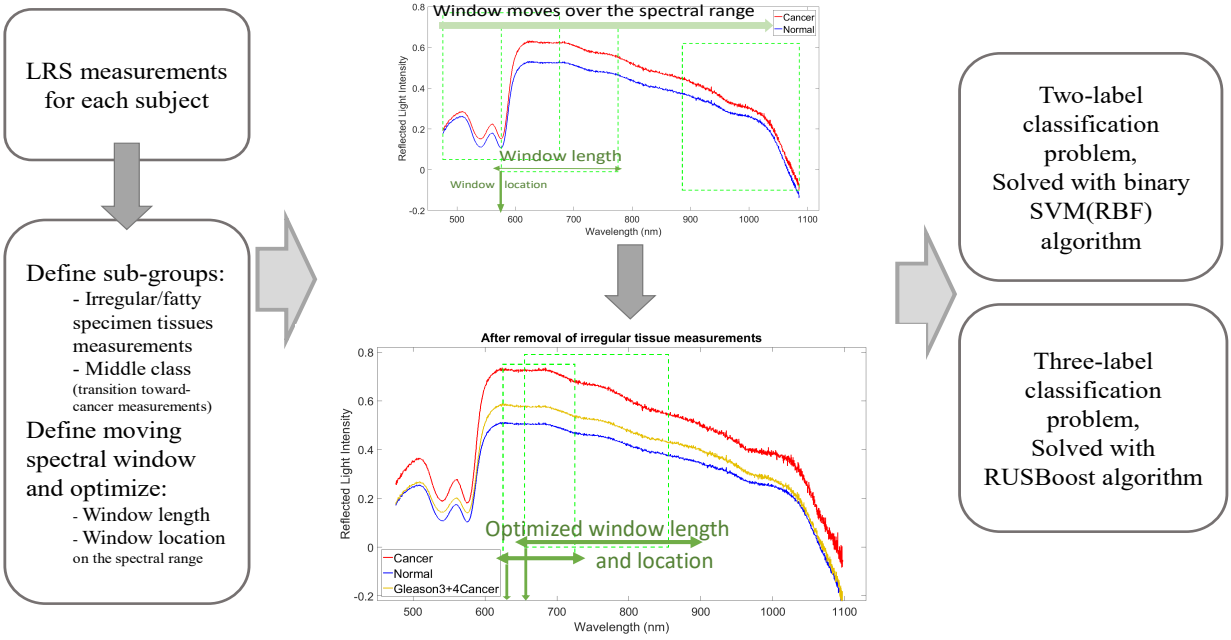


Figure 9: The improvement process describing how to start from (original data: 29 cancer samples vs. 155 normal) and achieve higher classification performance by having 3 classes with less imbalanced and overlapping samples and optimized spectral windows.

upward pattern indicates that for $\sigma \geq 10$ and penalty ≤ 0.1, performance is greater than 80 %, and on the other hand, for $\sigma \leq 1$, sensitivity ranges between 0 and 16 % and average performance is about 50 %.

In Figure (8), the grid search for the number of ensemble trees, learning rate, the maximum number of splits and minimum leaf size are shown. The goal of solving the multi-class problem was to achieve the most uniform accuracy in all of the three classes. Therefore, when the curves in each sub-plot of Figure (8), representing the accuracy of each class, were close to each other and above 50 %, it conveyed that the value for that parameter was ideal. But when the three curves started to diverge, it indicated that performance was not uniform among three classes. Figure (8) is indicating that to achieve uniform performance among classes, the best learning rate is 0.5, and best number of splits should be $\geq$ 16, and best value for the minimum number of leaf size should be $\leq$ 6. Regarding the number of ensemble trees, changing this parameter didn't improve performance significantly, therefore, we focused on reducing the computational time (the training and testing time of

RUSBoost algorithm). In the upper left sub-plot in Figure (8), test time increases significantly after 100 ensembles without an increase in average accuracy. The training time for 10, 100 and 1000 ensembles was 0.11, 0.95 and 9.29 (s) respectively, growing linearly without any improvement in performance. Hence, we decided to use 50 or 100 number of ensembles for the best-reported model. The parameters pertaining to best model with window length 10 (nm) in Figure (4) were 128, for the maximum number of splits, 2, for minimum leaf size and 50 ensembles. The ratio of majority classes (normal and GS(3 + 4)) to the small class (cancer) was 2, cost matrix was based on the frequency of each class (0.06 for cancer, 8.93 for normal, 0.60 for GS(3+4)), no pruning and merging was applied.

6. Summary and Conclusions

This study proposed an effective classification approach which is summarized in Figure (9) to discriminate between normal and cancerous LRS spectra. The proposed method consists of the following steps that contributed to a more discriminative spectrum between normal and cancer samples:

- Using Gleason score (3+4) to define a transition-to-cancer class which had similar values to the average of normal. This similar-to-normal behavior caused multi-label classification to be challenging and to perform in the range of low 60s %. On the other hand, removing the transition-to-cancer samples from cancer class, improved the binary classification performance (average of 91 %) by reducing the overlap.

- Removing the noisy data (i.e. the measurements with the irregular or fatty tissue). By comparing the numbers in Table (4), we conclude that this removal improves the specificity from (0.74 - 0.77) to (0.82-0.83) i.e. average of 6.5% increase for binary problem. Upward shift of the yellow curve when comparing Figure (5a and 5b), confirms reduction in overlap. By comparing the numbers in Table (5), irregular tissue removal yielded 1 % increase for the multi-class problem in window length of 200 nm which was not very significant.

- Defining the optimized spectral window through the wavelength range. For both binary and multi label classification, this window falls in the range between 600 to 700 (nm). The window is shown in Figures (5b) and (5a) for multi-class and binary-class respectively.

Regarding the binary classification, the best SVM model differentiating cancer from normal yielded about 91-92 % average of sensitivity and specificity for any window length less than 200 (nm). Specifically the spectral window of 634.65 - 664.65 (nm) was the best. However, all window lengths less that 300 nm through the range 515-815 (nm) yielded accuracy more than 90 %. We presented 3 alternatives to define the binary problem in addition to high risk v.s normal (model 3 in Table (6)). Based on the preference of the clinicians, any of the three models can be applied for positive margin detection with the average of sensitivity and specificity of 76 %, 95 % and 90 % respectively for model 1, 2 and 4, which were shown in Table (6). The training and testing time of the four mentioned binary models are 8.9 and 0.76 ms for model(1), 10.09 and 0.74 ms for model(2), 6.14 and 0.48 ms for model(3) and 8.07 and 0.67 ms for model(4), which confirms fast performance of the model. Moreover, a two step classification model could be applied by combining models 1 and 2 with augmented testing time of 1.51 ms; in a way that first step is to classify cancer from normal and then, if the prediction falls in cancer group, in the second step, model 2 can be applied to predict high risk from low risk.

References

[1] V1. Iremashvili, S.D Lokeshwar, MS. Soloway, L. Pelaez, SA. Umar, M. Manoharan, and M. Jord. Partial sampling of radical prostatectomy specimens: detection of positive margins and extraprostatic extension. *The American Journal of Surgical Pathology*, 37(2):219–25, 2013.

[2] Ma. B. Garnick. Positive surgical margins following radical prostatectomy. Online; accessed 28-Nov-2017.

[3] M. Wang, DB. Tulman, AB. Sholl, HZ. Kimbrell, SH. Mandava, K. Elfer, S. Luethy, MM. Maddox, W. Lai, BR. Lee, and JQ. Brown. Gigapixel surface imaging of radical prostatectomy specimens for comprehensive detection of cancer-positive surgical margins using structured illumination microscopy. *Scientific Reports*, 6(27419), 2016.

[4] M. S. C. Morgan, A. H. Lay, X. Wang, P. Kapur, A. Ozayar, M. Sayah, L. Zeng, H. Liu, C. G. Roehrborn, and J. A. Cadeddut. Light reflectance spectroscopy to detect positive surgical margins on prostate cancer specimens. *Journal of Urology*, 195(2):479–484, 2016.

[5] A. H. Lay, X. Wang, M. S. C. Morgan, P. Kapur, H. Liu, C.s G. Roehrborn, and J. A. Cadeddu. Detecting positive surgical margins: utilisation of light-reflectance spectroscopy on ex vivo prostate specimens. *BJU International*, 118(6):885–889, 2016.

[6] C. A. Giller, H. Liu, P. Gurnani, S. Victor, U. Yazdani, and D. C. German. Validation of a near-infrared probe for detection of thin intracranial white matter structures. *Journal of Neurosurgery*, 98:1299–306, Jun 2003.

[7] L. Qiu, D. K. Pleskow, R. Chuttani, E. Vitkin, J. Leyden, N. Ozden, S. Itani, L. Guo, A. Sacks, J. D. Goldsmith, M. D. Modell, E. B. Hanlon, I. Itzkan, and L. T Perelman. Multispectral scanning during endoscopy guides biopsy of dysplasia in barrett's esophagus. *Nature medicine*, 16 5:603–6, 1p following 606, 2010.

[8] V. Sharma, S. Shivalingaiah, Y. Peng, D. Euhus, Z. Gryczynski, , and H. Liu. Auto-fluorescence lifetime and light reflectance spectroscopy for breast cancer diagnosis: potential tools for intraoperative margin detection. *Biomed Opt Express*, 3:1825–40, Aug 2012.

[9] S. B. Kim, C. Temiyasathit, K. Bensalah, A. Tuncel, J. Cadeddu, W. Kabbani, A. V. Mathker, and H. Liu. An effective classification procedure for diagnosis of prostate cancer in near infrared spectra. *Elsevier, Expert Systems with Applications*, 37:3863 – 3869, 2010.

[10] A. Hoznek, Y. Menard, L. Salomon, and C.C Abbou. Update on laparoscopic and robotic radical prostatectomy. *Current Opinion In Urology*, 15(3):173 – 180, 2005.

[11] M. Menon, A. Shrivastava, and A. Tewari. Laparoscopic radical prostatectomy: conventional and robotic urology. *Urology*, 66(5):101 – 104, 2005.

[12] Understanding your pathology report: Prostate cancer, 2014. Online accessed 19-October-2014.

[13] Gleason grading system, 2014. Online accessed June-2017.

[14] N. Cristianini and J.S. Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, 2000.

[15] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, December 2001.

[16] B. Scholkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, pages 252–257. AAAI Press, 1995.

[17] V. Blanz, B. Scholkopf, H. Bulthoff, C. Burges, V. Vapnik, and T. Vetter. Comparison of view-based object recognition algorithms using realistic 3D models. In *Lecture Notes in Computer Science*, volume 1112, pages 251–256. Springer, 1996.

[18] M. Schmidt. Identifying speaker with support vector networks. In *Proceedings of Interface*, Sydney, 1996.

[19] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

[20] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of 10th European Conference on Machine Learning*, volume 1398, pages 137–142, Chemnitz, Germany, Apr. 1998.

[21] S. Lee and A. Verri. Pattern recognition with support vector machines. In *SVM 2002*, Niagara Falls, Canada, 2002. Springer.

[22] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugne, T. Furey, M. Ares, , and D. Haussler. Knowledge-base analysis of microarray gene expressiondata by using support vector machines. *PNAS*, 97(1):262267, 2000.

[23] T.N. Lal, M. Schroeder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf. Support vector channel selection in bci. *IEEE Transactions on Biomedical Engineering*, 51(6):1003–1010, 2004.

[24] W.S. Noble. *Support vector machine applications in computational biology*, chapter 3. Computational molecular biology. MIT Press, 2004.

[25] O. Seref, C. Cifarelli, O.E. Kundakcioglu, P.M. Pardalos, and M. Ding. Detecting categorical discrimination in a visuomotor task using selective support vector machines. In H. R. Arabnia, M. Q. Yang, and J. Y. Yang, editors, *Proceedings of the 2007 International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, volume 2, pages 580–587, 2007.

[26] Z. Huang, H. Chen, C.J. Hsu, W.H. Chenb, and S. Wuc. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37:543–558, 2004.

[27] T.B. Trafalis and H. Ince. Support vector machine for regression and applications to financial forecasting. In *International Joint Conference on Neural Networks (IJCNN'02)*, Como, Italy, 2002. IEEE-INNS-ENNS.

[28] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Proceedings of IEEE Neural Networks for Signal Processing*, pages 276–285, 1997.

[29] T. Joachims. Making large-Scale SVM Learning Practical. *Advances in Kernel Methods-Support Vector Learning*, pages 169–184, 1999.

[30] J.C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, 1999.

[31] C.C. Chang, C.W. Hsu, and C.J. Lin. The analysis of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 12(4):291–314, 1999.

[32] J. Weston and C. Watkins. Multi-class support vector machines, 1998.

[33] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2(5):265–292, 2002.

[34] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[35] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012.

[36] Leo Breiman. Bias, variance, and arcing classifiers. Technical report, 1996.

[37] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(1):185–197, Jan 2010.

[38] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156, San Francisco, 1996. Morgan Kaufmann.

[39] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Mateo, CA, 1993.

[40] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, Jul. 1948.