Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Statistical Analysis of a Lung Cancer Spectral Histopathology (SHP) Data Set

Xinying Mu,[1,2] Mark Kon,[1,2] Ayşegül Ergin,[2] Stan Remiszewski,[2] Ali Akalin,[3] Clay M. Thompson,[2,4] Max Diem[2,5]

[1]Department of Mathematics and Statistics and Program in Bioinformatics, Boston University, Boston, MA, USA

[2]Cireca Theranostics, 19 Blackstone St., Cambridge, MA, USA

[3]Department of Pathology, University of Massachusetts Medical School, Worcester, MA, USA

[4]Creative Creek, LLC., 2188 Lowell Point Rd, Camano Island, WA, USA

[5]Laboratory for Spectral Diagnosis, Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA

Abstract:

We report results on a statistical analysis of an infrared spectral dataset comprising 388 lung biopsies from a total of 374 patients. The method of correlating classical and spectral results and analyzing the resulting data has been referred to as spectral histopathology (SHP) in the past. Here, we show that standard bio-statistical procedures, such as strict separation of training and blinded test sets, result in a balanced accuracy of better than 95 % for the distinction of normal, necrotic and cancerous tissues, and better than 90 % balanced accuracy for the classification of small cell, squamous cell and adenocarcinomas. Preliminary results indicate that further sub-classification of adenocarcinomas will be feasible with similar accuracy once sufficiently large datasets have been collected.

October 12, 2014

1. Introduction: Spectral Histopathology (SHP)

SHP is an objective, reproducible optical method that utilizes the biochemical composition and disease-induced changes therein, rather than the cell morphology or tissue architecture, to render a diagnosis [1, 2]. SHP works on the principle that all biochemical components have their own distinct fingerprint infrared spectral signatures observable *via* infrared spectroscopy [3]. This technique is a well-established analytical method and widely employed in biophysical research to probe structure and dynamics of bio-molecular systems. The aforementioned compositional changes in tissue and cells then are manifested in small changes in the infrared spectra.

When observed through an infrared microscope, objects smaller than a human cell can be identified and their spectra can be acquired. Typically, spectra are collected from pixels about 5 to 6 μm on edge and *ca*. 25,000 of such 'pixel spectra' are collected for each 1 mm x 1 mm area of tissue. Thus, SHP samples the tissue at a spatial resolution (determined by the diffraction limit) lower than visible microscopy, but still sufficiently high to discern individual cells. Each pixel spectrum is a superposition of hundreds of individual component spectra, and represents a snapshot of the entire proteome, genome and metabolome of a cell or tissue pixel. It should be noted that other label-free methods, such as MALDI imaging, operate on a very similar principle of collecting spectral vectors – the mass distribution of proteins, for example – to construct images of cells and tissue based on biochemical composition, rather than morphology of the sample.

SHP has been developed over the past dozen of years [4-9], and has been shown to distinguish various tissue and cell types and normal from cancerous tissues by their biochemical composition. SHP can equally be applied to fresh frozen or de-paraffinized tissue sections [10, 11]. Results from this technique have been reported by a number of research groups worldwide, and have been verified for different cancers and different organs. However, only a few studies have collected datasets from a statistically significant number of samples to assess the sensitivity and specificity of SHP for a blinded test set. We present here initial results form a large study, involving a subset of the total of 388 patients and about 35 million pixel spectra, and a detailed evaluation of multivariate statistical methods used for analysis of this data set. The majority of the discussion will center on the statistical methods developed and applied to one half (the "non-vaulted" portion, see below) of the entire dataset.

2. Materials and Methods

In the past, many papers have been published in this journal and elsewhere [5, 11-14] to introduce the aims and methods of action of SHP, and the reader is referred to these references. Here, a detailed discussion of several new methodological steps – collectively referred to as the 'Cireca' steps (see Figure 2) – will be introduced that permit a pixel-based correlation of spectral features and classical histopathology, and thereby, traceability of any spectrum in the dataset to its original pathological diagnosis. This step is necessary to calculate a pixel-by-pixel balanced accuracy of SHP, where balanced accuracy is defined as the mean of sensitivity and specificity.

2.1 Sample Selection

The data reported in this study were derived from commercial tissue micro-arrays (TMAs) especially prepared for this work. In these TMAs, there were several patients represented more than once. These samples were removed to avoid training and testing of the algorithms by samples from the same patients. In addition, samples of benign lung tumors were derived from the archives of the Department of Pathology, University of Massachusetts Medical School with institutional permission. All samples were from formalin-fixed, paraffin embedded (FFPE) tissue blocks. This report deals with Part I of this study, involving normal and cancerous tissue samples in TMA format. Results for the benign tumors have been submitted for publication as well [15].

The TMAs were assembled to accommodate the goals of this study, which include

a) Distinction of normal (NOR) form cancerous lung tissue
b) Classification of lung cancers into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC)
c) Further classification of NSCLC into adenocarcinoma (ADC) and squamous cell carcinoma (SqCC)
d) Classification of ADC into several sub-classes of clinical relevance

Based on these goals, five TMAs were assembled at US Biomax, Inc. (Rockville, MD) that comprised 80 normal tissue (from cancer patients; biopsy free of cancerous tissue), 29 necrotic tissue from ADC, SqCC and SCLC patients, and 61 SCLC, 89 SqCC and 129 ADC cases for a total of 388 samples. Each tissue spot measured about 1.8 mm in diameter, and will be referred to as “patients” later in this report. These TMA sample numbers LC701 through LC706. From each TMA, three tissue sections were purchased, referred to as section A001, A002 and A003. Section A002 was mounted at Biomax on a standard microscope slide, de-paraffinized, stained and coverslipped. The other two sections were mounted on ‘low-e’ slides (see Section 2.3) and delivered as paraffin-embedded samples.

The number of tissue spots enumerated above included samples from another TMA (LC811, also from US Biomax, Inc.) that were included in this study to increase the total number of samples and to demonstrate that the results obtained were independent of TMA preparation and age of the samples.

2.2 Sample preparation

Slides for spectral data acquisition were de-paraffinized using standard procedure [16] and kept in a desiccator when not used. Original data were collected within a few days after de-paraffination except for the LC811 tissue spots that had been de-paraffinized *ca*. 6 months before repeat data acquisition; however, no discernible degradation of the spectral quality of the older samples was observed.

2.3 Infrared data acquisition and pre-processing

The methods of SHP, including data acquisition and data pre-processing, have been described in detail in the literature [2, 17, 18]. A block diagram of the required steps is shown in Figure 2. All spectroscopic studies reported here were carried out on ‘low emissivity’ (low-e) slides (Kevley Technologies, Chester-

field, OH) that are totally reflective toward infrared radiation, but are nearly totally transparent to visible light; thus, the same sample can be used both for infrared data acquisition and, after appropriate staining, for classical histopathology. The fact that the visible and infrared images were obtained from the same sample permits accurate registration of the two images. This step is necessary for accurate annotation (Section 2.5) of spectral features which, in turn, guarantees the traceability of pixel spectra and annotated tissue features.

Recent reports [19] have pointed out problems of infrared measurements on these reflective slides due to the occurrence of a standing electromagnetic wave on the reflective surface. However, subsequent simulations [20] have shown that the resulting intensity artifact is minimized when using high numeric aperture objectives. Furthermore, by using spectral derivatives rather than raw intensities and regional normalization [17], these effects can be minimized.

Infrared spectral data were acquired from pixels 6.25 μm on edge using a PerkinElmer (Shelton, CT, USA) model SpectrumOne/Spotlight 400 imaging infrared micro-spectrometer, resulting in 25,600 pixel spectra for each square millimeter of tissue. For each pixel, the spectral vector collected covers the wavenumber range from 700 to 4000 cm^{-1}, but only 501 intensity data points between 800 and 1800 cm^{-1} (with 2 cm^{-1} data point spacing) were used for the statistical analysis. The "hyperspectral data cube" contains the coordinate of each pixel, and the infrared spectrum associated with this pixel.

Each tissue spot produced about 10^5 individual pixel spectra that were pre-processed *via* a routine referred to as Cireca_SPP (spectral pre-processor, see Figure 2) to yield pseudo-color images of the tissue spots, as follows. First, the size of hyperspectral data cubes was reduced by a factor of four by co-adding four individual pixel spectra into a new spectrum with better signal-to-noise ratio, but larger pixel size, 12.5 μm on edge. The resulting set of *ca*. 25,000 pixels per tissue spot was corrected for confounding contributions such as noise, water vapor and resonance Mie (R-Mie) scattering (*via* a phase correction algorithm [21]) using procedures developed and reported previously in the literature [17]. In order to enhance the sensitivity of spectral methods toward specific changes of protein abundance, the broad and often unstructured raw spectra were converted to 2^{nd} derivatives. This process is known to reduce the half width of spectral bands, thereby providing better discriminatory power which provides for the ability to classify different tumor types. These second derivative spectra are the primary information obtained in an SHP experiment, and the task at hand is the decoding and correlation of the spectral information with the pathological diagnosis.

This was accomplished by converting the pre-processed datasets to pseudo-color images using unsupervised "hierarchical cluster analysis" (HCA); that is, no input from a pathologist was used in this step. A typical HCA-based pseudo-color image of a tissue spot is shown in Figure 1B. In this Figure, regions of the same color represent similar spectra. Visual inspection of Figures 1A and 1B immediately reveals a correspondence between the IR pseudo-color image and the H&E-stained image. This correspondence becomes even more pronounced at higher magnification of the visual image, and cellular details often can be identified and correlated with the corresponding features of the HCA image. At the four-cluster level shown in Figure 1B, necrosis (gray), adenocarcinoma (blue and green) and endothelium (red) can be clearly distinguished. Regions denoted in the HCA images were correlated to tissue pathology in a step referred to as "annotation", described in Section 2.5 below.

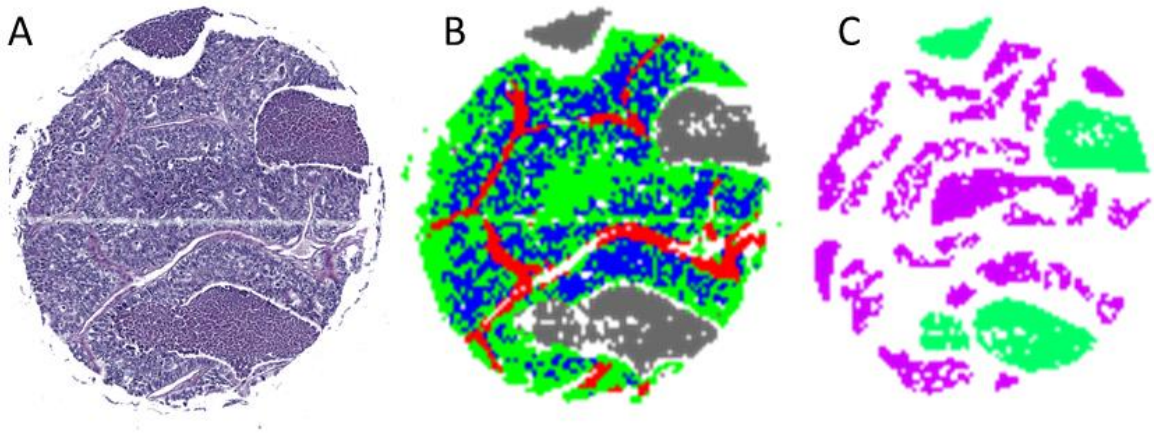

Figure 1. (A) Photomicrograph of an H&E stained TMA tissue stop. (B) Pseudo-color image constructed from the spectral dataset *via* unsupervised hierarchical cluster analysis. (C) Pathology-annotated regions (pathology-spectroscopy correlation). Green: necrosis, purple: adenocarcinoma)

2.4 Visual image acquisition

After infrared data acquisition, the tissue sections were stained at the Department of Pathology at University of Massachusetts Medical School, using hematoxylin/eosin (H&E) and following standardized and validated methods. After coverslipping, the tissue sections were imaged using an Olympus (Center Valley, PA) BX51 microscope equipped with a computer-controlled microscope stage with linear stepping motors (0.1 μm resolution). Images were taken *via* a Qimaging (Surrey, BC, Canada), model QICAM high resolution digital camera. The microscope was operated using Media Cybernetics [Rockville, MD, USA] Image Pro Plus software. The tissue spots were imaged at 20x magnification, producing large mosaic visual image data files at sufficiently high spatial resolution for pathological interpretation. Figure 1A depicts a composite visual image of an H&E stained tissue spot. Comparison of Figures 1A and 1B demonstrates that the tissue types and disease stated detected by infrared imaging directly correspond to the features detected by visual pathology.

Registration of the slide position for visual and infrared microscopy was aided by mounting the slides in a specially designed and manufactured slide holder that was equipped with three reticles whose positions in the particular microscope table were read and recorded at 0.1 μm accuracy.

2.5 Annotation

The annotation process correlates unambiguously assignable tissue areas from the H&E stained visual images with corresponding regions of the pseudo-color infrared image. To this end, software was created (the Cireca_ANNOTATE software of Figure 2) that created a registered semi-transparent overlay of the two images such that tissue features still can be perceived, but are displayed on a color background that corresponds to the HCA clusters. The two images were coarsely registered with respect to each other using the reticle coordinates collected on the infrared and visual microscopes. Subsequently, fine

registration was achieved by optimizing the parameters of a rigid body (image) transform that minimizes the least square error of the difference between the two images [22].

Subsequently, the participating pathologist selected areas of diagnostic interest, for example, the most typical regions of a disease, or normal tissue types. Subsequently, the Cireca_ANNOTATE software checked the selected regions for uniform cluster membership, and eliminated pixels that did not conform to the majority assignment within one selected area. The results of the annotation process for a particular tissue spot are shown in Figure 1C. Here, the four green areas were identified to be due to necrosis, while the purple areas were deemed prototypical for solid mucinous adenocarcinoma. The number of selected areas in Figure 1C is typical for an annotated tissue spot, and yielded, on average, about 1400 pixel spectra for each tissue spot, corresponding to about 350 cells. This latter assessment was based on an estimate of a cell's size (*ca*. 25 μm in diameter) and the aggregated pixel size (12.5 μm on edge).

Pixel spectra from the areas selected by the pathologist were entered into a traceable database of spectra; pixel spectra from different patients but with the same tissue code were combined into tissue classes. Table I summarizes the total number of pixel spectra and those annotated for this study.

Table I. Number of pixel spectra, processed spectra and annotated spectra in entire dataset

Total spectra:	~ 39 x 10^6 (388 spots, ~ 100,000 spectra per spot)
Processed spectra:	4 x 10^6 pixel spectra (2x2 pixel averaging and elimination of blank pixels)
Annotated spectra:	5.5 x 10^5 spectra (see Table II)
Annotation regions:	9.3 x 10^3 annotation regions (see Table II)
Annotation regions/spot:	24 regions/spot (average)
Pixels /annotation region:	60 pixels (average
Main tissue types:	168 (54 malignant, 114 normal classes)

2.6 Data Traceability

In the dataset of *ca*. 550,000 annotated pixels (see Table I), each pixel spectrum is uniquely identified and traceable to the tissue micro-array name (*e.g*., LC706), the particular section (e.g., A001, see Section 2.1), the individual tissue spot identified by row and column (*e.g.*, C 3) and the coordinate of the pixel spectrum. This coordinate was uniquely defined by the pixel X, Y address, and the pixel size. The pixel X, Y address was referenced against the reticle positions in the slide holder (see Section 2.4). Each annotated pixel spectrum, in addition, was tagged with a code that identified the pathology diagnosis. Thus, any pixel spectrum can be relocated and traced, and may be compared to the corresponding region of the visual image that was used for annotation. The export of annotated pixels and their incorporation into the database was accomplished using software referred to as "Cireca_TracDat", see Figure 2.

2.7 Construction of the database

All pre-processed pixel spectra, whether from the annotated areas or from the remainder of the tissue spot, were subsequently stored in a database created by the commercial software "Filemaker Pro" [Filemaker, Inc., Santa Clara, CA] from which spectra could be exported, *via* queries, into datasets that

were used for training and testing of the algorithms. Pixel spectra to be included in a particular dataset could be queried by disease or tissue type code, patient ID, spot position, *etc*. The process of exporting the annotated spectra was carried out be software referred to as 'Cireca_TracDat'. Figure 2 depicts the entire process flow from image acquisition, pre-processing, annotation, and traceable database construction. Note that this process ascertains the complete traceability of each pixel spectrum and its association with a medical code determined by the annotation step. The database serves as an information hub for subsequent classifier training, or creating a prediction report on an unknown sample.

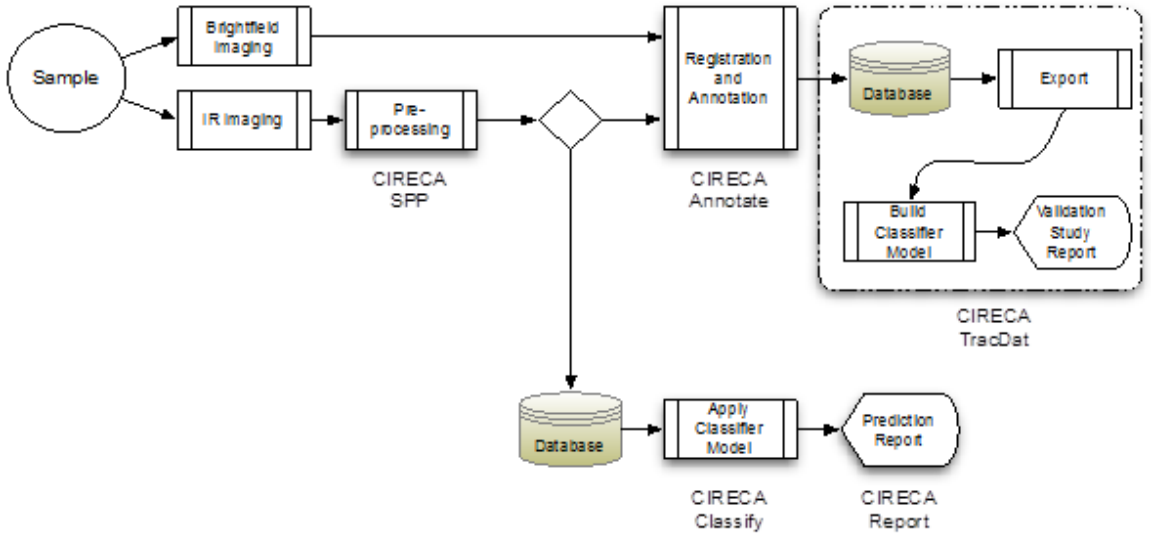

Figure 2. Block diagram of the Cireca process.

3. Computational Aspects

As pointed out above, the main emphasis of this report is the discussion of the statistical methods and metrics employed in this study that demonstrates the classification potential of SHP for lung cancers. Most of the work reported here was carried out on a data subset of 188 patients; the remainder of the dataset (200 patient samples) was kept as a completely blinded test set referred to as the "vaulted" dataset. More diagnostic and prognostic results from the overall study will be reported simultaneously in a medically oriented journal [15].

Thus, the main emphasis of this study was the evaluation of several machine learning algorithms (MLAs, or "machines") that were applied to the non-vaulted dataset, and the development of tests and metrics for the evaluation of sensitivities and specificities of the various procedures. In general, the multivariate statistical tests were carried out in a "pixel-based", a "patient-based" or "image-based" fashion, to be introduced below. The evaluation of the MLAs was carried out mostly on the pixel-based level, since it was argued that this approach afforded the largest datasets (see below) and hereby, the best assessment of the MLAs' accuracies.

3.1 Hardware

All computations were carried out on a Dell workstation equipped with a 12-core Intel processor and 56 GByte memory, running Linux and 64 bit Windows operating systems. This workstation was connected to a server with 20 TByte hard drive space and 100% local and Amazon S3 cloud backup. All computations were carried out in MATLAB version R2013b (The Mathworks, Natick, MA) using scripts developed in house with complete revision history control on an SVN server. The scripts and input data-sets were archived and documented in a work log for each of the "studies" carried out to support complete reproducibility of the work.

3.2 Data subsets for training, testing and evaluation of MLAs.

The composition of the entire dataset is presented in Table II. Notice that the total patient number in Tables II A and II B (as shown in Table II C) exceeds the patient number in the study; this is due to the fact that for most patients, at least two, often three diagnostic classes were rendered, often including necrosis and normal tissue types. Thus, the patient numbers in Tables II A, B and C reflect the number of patients contributing to each class, rather than the total patient number in the study. Subsequently, tissue subtypes with less than 3 patients or less than 50 pixels were dropped from the study.

As pointed out above, the results reported here were obtained from a subset of the entire dataset. This subset contained data from 188 patients, and was further divided into a sub-training and a sub-test set, each comprising *ca.* 125,000 spectra. The evaluation of the MLA performance was carried out on these subsets such that the blinded character of the "vaulted" dataset was maintained. This vaulted dataset originally contained 200 patient samples, and was not used for training and testing of the classifiers which was performed on the sub-training set and sub-test set, respectively. Thus, the vaulted data were not utilized until late in summer 2014 and revealed accuracies very similar to those obtained for sub-training and sub-test datasets. These results have been submitted for publication in a pathology journal [15].

Since even the sub-training and the sub-test datasets contained *ca.* 125,000 pixel spectra, which would require long computation times, smaller benchmark data sets were created. For example, a balanced dataset referred to as the 'benchmark' data set was used to test various MLAs (see below). These datasets were created from the overall dataset by database queries that have been archived. For training and testing, a predetermined number of pixel spectra were randomly selected from the balanced training classes; the optimum number of spectra to be included was optimized experimentally (see Section 3.3.3).

3.3 Computational methods

3.3.1 Benchmark data sets

Although multi-class identification of prokaryotic cells (bacteria) by spectral methods have become common-place, no large multi-classification study has been performed for complex tissues. The overall goals of the study presented here were the reliable distinction between normal (NOR) and diseased tis-

sue, the distinction between necrotic (NECR) and cancerous tissue and between small cell lung cancer (SCLC) and non-small cell cancers (NSCLC). Furthermore, the latter category was to be distinguished into squamous cell lung cancer (SqCC) and adenocarcinomas (ADC) which themselves had several grades and/or subclasses.

Table II A. Number of Annotated Regions, Pixel Spectra and Patients Contributing to Pathological Tissue Classes in Entire Dataset (Pathological Tissue Types Only)

Annotated Regions	Pixel Spectra	Number of patients	Description
1	60	1	Lung/ADC/Acinar/mixed mucinous & non-mucinous
734	35516	39	Lung/ADC/Acinar/mucinous
503	17484	27	Lung/ADC/Acinar/non-mucinous
24	1324	3	Lung/ADC/Colloid/mucinous
1	189	1	Lung/ADC/Lepidic/mixed mucinous & non-mucinous
140	5980	9	Lung/ADC/Lepidic/mucinous
166	6258	8	Lung/ADC/Lepidic/non-mucinous
99	4189	10	Lung/ADC/Micro-papillary/mucinous
58	1056	3	Lung/ADC/Micro-papillary/non-mucinous
18	3279	2	Lung/ADC/Other/non-mucinous
1	142	1	Lung/ADC/Papillary/mixed mucinous & non-mucinous
160	18704	11	Lung/ADC/Papillary/mucinous
68	3419	7	Lung/ADC/Papillary/non-mucinous
503	38335	29	Lung/ADC/Solid/mucinous
355	26476	18	Lung/ADC/Solid/non-mucinous
20	4777	3	Lung/SqCC/Basaloid/grade 1
37	5494	4	Lung/SqCC/Basaloid/grade 2
266	31802	23	Lung/SqCC/Other/grade 1
484	58749	40	Lung/SqCC/Other/grade 2
418	48030	31	Lung/SqCC/Other/grade 3
1465	77685	70	Lung/SCLC/Other/
284	34913	41	Lung/Necrosis/necrotic ADC
322	20913	29	Lung/Necrosis/necrotic SCLC
462	32722	56	Lung/Necrosis/necrotic SqCC
111	5769	17	Lung/Keratin Pearl
22	197	3	Lung/Macrophages/Non-Mucinous
6722	483462	486	26 Pathological Tissue Types

Table II B. Number of Annotated Regions, Pixel Spectra and Patients Contributing to Normal Tissue Classes in Entire Dataset (Normal Tissue Types)

Annotated Regions	Pixel Spectra	Number of patients	Description
45	1602	10	Lung/Conn.Tiss/Black Carbon Pigment
24	2341	8	Lung/Conn.Tiss /Dense CT/Abundance of Fibroblast
4	514	3	Lung/Conn.Tiss /Dense CT/Abundance of Lymphoid Cells
22	925	3	Lung/Conn.Tiss /Loose CT/Abundance of Fibroblast
1	112	1	Lung/Conn.Tiss /Loose CT/Abundance of Lymphoid Cells
2	130	1	Lung/Conn.Tiss /Loose CT/Other
836	23336	59	Lung/Alveolar/Normal histomorphology)
342	5552	43	Lung/Alveolar/Wall thickened by fibrosis
632	12667	56	Lung/Alveolar/Wall with congested capillaries
11	140	4	Lung/Bronchiole/Wall Adventitia
27	481	7	Lung/Bronchiole/Wall Columnar Epithelium
4	151	1	Lung/Bronchiole/Wall Cuboidal Epithelium
8	156	4	Lung/Bronchiole/Wall Muscle
202	3097	37	Lung/Blood/Red Blood Cell (Erythorocyte)
3	107	1	Lung/Blood/WBC - Lymphocytes
63	1854	16	Lung/Blood Plasma/Plasma (with Fibrinogen)
4	35	2	Lung/Blood Plasma/Serum (without Fibrinogen)
92	1751	24	Lung/Blood Plasma/Serum with Blood Cells
60	2363	15	Lung/Blood Vessel/Wall Adventitia
13	287	3	Lung/Blood Vessel/Wall Endothelium
202	7301	41	Lung/Blood Vessel/Wall Muscle

2597	64902	339	21 Normal Tissue Types

Table II C. Total Number of Annotated Regions, Pixel Spectra and Patient Contributions

	Annotated Regions	Pixel Spectra	Number of patients
Pathological Tissue Types	6722	483462	486
Normal Tissue Types	2597	64902	339

Total	9319	548364	825

Table III A. Example of Raw Training and Test Datasets for MLA Comparison (S011)

Annot'ed Regions	Pixel Spectra	Number patients	Description	
Training				
281	12572	15	Lung/ADC/acinar/muc. & non-muc.	
63	2809	4	Lung/ADC/lepidic/muc. & non-muc.	
13	247	4	Lung/ADC/micro-papillary	
33	2762	3	Lung/ADC/papill./muc. & non-muc.	
179	10568	10	Lung/ADC/solid/muc. & non-muc.	Sum ADC= 28958
25	907	5	Lung/keratinpPearl	
20	1647	1	Lung/SqCC/basaloid	
54	4830	5	Lung/SqCC/grade 1	
122	12066	12	Lung/SqCC/grade 2	
83	5942	7	Lung/SqCC/grade 3	Sum SqCC= 25392
Testing				
301	12252	17	Lung/ADC/acinar/muc. & non-muc.	
84	2562	5	Lung/ADC/lepidic/muc. & non-muc.	
24	331	5	Lung/ADC/micro-papillary	
85	5259	5	Lung/ADC/papill./muc. & non-mucs	
196	10400	10	Lung/ADC/solid/muc. & non-muc.	Sum ADC= 30804
9	1124	3	Lung/keratin pearl	
18	3578	2	Lung/SqCC/basaloid	
56	7978	6	Lung/SqCC/grade 1	
104	13850	8	Lung/SqCC/grade 2	
130	12828	8	Lung/SqCC/grade 3	Sum SqCC= 39358

The dataset for this study contained pixel spectra that exhibited enormous spectral changes (*e.g.*, NECR) whereas others exhibited enormously small spectral changes (such as some of the ADC subclasses). In a previous lung cancer pilot study, we used hierarchical binary classifiers based on artificial neural networks, to consecutively classify and remove the most different spectral classes in the dataset and arrived at good classification accuracies [23]. However, the ANNs used displayed a relatively large variation in classification accuracy in consecutive runs; thus, a systematic comparison of various classification algorithms was carried out, discussed next.

Preliminary work [23]had indicated that the classification of NSCLC into SqCC and ADC was the most difficult major classification task; thus, the performance of several MLAs as well as the methods of balancing the datasets, were evaluated for this classification task. To this end, benchmark training and test sets were constructed from the sub-training and sub-test sets. The composition of these datasets is shown in Table III A and B. Here, Table III A is an example of the result of an automatic data base query to select all subclasses/grades of ADC and SqCC. The number of patients, annotation regions and a count of the pixel spectra are given in this Table.

For the evaluation of the various MLAs, it is desirable to have the training dataset balanced both in terms of patients as well as pixels. The (imbalanced) patient numbers in Table IIIA were balanced by randomly eliminating patients until both training and test sets had equal number of patients in all clas-

ses. In the case of odd total patient numbers, the number of patients in the classes was allowed to differ by one, see Table III B.

Table III B. Example of Balanced Training and Test Datasets for MLA Comparison (S011)

Annot'ed Regions	Pixel Spectra	Number patients	Description	
Training				
281	12572	15	Lung/ADC/acinar/muc. & non-muc.	
63	2809	4	Lung/ADC/lepidic/muc. & non-muc.	
13	247	4	Lung/ADC/micro-papillary	
33	2762	3	Lung/ADC/papill./muc. & non-muc.	
169	9829	9	Lung/ADC/solid/muc. & non-muc.	Sum ADC= 28219
14	728	3	Lung/keratinpPearl	
20	2306	1	Lung/SqCC/basaloid	
54	6763	5	Lung/SqCC/grade 1	
85	10100	9	Lung/SqCC/grade 2	
83	8320	7	Lung/SqCC/grade 3	Sum SqCC= 28217
Testing				
234	13187	15	Lung/ADC/acinar/muc. & non-muc.	
78	3013	4	Lung/ADC/lepidic/muc. & non-muc.	
17	346	4	Lung/ADC/micro-papillary	
48	4166	3	Lung/ADC/papill./muc. & non-mucs	
196	13635	10	Lung/ADC/solid/muc. & non-muc.	Sum ADC= 34347
9	1124	3	Lung/keratin pearl	
8	2127	1	Lung/SqCC/basaloid	
49	6316	5	Lung/SqCC/grade 1	
104	13850	8	Lung/SqCC/grade 2	
111	10937	7	Lung/SqCC/grade 3	Sum SqCC= 34354

Eliminating these patients, however, created a discrepancy in the number of pixel spectra in the sub-classes. This was then addressed by oversampling the datasets. Oversampling and undersampling are methods commonly employed in statistical analyses, and will be explained next.

If in a two class problem (classes A and B) the number of samples (spectra) n_A and n_B are different, with $n_A > n_B$, then there are two ways to balance the datasets. In one method, referred to as oversampling, all n_A spectra of group A are used, and data in class B are used repeatedly until $n_A \approx n_B$. In undersampling, spectra are randomly selected from group A such that $n_A \approx n_B$ [24]. Undersampling has the disadvantage that part of the heterogeneity in class A is ignored by the random selection process, whereas in over-sampling, no data are omitted but the smaller dataset is used repeatedly. When SVM classifiers are used for data analysis, oversampling may be visualized as producing more data on one side of the hyperplane separating the data, but these additional data have no effect on the position and orientation of this hyperplane. Omitting data in undersampling, on the other hand, can affect the hyperplane due to reduced sample heterogeneity. Experimentally, it was found that oversampling generally performed better than undersampling; thus, the number of spectra in each class was equalized by oversampling. The resulting

S011 balanced training and test sets are shown in Table III B. Here, the number of patients in the training and test set are equal (except, as noted, for odd numbers of patients) and the number of spectra in the ADC and SqCC classes was equalized by oversampling.

As mentioned above, a prior study used hierarchical ANN classifiers [23] to eliminate classes of higher spectral variance consecutively. The SVMs reported here performed better when used as multi-classifiers rather than as hierarchical, binary classifiers for the benchmark datasets tested. This can be visualized as being due to the fact that the multi-classifier has a broader range of options for classification, whereas each binary classifier has only two options available. Several benchmark datasets were created to allow optimization and testing of MLAs for specific tasks.

3.3.2 Machine-learning algorithms (MLAs) evaluation and optimization

During the course of this study, several types of MLAs were tested for their suitability to classify the spectral data. The 80-patient study published previously [23] used artificial neural networks (ANNs) as the MLA of choice. In subsequent efforts, it was found that – although their average accuracy was slightly higher than that of other MLAs – they proved to be more costly in terms of computer times. This is, in part, due to the fact that ANNs start the analysis with a randomized weight matrix connecting input spectral features with the desired outputs. This randomized initial step led to variations in the output accuracy which required averaging 10 independent trial runs. This step could be avoided using support vector machines (SVMs) that proved to be more stable and reproducible. In addition to SVMs and ANNs, random forest (RF), naïve Bayes (NB) and k-nearest neighbor (KNN) classification algorithms were used on the same benchmark dataset. The overall accuracy of these MLAs, as applied to the benchmark set shown in Table III, is displayed in Table IV. Notice that the MLAs listed here were un-optimized at this point; *i.e.*, the default settings for the MLAs were used.

Although the un-optimized ANN slightly outperformed the SVMs, the latter were selected because of faster execution times, their higher reproducibility, their broad acceptance in the scientific community, and their well-understood mode of action. In ANNs, on the other hand, the mode of action – the connectivity between input and output – is somewhat more random and harder to reproduce.

Table IV. Accuracy of MLAs Used in Benchmark Study

MLA	Accuracy
SVM (lkf)	87.4%
SVM (qkf)	86.8%
ANN	88.1%
RF	85.0%
KNN	78.1%
NB	76.7%

SVM: support vector machine; lkf: linear kernel function; qkf: quadratic kernel function; ANN: artificial neural network; RF: random forest; KNN: k-nearest neighbors; NB: naïve Bayes

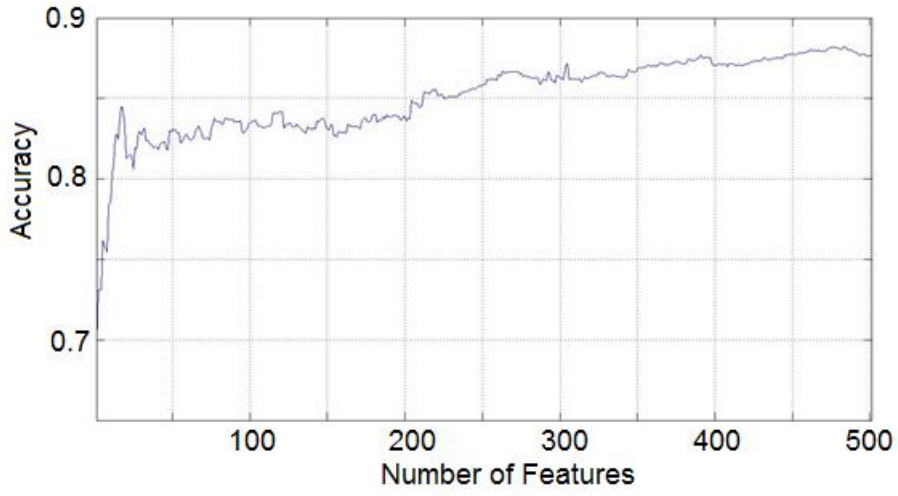

Figure 3. Accuracy of the r-SVM classification as a function of feature number

After further evaluation, a SVM with radial kernel or basis function (rbf) was used for classification. In the case of radial basis function, two parameters, "c" (penalty weight on mis-classification error) and "γ" (width of the radial basis kernel) [25] were optimized by varying them independently from 0.000061 (2^{-14}) to 0.031 (2^{-5}) for γ and 0.0625 (2^{-4}) to 32 (2^{5}) for C. This resulted in an accuracy of 92.4 $\pm$ 0.85 % for a benchmark data set consisting of *ca*. 190,500 training spectra and 48,600 test spectra (the S014 training/test datasets) on a pixel base and 94.0 $\pm$ 2.6 % on the leave-one-out-cross-validation (LOOCV) patient level. The differences in pixel-based and LOOCV computations, as well as the methods to estimate the confidence limits, will be discussed below.

Feature selection [26] sometimes can improve classification accuracy by eliminating confounding features. Thus, feature selection was implemented in the present study using the reduced SVM ("r-SVM") implementation in MATLAB. Here, the number of features was decreased from 501 second derivative intensity points (in the range from 1800 – 800 cm^{-1}, with 2 cm^{-1} data point spacing) to below 50 features, see Figure 3. This process resulted in a gradual reduction of the classification accuracy, with a steep drop-off at about 35 features. The gradual decrease in accuracy between 501 and 100 features suggested that the use of all features in the spectral vectors is advantageous. This result is interesting and deserves further discussion. The spectral resolution of the spectrometers (4 cm^{-1}, or about 2 data points) and the inherent line width of the observed 2nd derivative bands (> 15 cm^{-1}, or about 7 data points) determine that only about 70 data points of the 501 intensity points are linearly independent. Since the accuracy of the MLAs increases above this limit of features, one may conclude that the algorithm detects slight variations in band shapes and uses them for classification. Furthermore, the larger number of features does not decrease the accuracy due to introduction of noise. As expected, the variation (noise) in the plot of accuracy *vs.* number of feature decreases toward higher number of features.

3.3.3 Dependence of accuracy on number of pixel spectra

The dependence of the classification accuracy on the number of included pixel spectra was tested using training and test sets that had sufficient numbers of individual pixel spectra to allow a broad range of inputs. Thus, S011 introduced in Table III B and S014, the entire sub-training and sub-test sets were used, with number of randomly selected spectra varying from 1,000 to 20,000 at constant patient count. These results, summarized in Table V, suggest that a plateau of balanced accuracy is reached at or before 1000 pixel spectra for this dataset, and that increasing this number increases computation time enormously at no gain in accuracy.

Table V. Dependence of Balanced Accuracy on Number of Pixel Spectra Used for Training (S014)

Number of Pixels	Balanced Accuracy (Pixel-based)	Balanced Accuracy (Patient-based)
1000	88.60%	89.30%
2000	90.60%	90.30%
5000	90.00%	90.20%
10000	90.60%	90.60%
15000	90.80%	90.80%
20000	90.80%	90.80%

3.3.4 Confidence intervals (CI)

In statistical analyses, it is common to include confident intervals (CI) defined in terms of the standard error. Here, estimates for the confidence intervals were obtained by two methods. Figure 4b shows the results obtained for a simulation of the confidence limits by varying the number of patients in the training set, at a constant level of input spectra per classification class (2000), and perform 10 independent SVM training and test runs by randomly selected 10,000 training spectra from the entire training dataset. These 10,000 training spectra were randomly selected from 5 classes (NOR, NECR, SCLC, SqCC and ADC) with 2000 spectra per class. The size of the entire training set increased as the number of patients increased, but the number of spectra used in the training set remained constant (at 10,000). As can be seen from Figure 3b, the overall accuracy increases, as expected, as the number of patients in the training set increases, from about 85 to over 90%, and the scatter in the accuracy for 10 independent runs decreased by a factor of about five.

This result can be modeled, using the analytical expression for the sensitivity and standard error. The 95 % confidence interval CI is given by

$$CI = S - 1.96\sqrt{SE} \quad (1)$$

$$SE = \frac{S(1-S)}{n} \quad (2)$$

Here, S is the sensitivity of the measurement, SE the standard error and n is the number of samples. A plot of this confidence interval as a function of sensitivity and number of patients is given in Figure 4a.

The sensitivity (blue line) was estimate from the balanced accuracy from Figure 4b. The green line in Figure 3a denotes the confidence interval. The similarity in shape and magnitude of the predicted confidence interval and the results from the simulations suggests that the dataset faithfully reproduces the variance between patients, and that the selected number of pixels per class (2000) was sufficient to represent each class adequately.

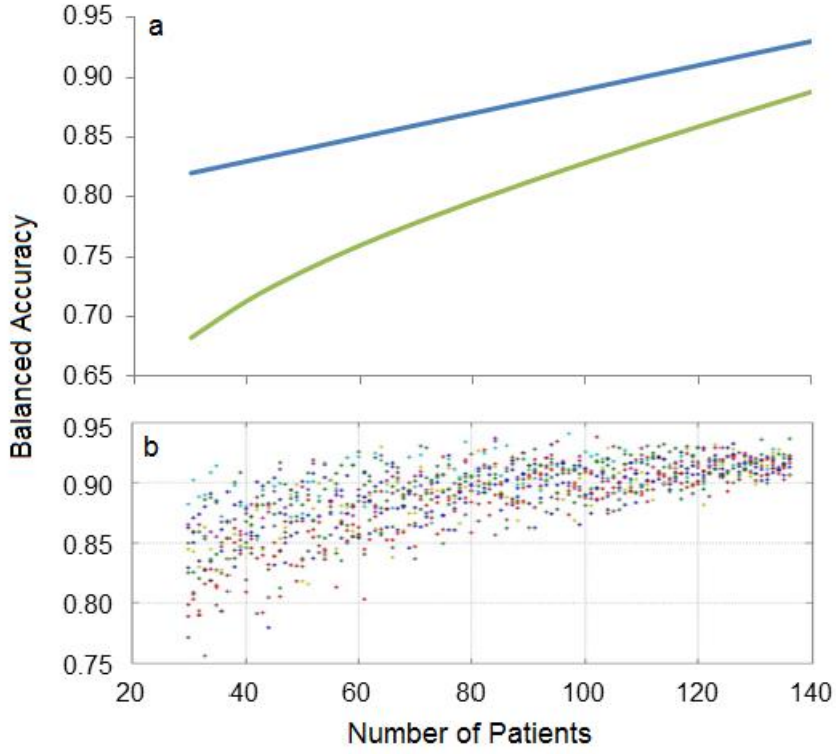

Figure 4 (a). Confidence interval (green trace) for given accuracy (blue trace) and patient number (b) Results of a simulation of the confidence interval and the accuracy of a five-class SVM classifier as a function of the number of patients in the training set.

The results of this simulation also suggest that the annotation method described earlier that often yields hundreds or thousands of individual pixel spectra for each annotated spot produces a representative sampling of tissue homogeneity and patient-to-patient variance. This is in contrast to other cancer diagnostic methods that yield one data point per patient whereas in SHP, thousands of data points are created for each patient. Furthermore, these results indicate that repeated random selection of spectra may reduce the impact of having only one pathologist (in addition to the surgical pathologist at Biomax) review /annotate the samples; that is, that the repeated random selection of pixel spectra produces a heterogeneity in the training set that is comparable in magnitude to the heterogeneity due to different pathologists.

4. Results for sub-training and sub-test sets

As indicated earlier, the sub-test set was analyzed by SVMs trained on the sub-training set. Together, these datasets contained 173 patients, and the composition of the sub-training set is shown in Table VI. As discussed previously, the number of patients listed in Table VI exceeds the number of patient samples since the annotation of most spots resulted in two or three tissue classes. The sub-test set had a

Table VI. Composition of S010 Sub-training Set

Annotated Regions	Number of Pixel Spectra	Number of patients	Description
197	9700	10	Lung/ADC/Acinar/mucinous group
84	2872	5	Lung/ADC/Acinar/non-mucinous
14	879	2	Lung/ADC/Lepidic/mucinous group
49	1929	2	Lung/ADC/Lepidic/non-mucinous
13	247	4	Lung/ADC/Micro-papillary group
10	1383	2	Lung/ADC/Papillary/mucinous group
23	1379	1	Lung/ADC/Papillary/non-mucinous
113	7007	7	Lung/ADC/Solid/mucinous
66	3561	3	Lung/ADC/Solid/non-mucinous
350	15689	17	Lung/SCLC/Other/grade 4
20	1647	1	Lung/SqCC/Basaloid group
54	4830	5	Lung/SqCC/Other/grade 1
122	12066	12	Lung/SqCC/Other/grade 2
83	5942	7	Lung/SqCC/Other/grade 3
108	9785	12	Lung/Necrosis/necrotic ADC
159	11615	13	Lung/Necrosis/necrotic SCLC
140	6903	16	Lung/Necrosis/necrotic SqCC
306	8073	20	Lung/Alveolar/Normal (histomorphologically)
99	1368	18	Lung/Alveolar/Wall thickened by fibrosis
220	4026	19	Lung/Alveolar/Wall with congested capillaries
45	656	12	Lung/Blood Cells group
14	110	3	Lung/Blood Plasma/Plasma (with Fibrinogen)
29	388	10	Lung/Blood Plasma/Serum group
14	357	6	Lung/Blood Vessel/Wall group
78	2238	13	Lung/Blood Vessel/Wall Muscle
6	228	2	Lung/Bronchiole/Wall Columnar Epithelium
9	155	4	Lung/Bronchiole group
25	1171	3	Lung/Connective Tissue/Black Carbon Pigment
26	1439	8	Lung/Connective Tissue group
25	907	5	Lung/Keratin Pearl/Epithelial Pearl
2501	118550		

similar composition in terms of patient numbers. These datasets were analyzed in four different ways, referred to here as Tests 1-4. Test 1 was an overall pixel-based analysis; that is, all pixels from the annotation groups to be classified were combined into five balanced data sets from which pixel spectra were picked randomly for the training of the algorithm. Test 1 was carried out in a cumulative and a non-cumulative manner. In the former, one class, *e.g.*, SCLC, was tested against a combined dataset of all other classes, whereas in the non-cumulative test, the number of pixel spectra and patients remaining after each step decreased by the number of pixel spectra that were classified in the previous step.

Test 2 was based on a leave-one-out cross validation (LOOCV) methodology. Here, data were analyzed on a pixel-base as in Test 1, but the "test set" was the one patient in the dataset not used for training.

Test 3 was implemented in order to assess the accuracy of SHP on a patient-by-patient level against the pathological diagnosis from US Biomax. For this test, rules had to be created that define "agreement" and "disagreement" with the pathological diagnoses. These rules will be discussed below (Section 4.3)

Test 4 finally was a graphic rendering of the SHP diagnosis, as compared to the pathologist's annotation. It is envisioned that such a graphical rendition will constitute the report that will be provided to the pathologist by the Cireca SHP product. Such a report would also include a statistical summary of all pixels in a biopsy.

4.1 Pixel-based results (Test 1)

The results of the non-cumulative pixel-based test are shown in Table VII A. The order of the test procedure is given in the Table, *i.e.,* from Cancer vs. Normal to SqCC vs. ADC. The combined accuracy of all classification steps is 93.2 %

Table VII A. Non-cumulative Pixel-based Classification of Pixel Spectra in the Sub-test set (S010)

Classification	Average accuracy
Cancer vs. Normal	98.3%
SCLC vs. Not SCLC	92.4%
Necrosis vs. Not	94.7%
SqCC vs. ADC	87.5%

The results of the cumulative pixel-based test are shown in Table VII B. Here, the total true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classifications against all other classes are listed for the Sub-test dataset that contained a total of 120,145 pixel spectra. The overall accuracy of the cumulative test, 92.5 %, was similar to the step-wise accuracy. In both cases, the same trend was observed that the classification of cancer *vs*. normal had the highest accuracy, indicating that all disease states *(i.e*., all the cancer classes as well as necrosis) differ spectrally quite significantly from the spectra of normal tissue. These differences often involve changes in the nucleic acid spectral envelopes, as well as a more complex pattern in the amide I region [27]. The classification of necrosis had the second highest accuracy; again, classification can be understood by the occurrence of an additional band in the sec-

ond derivative spectra at ca. 1635 cm^{-1} in the amide I manifold which is a spectral signatures of denatured and precipitated proteins [28, 29].

Table VII B. Cumulative Pixel-based Classification of Pixel Spectra in the Sub-test set (S010)

	Normal	SCLC	Necrosis	SqCC	ADC
true_positives	7665	18510	14771	32653	24021
false_positives	1989	1328	4051	7147	8010
true_negatives	110427	93073	99584	73640	81331
false_negatives	64	7234	1739	6705	6783
accuracy	98.3%	92.9%	95.2%	88.5%	87.7%

SCLC was detected with 92.4 % and 92.9 % accuracy in the non-cumulative and the cumulative tests, respectively. Finally, the discrimination between SqCC and ADC was achieved with 87 % accuracy. This classification is also the most difficult to carry out in classical histopathology, in particular in the case of poorly differentiated carcinomas. Furthermore, these two cancer types can occur as mixed adeno-squamous carcinomas which will aggravate both the annotation process as well as the SHP classification.

Cancer sub-classification of the ADC group was carried out as well. An unsupervised analysis of the mean class spectra of acinar, lepidic, solid, papillary and micro-papillary sub-classes indicated that these classes split according the International Association for the Study of Lung Cancer (IASLC) [30] categories of *low grade* (100 % five year survival), *intermediate grade* (80-90 % five year survival) and *high grade* (60-75 % five year survival). A subsequent SVM classifier for the purpose of discriminating between ADC subtypes achieved an overall accuracy of about 90 %. However, in view of the previous discussion of patient number *vs.* accuracy (Section 3.3.3 and 3.3.4), these results should be viewed as preliminary, and efforts are underway to verify these results with vastly increased datasets.

4.2 Pixel-based LOOCV results (Test 2)

The pixel-based LOOCV was carried out on the entire sub-training set of 173 patients. The (very time consuming) LOOCV yielded classification accuracies, shown in Table VIII that were nearly identical to those reported in Table VII A above. LOOCV is used frequently in medical statistics when the size of the datasets is small since the training of the classifier is being carried out for nearly the entire dataset. The fact that the LOOCV results here are nearly identical to those from the 50:50 split of the dataset indicates that the patient number is sufficient to ascertain a statistically significant result.

4.3 Whole-spot results referenced to the Biomax diagnosis (Test 3)

The previous discussion presented results on a pixel spectrum basis, where the gold standard was the annotation by the pathologist. Thus, only annotated pixels were included in this test, and the annotated pixels, as indicated in Table I, represented a relatively small fraction (ca. 14 %) of the entire data collected. In order to assess whether or not SHP properly diagnosed the majority of each tissue area, another

test procedure was established that used all pixels in a tissue spot and the Biomax diagnosis as the gold standard.

Table VIII. Leave-one-out Cross Validation Results of the Combined Sub-training and Sub-test Datasets

Classification	Average Accuracy	Number of Patients
Cancer vs. Normal	98.60%	174
SCLC vs. Not SCLC	95.50%	134
Necrosis vs. Not	94.20%	99
SqCC vs. ADC	86.40%	99

To determine the performance of the multi-classifier on full samples, all pixel spectra of each of the 94 tissue spots within the sub-testing set were run through the multi-classifier. This produced a classification for every pixel spectrum in a tissue spot, both annotated as well as un-annotated. Each spot was assigned class numbers for the five major classes (1-Normal, 2-SCLC, 3-Necrosis, 4-SqCC and 5-ADC) as follows. A positive class number was assigned to pixel spectra when their SHP prediction agreed with the major Biomax pathology, and a negative class number otherwise. Thus, a negative number (-1) would be assigned to pixels in normal tissue regions in a tissue spot that was diagnoses cancerous by Biomax. Similarly, a negative value (-3) would be assigned to necrotic regions in a cancerous spot when necrosis was not explicitly identified in the Biomax diagnosis, but a number of +3 for an equivalent region of a cancerous spot where necrosis was diagnosed by Biomax. This step resulted in a correlation of all major Biomax diagnoses found in the TMAs with SHP as summarized in Table X.

Table IX Biomax pathology to SHP correlation

Biomax pathology	Classification rule*	Label
Normal (cancer adjacent)	[1]	Normal (no cancers or necrosis)
SCLC	[-1 2 -3 -4 -5]	SCLC with optional normal
SCLC (undifferentiated)	[-1 2 -3 -4 -5]	SCLC with optional normal, necrosis, SqCC or ADC
SCLC with necrosis'	[-1 2 3 -4 -5]	SCLC with optional normal, necrosis, SqCC or ADC
SqCC	[-1 -2 -3 4 -5]	SqCC with optional normal
SqCC with necroosis	[-1 -2 3 4 -5]	SqCC and necrosis with optional normal
ADC	[-1 -2 -3 -4 5]	ADC with optional other classes
ADC with necrosis	[-1 -2 3 -4 5]	ADC and necrosis with optional normal
ADC (mucinous)	[-1 -2 -3 -4 5]	ADC with optional normal
ADC (papillary)	[-1 -2 -3 -4 5]	ADC with optional normal

*at least one positive classification is required, negative ones are optional

The following rules for whole spot analysis were established:

- A threshold value of pixels (see below) was required before any class was considered significant.
- At least one positive class value was required to be present for a given tissue spot; that is, the tissue class corresponding to the major Biomax diagnosis had to be represented in the SHP result with a pixel number above the threshold value.
- Classes with negative values were optional and could be present within the spot without triggering a "non-match" (for example, normal areas in a cancerous spot); however, if the number of pixel spec-

tra in one of the 'negative' classes exceeded the threshold and the primary 'positive' diagnostic class, a "non-match" was recorded for that spot.

The threshold values were applied to all classes, and varied between 200 and 2000 pixel spectra. For a threshold of 400 pixel spectra, an optimum performance of the method was realized. At the spatial resolution of the processed data, 400 pixels correspond to approximately 100 cells in an area of *ca.* 200 μm x 200. This corresponds roughly to the field of view of a visible microscope at 20X magnification, and 100 cancerous cells in the field of view would certainly trigger a pathologist's response as to the presence of cancer. A lower threshold would increase the sensitivity of the SHP classification at the expense of the specificity. The results of this whole spot analysis are presented in Table X.

Table X. Results of the whole spot classification

	Normal	NOT Normal
Normal	45	2
NOT Normal	0	47
sensitivity	95.7%	100.0%
specificity	100.0%	95.7%
accuracy	97.9%	97.9%

	SCLC	NOT SCLC
SCLC	22	1
NOT SCLC	0	71
sensitivity	95.7%	100.0%
specificity	100.0%	95.7%
accuracy	98.9%	98.9%

	Necrosis	NOT Necr.
Necrosis	37	4
NOT Necr.	0	53
sensitivity	90.2%	100.0%
specificity	100.0%	90.2%
accuracy	95.7%	95.7%

	SqCC	NOT SqCC
SqCC	41	0
NOT SqCC	0	53
sensitivity	100.0%	100.0%
specificity	100.0%	100.0%
accuracy	100.0%	100.0%

	ADC	NOT ADC
ADC	53	1
NOT ADC	0	40
sensitivity	98.2%	100.0%
specificity	100.0%	98.2%
accuracy	98.9%	98.9%

4.4 Label-image based whole spot analysis (Test 4)

The graphic depiction of classification results is of enormous importance, since it may be used by a pathologist as an ancillary tool for the diagnosis of some biopsies that are not uniform, or do not present a clear answer. Figure 5 shows the results of three different label-images from the test set for different necrosis and cancer classes. The left row panels show the H&E stained images of the tissue spots, the second row panels the regions annotated by the pathologist, depicting the appropriate diagnosis by color code (see below), and the third column the SHP prediction, depicted in the same color code. The right-most column depicts false positive and false negative SHP diagnoses in red. The color code utilized in Figure 4 is as follows: green; necrosis; yellow: SCLC; blue: SqCC purple: ADC; red: misclassifications. The results presented in Figure 5 are typical for those obtained for the label-image approach, and the quality of agreement represents the majority of the spots analyzed. It should be re-emphasized that the results in the third column, *i.e.*, the SHP prediction on a pixel-basis was obtained in a blinded fashion: the datasets of the spots analyzed were pre-processed and annotated such that a true-false decision of

the SHP results could be rendered; however, these spots were not used in the training of the classifier at all, but were analyzed in a blinded manner. Thus, the agreement shown between the second and the

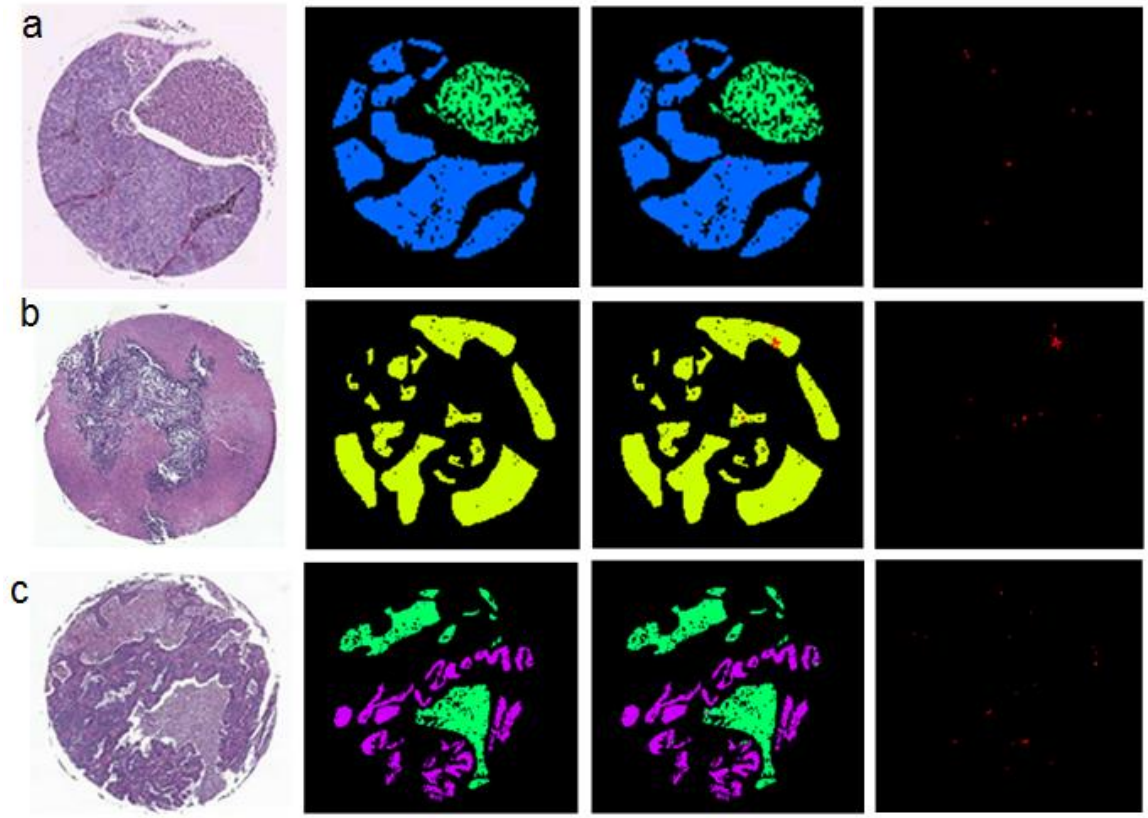

Figure 5. Row (a): Tissue spot with necrosis (green) and squamous cell carcinoma (blue). Row (b): Tissue spot with small cell carcinoma (yellow). Row (c): Tissue spot with adenocarcinoma (purple) and necrosis (green). Red regions: false positive and false negative misclassification. See text for detail

third column of images truly represents the ability of the classifier to distinguish the different cancer types. It also should be noted that the misclassifications, shown both in the third and fourth column images, occur mainly at edges of tissues and thus, are most likely due to poor signal quality of the pixel spectra. There are, of course, a few tissue spots where the US Biomax pathology and the SHP prediction disagree; in these, it is always the cancer type, and not the presence or absence of cancer, which is not properly predicted.

Figure 6B shows an overlay of the annotation regions selected by the pathologist (bright green), and the whole spot label images obtained by SHP. This image demonstrates the agreement between the regions selected by the pathologist and the overall regions of disease. Figure 6A shows the corresponding HCA cluster image that was used by the pathologist in the annotation process. The homogeneity of the annotation regions selected by the pathologist critically depends on the overlay of the visual and HCA images, and Figure 5 demonstrates that the annotation procedure developed for the establishment of the databases produces highly specific and homogeneous databases. This particular image demonstrates why the SHP results seem to be more accurate than expected from an accepted level of diagnostic accuracy

of classical histopathology: The algorithms reported here were trained on datasets that contained very carefully selected regions of disease upon which more than one pathologist agreed. The selected regions may be considered quintessential for a disease type/stage and thus, may enable an MLA to detect specific, recurring features in these typical spectra from other samples.

5 Results of blinded dataset

The vaulted dataset of a total of 194 patients yielded excellent accuracy as well. When this dataset was analyzed by an algorithm trained on the entire training set using the Test 3 procedure introduced above, only 9 patients were misclassified (1 false positive and 8 false negatives). Among the false negatives, there were 3 patients for whom the cancer was predicted correctly, but the necrotic co-diagnosis was missed by SHP. Thus, there were only 5 patient samples out of 194 in which SHP produced an incorrect cancer diagnosis; this corresponds to an overall cancer detection accuracy of 97 %. In all the five false negative diagnoses, SHP detected cancer, but not the cancer diagnosed at the US Biomax level: 3 patients diagnosed by Biomax as ADC with necrosis were assigned by SHP to SCLC classes, one SqCC with necrosis patients was assigned to ADC with necrosis (with a smaller percentage of SqCC) by SHP, and one SqCC sample was assigned as ADC. A final re-evaluation of these misdiagnosed cases is pending.

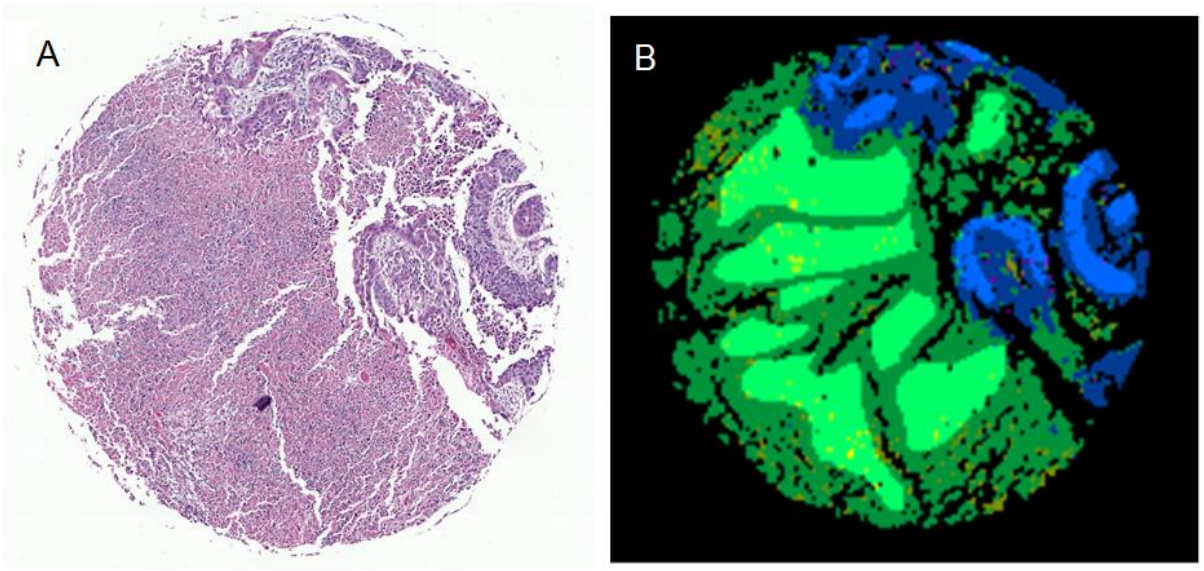

Figure 5 (a): Visual photomicrograph of an H&E stained tissue spot diagnosed as SqCC with necrosis.(b) SHP label image (pale blue and green) and annotation regions (bright green and blue). This image demonstrates the excellent selection of the annotation regions within areas of a particular diagnostic attributes.

To our knowledge, this is the first report of a large, truly blinded dataset analyzed by SHP, and the first large dataset classified into multiple disease classes. The overall accuracy achieved in this study is indicative of the sensitivity of SHP in detecting and classifying disease by a snapshot of the overall chemical composition of various cancers. A more detailed review of the analysis of the entire dataset have been submitted simultaneously to a more medically oriented journal [15].

6 Conclusions

The results presented in this paper summarize the statistical protocols that were developed for the analysis of the lung data set described here. Subsequently, the vaulted (blinded) dataset was analyzed and the prediction accuracy of SHP was found to be comparable to that reported here for the sub-training and sub-test datasets (*cf.* Section 5).

The studies of this report, as well as the results of the blinded dataset that also included benign lung tumors, represent the first SHP-based analyses of a large patient number investigation carried out following standard biomedical statistical procedures of strictly divided and balanced training and test sets, optimization of the statistical procedures, and analysis on a patient-by-patient basis. This study also reported, for the first time, a complex mixture of normal, cancer-adjacent, necrotic, and benign and malignant tumor tissue types found in a pathologically realistic setting. The SHP-based discrimination of normal tissue types and benign lesions from cancer was achieved with very high accuracy, as was the detection of necrosis and SCLC. The lowest accuracy was observed for the classification of SqCC and ADC, two cancer types that are known for their heterogeneity. Furthermore, these two carcinomas are known to occur frequently as mixed type cancers; thus, the lower classification accuracy of SHP could well be due to the inherent similarity of the two cancer types.

Acknowledgement

Initial aspects of the work, in particular, the development of SHP methodology, were partially funded by a grant from the National Cancer Institute, CA 111330, to MD.

References

1. Diem, M., et al., *Molecular pathology via Infrared and Raman spectral imaging*, in *Ex-vivo and in-vivo Optical Pathology*, M. Schmitt and J. Popp, Editors. 2013, Wiley-VCH.
2. Diem, M., et al., *Applications of Infrared and Raman Micro-spectroscopy of Cells in Medical Diagnostics: Present Status and Future Promises.* Spectroscopy – Biomedical Applications, 2012. **27**(5-6): p. 463-496.
3. Diem, M., *Introduction to Modern Vibrational Spectroscopy* 1993, New York: Wiley-Interscience.
4. Diem, M., et al., *IR Spectroscopic Imaging: from Cells to Tissue*, in *Spectrochemical Analysis using Infrared Multichannel Detectors*, R. Bhargava and I.W. Levin, Editors. 2005, Blackwell Publishing: Sheffield (UK). p. 189-203
5. Diem, M., et al., *A decade of vibrational micro-spectroscopy of human cells and tissue (1994-2004).* Analyst, 2004. **129**(10): p. 880-5.
6. Lasch, P., et al., *Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis.* Biochim Biophys Acta, 2004. **1688**(2): p. 176-86.
7. Bassan, P., et al., *Automated high-throughput assessment of prostate biopsy tissue using infrared spectroscopic chemical imaging.* Proc. of SPIE 2014. **9041**.
8. Bhargava, R., Fernandez, D. C., Schaeberle, M. D., Levin, I. W. *FTIR Imaging of Biological Tissue for Histopathological Analysis*. in *PittCon*. 2002. New Orleans.
9. Pounder, F.N. and R. Bhargava, *Toward Automated breast Histopathology Using Mid-IR Spectroscopic Imaging*, in *Vibrational Spectroscopic Imaging for Biomedical Appications*, G. Srinivasa, Editor. 2010, McGraw Hill: New York.
10. Ly, E., et al., *Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies.* Analyst, 2008. **133**: p. 197-205.
11. Bird, B., et al., *Detection of Breast Micro-Metastases in Axillary Lymph Nodes by Infrared Micro-Spectral Imaging.* Analyst, 2009. **134**: p. 1067-1076.
12. Hackett, M.J., et al., *Chemical alterations to murine brain tissue induced by formalin fixation: implications for biospectroscopic imaging and mapping studies of disease pathogenesis.* Analyst, 2011. **136**: p. 2941-2952.
13. Holton, S.E., M.J. Walsh, and R. Bhargava, *Subcellular localization of early biochemical transformations in cancer-activated fibroblasts using infrared spectroscopic imaging.* Analyst, 2011. **136**: p. 2953-2958.
14. Krafft, C., et al., *Analysis of human brain tissue, brain tumors and tumor cells by infrared spectroscopic mapping.* Analyst, 2004. **129**(10): p. 921-925.
15. Akalin, A., et al., *Classification of Malignant and Benign Tumors of the Lung by Infrared Spectral Histopathology (SHP).* Laboratory Investigations, 2015. **in press**.

16. foliobio.com. *Deparaffin Protocol*. Available from: http://foliobio.com/media/pdf/Deparaffin%20protocol.pdf.

17. Miljković, M., et al., *Spectral Cytopathology: new aspects of data collection, manipulation and confounding effects.* Analyst, 2013. **138**: p. 3975-3982.

18. Diem, M., et al., *Molecular pathology via Infrared and Raman spectral imaging.* J.Biophotonics, 2013. **6**(11-12): p. 855-886.

19. Bassan, P., et al., *The inherent problem of transflection-mode infrared spectroscopic microscopy and the ramifications for biomedical single point and imaging applications.* Analyst, 2013. **138**: p. 144-157.

20. Wrobel, T.P., et al., *Electric field standing wave effects in FT-IR transflection spectra of biological sections: Simulated models of experimental variability.* Vibr.. Spectrosc., 2013. **69**: p. 84-92.

21. Miljković, M., B. Bird, and M. Diem, *Dispersive line shape effects in infrared spectroscopy.* Analyst, 2012. **137**: p. 3954-3964.

22. Remiszewski, S., et al., *A multi-centre, multi-platform spectral histopathology (SHP) study for lung cancer diagnostics and prognostics* Analyst, 2015. **submitted**.

23. Bird, B., et al., *Infrared Spectral Histopathology (SHP): A Novel Diagnostic Tool for the Accurate Classification of Lung Cancers.* Laboratory Investigations, 2012. **92** p. 1358-1373.

24. Rahman, M.M. and D.N. Davis, *Addressing the Class Imbalance Problem in Medical Datasets.* Intern. J. machine Learning Computing, 2013. **3**(2): p. 224-228.

25. Statnikov, A., et al. *A Gentle Introduction to Support Vector Machinesin Biomedicine*. 2009.

26. Zhang, X., et al., *Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data.* BMC Bioinformatics, 2006. **7**(1): p. 197.

27. Diem, M., et al., *Infrared and Raman Spectroscopy and Spectral Imaging of Individual Cells*, in *Infrared and Raman Spectroscopic Imaging* R. Salzer and H.W. Siesler, Editors. 2014, Wiley – VCH Publishing,: Weinheim, Germany. p. 181-223.

28. Lasch, P., et al., *Antemortem Identification of Transmissible Spongiform Encephalopathy (TSE) from Serum by Mid-infrared Spectroscopy*, in *Vibrational Spectroscopy for Medical Diagnosis*, M. Diem, P.R. Griffiths, and J.M. Chalmers, Editors. 2008: Chichester, UK. p. 97-122.

29. Kurouski, D., et al., *Is Supramolecular Filament Chirality the Underlying Cause of Major Morphology Differences in Amyloid Fibrils?* J.Amer.Chem.Soc., 2014. **136**: p. 2302−2312.

30. Travis, W.D., et al., *International Association for the Study of Lung Cancer / American Thoracic Society / European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma.* J.Thoracic Oncology, 2011. **6**(2): p. 244-285.