Directional Kernel Density Estimation for Classification of Breast Tissue Spectra

Arturo Pardo*, Eusebio Real, Venkat Krishnaswamy, José Miguel López-Higuera, *IEEE Senior,* Brian W. Pogue, and Olga. M. Conde

***Abstract*— In Breast Conserving Therapy, surgeons measure the thickness of healthy tissue surrounding an excised tumor (surgical margin) via post-operative histological or visual assessment tests that, for lack of enough standardization and reliability, have recurrence rates in the order of 33%. Spectroscopic interrogation of these margins is possible during surgery, but algorithms are needed for parametric or dimension reduction processing. One methodology for tumor discrimination based on dimensionality reduction and nonparametric estimation – in particular, Directional Kernel Density Estimation –is proposed and tested on spectral image data from breast samples. Once a hyperspectral image of the tumor has been captured, a surgeon assists by establishing Regions of Interest where tissues are qualitatively differentiable. After proper normalization, Directional KDE is used to estimate the likelihood of every pixel in the image belonging to each specified tissue class. This information is enough to yield, in almost real time and with 98% accuracy, results that coincide with those provided by histological H&E validation performed after the surgery.**

***Index Terms* — Surgical guidance/navigation, breast, dimensionality reduction, Image reconstruction, machine learning, pattern recognition and classification, probabilistic and statistical methods, quantification and estimation, ROC analysis, segmentation.**

This paper was reviewed and sent on xxxx. Research reported in this paper was funded by projects DA2TOI (codename FIS 2010-19860), FOS4 (codename TEC 2013-47264-C2-1-R) and an undergraduate Research Assistant Fellowship (Beca de Colaboración) entitled "Multispectral enhancement systems for tissue diagnosis in oncology and cardiovascular medicine", the latter granted to the main author by the Spanish Ministry of Education, Culture and Sports.

*Arturo Pardo is with the Photonics Engineering Group, Universidad de Cantabria, Universidad de Cantabria, Santander -- 39006 Cantabria, Spain (e-mail: pardofrancoa@unican.es).

Eusebio Real is with the Photonics Engineering Group, Universidad de Cantabria, Universidad de Cantabria, Santander -- 39006 Cantabria, Spain (e-mail: eusebio.real@unican.es).

Venkat Krishnaswamy is with the Department of Surgery in the Geisel School of Medicine, Hanover NH 03755 USA (e-mail: venkataramanan.krishnaswamy@dartmouth.edu)

José Miguel López-Higuera is with is with the Photonics Engineering Group, Universidad de Cantabria, Santander -- 39006 Cantabria, Spain (e-mail: lopezhjm@unican.es).

B. W. Pogue is with the Department of Surgery in the Geisel School of Medicine, Hanover NH 03755 USA (e-mail: Brian.W.Pogue@dartmouth.edu).

Olga M. Conde is with the Photonics Engineering Group, Universidad de Cantabria, Santander -- 39006 Cantabria, Spain (e-mail: olga.conde@unican.es).

I. Introduction

Breast Conserving Therapy refers to surgical procedures on breast cancer patients where only malignant tissues are removed keeping the resected margin to just the malignantly defined regions from radiologic imaging. In particular, onco-cosmetic surgeries such as lumpectomies are a much more moderate way to extract localized carcinomas as compared to more extensive diseases which might require full mastectomy. One issue with this less invasive surgery though is that the way surgeons assess whether or not the intervention was successful is through careful evaluation of the *surgical margins* present in the tumor after surgery. During the extraction procedure, the practitioner removes the tumor with a layer of healthy tissue surrounding it. It is the thickness of this layer that receives the name 'surgical margin', and this thickness measurement provides insight about whether or not the procedure went well. These margins must be tumor-free; otherwise it is said that the tumor has a *positive margin*, and it is very likely that some cancer is left inside the intraoperative cavity. About 20-40% of all BCT surgeries fall under this classification and patients whose tumors have positive margins will have a high likelihood of undergoing surgery again [1 – 3].

The main problems a surgeon faces when assessing surgical margins are twofold, related to the margin thickness and evaluation methodology. Firstly, there is no single agreed standardization on how thick the surgical margin must be. Although there have been several studies analyzing the long-term consequences of performing BCT procedures with higher or lower thickness standards, in practice surgeons leave a 5 mm thick surgical margin when extracting invasive carcinomas and at least 10 mm for *in situ* cases [3]. Secondly, surgeons rely on their visual acuity and palpation when evaluating a surgical margin, since intrasurgical margin evaluation methods – such as Frozen Section Analysis (FSA) or preparation cytology – either take longer than the surgical procedure to provide a viable result, or are not precise or reliable enough. Depending on the method, the complete assessment may last up to 30 minutes, if there is a pathologist available to begin the procedure in the operating room [1].

Therefore, the main objective of this study was to find an

assessment tool with which surgeons may reliably support their clinical decisions quickly, easily and in a non-invasive way, so they can avoid the risk of closing the incision when they could continue the procedure and extract any remaining cancerous tissue still remaining inside. Current state-of-the-art spectroscopy-based classification procedures make use of signal processing methods such as k-Nearest Neighbors classification [1, 4], Principal Component Analysis (PCA), alone or combined with Independent Component Analysis [5]. The conundrum is that these algorithms that provide high reliability often require a large database of cases with similar characteristics – which cannot ensure to be sufficient for every single, specific case – and, at the same time, blind separation techniques are highly time-efficient, but not as reliable.

In this article, we propose a different approach: we are not attempting to find a model that properly fits the spectral characteristics of every case and sample, but instead, we seek information about what makes every spectrum different from the rest in each particular scenario. Here, multivariate estimation comes to good use. It does require some kind of training, but not with respect to other tissue samples; regions within the sample itself are used to establish which spectral characteristics are differentiable in each particular case, only to then classify the whole picture, providing a quick, simple and reliable image of spectral differentiability in any tissue. Five steps are required to achieve this, namely (a) calculating the Spectral Normal Variate; (b) applying Singular Value Decomposition to the data matrix; (c) preserving directional data; (d) finding the likelihood of every pixel to be of either malignant or non-malignant tissue by finding a Directional Kernel Density Estimate (d-KDE) of the directionality that represents cancerous and non-cancerous tissue, using small subsets of hyperspectral pixels from the image pinpointed by the surgeon; and (e) classifying every pixel as either healthy or malignant according to that estimate. Finally, classification results need to be represented in a user-friendly way. For that exact purpose, two methods based on hyperspectral-to-RGB transformation techniques are described as well. The surgical team may then use this graphical information to perform the excision quickly and reliably, without having to take care of the system itself.

II. Materials and Methods

A. The hyperspectral imaging platform

In order to acquire hyperspectral images of excised samples, Krishnaswamy et al. [6] devised at the Thayer School of Engineering at Dartmouth College an effective spectrum retriever with raster scanning capabilities. This device is composed of two parts, namely a raster-scanning platform and a confocal spectroscopy setup that exposes the sample to white light, and then retrieves backscattered spectra. Figure 1 depicts a simplified schematic of this setup.

White light produced by a tungsten-halogen lamp (HL) – which was coupled to a 50 μm fiber (F1) – was aimed at an achromatic lens (L1), which then directed most power through a beam splitter (BS). On top of BS, a second lens (L2) was located to properly focus light at the sample with a focal spot not greater than 100 μm. A moving transparent platform (XY) was designed to provide displacement in a plane normal to the focal axis. To avoid specular reflections coming from the platform, XY was rotated 45 degrees with respect to the focal axis of L1 and L2. Backscattered light then returned from the sample, reaching the beam splitter and was thus diverted to lens L3, which was optically coupled to a fiber (F2); this fiber was connected to a CCD-based spectrometer (SPEC), calibrated in the 510-785 nm range with a spectral resolution of 1 nm. A computer (COMP) controlled the XY location and stored data provided by the SPEC [1,6].

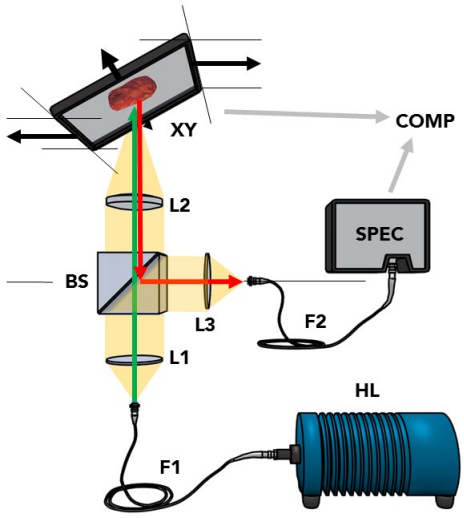

Fig. 1. Confocal microscopy setup designed by Krishnaswamy et al. [6]

Additionally, this setup required proper calibration, by means of referencing to a spectrally flat material. Reflectance values were obtained by normalizing the received spectrum with reference to the whole setup response:

$$R(\lambda) = \frac{I_{\text{meas}}(\lambda) - I_{\text{bg}}(\lambda)}{I_{\text{ref}}(\lambda) - I_{\text{bg}}(\lambda)} \quad (1)$$

where all variables are measurements of light intensity: $I_{\text{meas}}(\lambda)$ was the light intensity reflected by the current sample under analysis, $I_{\text{bg}}(\lambda)$ was the background spectrum (found by taking a measurement without illumination), and $I_{\text{ref}}(\lambda)$ was the reference spectrum, acquired by placing *Spectralon* (Labsphere, Inc., North Sutton, New Hampshire) on XY and taking a proper measurement of the intensity spectrum that reached the spectrometer.

B. Breast tissue specimens and Regions of Interest

Here, Laughney et al.'s tissue database [1, 2] was employed to test the overall classification performance of the proposed algorithm. It is composed of 29 imaged samples of about $10 \times 10 \times 4$ mm of volume. These samples obtained by the

Department of Pathology at DHMC were imaged with the imaging platform described in Section II.A, following a clear-cut protocol that would minimize tissue degradation and damage [1]. Once the image was taken, a pathologist used standard histological analysis procedures with the intention of seeking up to seven different tissue categories in the sample, namely (a) normal tissue, (b) benign tissue (i.e. benign tumoral tissue), (c) ductal carcinoma in situ (DCIS), (d) invasive ductal carcinoma (IDC), (e) invasive lobular carcinoma (ILC), (f) inflammation and (g) adipose tissue. For every tissue type found in a sample, a binary mask (or region of interest, ROI) was included to the corresponding sample profile in the database, providing an expert description of every tissue class certainly present in each extracted sample at specific locations. This description will be the basis on which to support the results of our classifier.

C. Finding spectral directionality

1) Spectral Normal Variate

The imaging system described in the previous section took a continuous spectrum – represented by an unknown, continuous real-valued function $R(\lambda)$ of real variable λ – and sampled it at discrete wavelength numbers with enough resolution to recover relevant properties via interpolation. The first step in this method was to remove multiplicative variations in reflectance due to differences in sample particle size, path length, substance concentration and/or thickness, and focus on the spectral properties of any given dataset. The Spectral Normal Variate serves this purpose well in diffuse reflectance spectroscopy [7, 8]. Given a spectrum $\boldsymbol{r}_k \in \mathbb{R}^m$, its SNV can be easily found with the expression

$$\boldsymbol{g}_k = \frac{\boldsymbol{r}_k - \mu_k}{\sigma_k}, \tag{2}$$

where μ_k and σ_k are the sample average reflectance and the sample standard deviation of the reflectance vector elements, respectively. This transformation allowed the expression of every spectrum $\boldsymbol{r}_k$ as reflectance variations of a pixel with respect to its average reflectance, in standard deviation units.

2) Singular Value Decomposition of the data matrix

Every corrected spectrum would then be a vector in a high-dimensional space, and further calculations require the usage of a dimensionality reduction procedure to deal with a smaller amount of data per pixel. The SVD of a real matrix $A \in \mathbb{R}^{p\times m}$ is the factorization

$$A = U\Sigma V^T, \tag{3}$$

where U and V are orthogonal matrices, whose columns are referred to as the left-singular and right-singular vectors of A, respectively, and Σ is a diagonal matrix whose nonzero elements are referred to as the *singular values* of A. If we stack all spectra $\boldsymbol{g}_1, \dots, \boldsymbol{g}_p \in \mathbb{R}^n$ as row vectors in a matrix G, the decomposition $G = U\Sigma V^T$ can be written as a sum of orthogonal matrices of rank one, namely

$$G = E_1 + E_2 + \cdots + E_r, \tag{4}$$

where $r = \text{rank}(G) \leq n$, and $E_i = u_i \sigma_i v_i^T$, $E_j E_k = 0$, $\forall j \neq k$. This sum can be truncated, obtaining an approximation $\tilde{G}_L = E_1 + \cdots + E_L$, with $L \leq r$. This approximation has an error that is known and can be found by finding the Frobenius norm of the matrix difference [9]

$$\left\| G - \tilde{G}_L \right\|_F = \sigma_{L+1}. \tag{5}$$

Moreover, this truncation makes it possible to express every spectrum $\boldsymbol{g}_k$ as a linear combination of the first $L \leq r$ right-singular vectors of G

$$\boldsymbol{g}_k \approx \sigma_1 u_{k1} \boldsymbol{v_1} + \sigma_2 u_{k2} \boldsymbol{v}_2 + \cdots + \sigma_L u_{kL} \boldsymbol{v}_L, \tag{6}$$

which in turn implied that every pixel could be expressed by its coordinates in the lower-dimensional subspace defined by the first right-singular vectors of G, i.e. $\boldsymbol{a}_k = (a_{k1}, \dots a_{kL})$. When selecting an appropriate value for L, two methods were proposed. L was either assigned a constant value (for instance, $L = 10$), or chosen dynamically such that the contribution of the L-th singular value to the sum of the $L-1$ elements in the diagonal of Σ is lower than a constant value [5]:

$$\sigma_{\text{contrib}}(k) = \frac{\sum_{i=1}^{k} \sigma_i^2 - \sum_{i=1}^{k-1} \sigma_i^2}{\sum_{i=1}^{k-1} \sigma_i^2}, \tag{7}$$

being σ_i the i-th element in the diagonal of Σ. In our case, for each image, a value of L for which $\sigma_{\text{contrib}}(L) < 0.01$ holds was deemed an appropriate value.

3) Vector normalization

Different spectra will be considered from now on as spectra represented by non-proportional vectors in the lower-dimensional space. Thus, the magnitude of every vector can be ignored, and then dividing every vector in the lower-dimensional space by its norm with

$$\boldsymbol{x}_k = \frac{\boldsymbol{a}_k}{\|\boldsymbol{a}_k\|} \tag{8}$$

leaves only the directional information of every spectrum in the lower-dimensional space.

4) Directional Kernel Density Estimation

Now that only the *direction* of every sampled spectrum is taken into account, it seems appropriate to find a way to quantify the orthogonality of every spectrum with respect to a set of labeled spectra in the same lower-dimensional space. In this case, two classification hypotheses were defined, so that

every spectrum was to be classified as either $H_0 \coloneqq$ "Healthy tissue" or as $H_1 \coloneqq$ "Malignant tissue". In order to arrive to a classification rule, the *conditional directional probability densities* of any vector $\boldsymbol{x} : \|\boldsymbol{x}\| = 1$ given that it is healthy tissue ($\hat{f}(\boldsymbol{x}|H_0)$) and given that it belongs to the malignant tissue class ($\hat{f}(\boldsymbol{x}|H_1)$) must be estimated. A set of sample vectors $\boldsymbol{X}_{1,0}, \boldsymbol{X}_{2,0}, \dots, \boldsymbol{X}_{n,0}$ from the hyperspectral image that are known to be of healthy tissue, and another subset of sample vectors $\boldsymbol{X}_{1,1}, \boldsymbol{X}_{2,1}, \dots, \boldsymbol{X}_{n,1}$ known to represent cancerous tissue are used to estimate each directional PDF as follows:

$$\hat{f}_h(\boldsymbol{x}|H_k) = \frac{c_{h,L}(K)}{n} \sum_{i=1}^{n} K\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_{i,k}}{h^2}\right), \tag{9}$$

with $\boldsymbol{x},\ \boldsymbol{X}_{i,k} \in S^{L-1}$, for all values of i, and $k = 0, 1$. Here, $K(.)$ is known as the *directional kernel*, $h > 0$ is a constant usually referred to as the *estimator bandwidth*, and $c_{h,L}(K)$ is a normalization constant, which is dependent of the selected kernel $K(.)$ and parameters bandwidth h and dimension L. It must be noted that, albeit $\boldsymbol{X}_{1,i}, \boldsymbol{X}_{2,i}, \dots, \boldsymbol{X}_{n,i}$ are referred to as *training vectors*, there is no training taking place: vector $\boldsymbol{x}$ in (9) simply corresponds to the direction at which the estimate of each probability density function is evaluated, by using $\boldsymbol{X}_{1,i}, \boldsymbol{X}_{2,i}, \dots, \boldsymbol{X}_{n,i}$ for each hypothesis $i = 0,1$. The kernel function must be a non-negative function defined in $\mathbb{R}^+$ that satisfies

$$0 \leq \int_0^\infty K(v)\, v^{\frac{k-3}{2}}\, dv < \infty \tag{10}$$

and $c_{h,L}(K)$ must be a positive constant such that the integral of the kernel over the surface Ω of the L-sphere is such that

$$\frac{h^{L-1}}{c_{h,L}(K)} = \int_\Omega K\left(\frac{1 - x^T y}{h^2}\right) d\omega(y) \tag{11}$$

holds. With these conditions, the PDF estimate has the properties of a probability density function [10, 11]. Although any kernel that satisfies these conditions is viable, the simplest and most convenient usable kernel is $K(r) = e^{-r},\ r \geq 0$, which is commonly referred to as the *von Mises kernel*, created with the von Mises - Fisher directional probability distribution on the L-sphere in mind:

$$f_{vM}(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = C_L(\kappa)\, e^{\kappa\, \boldsymbol{x}^T \boldsymbol{\mu}},$$
$$C_L(\kappa) = \frac{\kappa^{\frac{L-1}{2}}}{(2\pi)^{\frac{L+1}{2}}\, \mathcal{I}_{\frac{L-1}{2}}(\kappa)}. \tag{12}$$

Here, $\mathcal{I}_\nu(z)$ is the modified Bessel function of the first kind and order ν. The von Mises kernel is of particular convenience in a directional estimator, since $f_{vM}(\boldsymbol{x}; \boldsymbol{\mu}, \kappa)$ tends to its maximum value $C_L(\kappa)e^\kappa$ as $\boldsymbol{x}$ tends to the average direction vector $\boldsymbol{\mu}$, and tends towards zero as $\boldsymbol{x}$ becomes orthogonal to $\boldsymbol{\mu}$. If we let K be the von Mises kernel, (10) is simplified greatly, allowing for greater comprehension of what the estimator does with our data, since the inverse of the normalization constant $c_{h,L}(K)^{-1}$ becomes $c_{h,L}(K) = C_L\left(\frac{1}{h^2}\right) e^{\frac{1}{h^2}}$, allowing (9) to be rewritten as

$$\begin{aligned} \hat{f}(\boldsymbol{x}|H_k) &= C_L\left(\frac{1}{h^2}\right) e^{\frac{1}{h^2}} \frac{1}{n} \sum_{i=1}^{n} e^{-\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_{i,k}}{h^2}} \\ &= \frac{1}{n} \sum_{i=1}^{n} C_L\left(\frac{1}{h^2}\right) e^{\frac{\boldsymbol{x}^T \boldsymbol{X}_{i,k}}{h^2}} \end{aligned} \tag{14}$$

or, in other words, using the von Mises kernel provides a generalized estimator of directional densities, being the estimate a mixture of von Mises - Fisher directional density probability functions [10]:

$$\begin{aligned} \hat{f}_h(\boldsymbol{x}\,|H_k) &= \frac{1}{n} \sum_{i=1}^{n} f_{vM}\left(\boldsymbol{x};\, \boldsymbol{X}_{i,k}, \frac{1}{h^2}\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} C_L\left(\frac{1}{h^2}\right) e^{\frac{\boldsymbol{x}^T \boldsymbol{X}_{i,k}}{h^2}}. \end{aligned} \tag{15}$$

This implies that $\tilde{f}_h\,(\boldsymbol{x}|H_k)$ will take greater values at the directions x where most training vectors $\boldsymbol{X}_k$ are pointing. The resolution of the estimation is highly dependent on the value of h, the *estimator bandwidth*. Its value must be chosen so that the $\mathcal{L}_2$ distance (mean squared error, MSE) between the estimation $\hat{f}_h$ and f (the actual, yet unknown PDF) is minimal. This error happens to be a random variable, since it depends on the number of training vectors used in the estimation (n) and on the fact that those training vectors will be different in each hyperspectral image. Thus, the *Mean Integrated Square Error* (MISE) is employed to find the expected value of the MSE of the approximation

$$\mathrm{MISE}(h) = E\left[\int_{\Omega_q} \left(\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\, \omega_q(d\boldsymbol{x})\right]. \tag{16}$$

In this case, Eduardo García - Portugués's rule-of-thumb estimator bandwidth h_{ROT} was used, as it minimizes the mean integrated square error if the von Mises kernel is used [11]:

$$h_{ROT} = \begin{cases} \left(\dfrac{4\pi^{\frac{1}{2}}\, \mathcal{I}_0(\hat{\kappa})^2}{\hat{\kappa} n \left(2\mathcal{I}_1(2\hat{\kappa}) + 3\hat{\kappa}\mathcal{I}_2(2\hat{\kappa})\right)}\right)^{\frac{1}{5}}, & L = 1, \\ \left(\dfrac{8\sinh^2(\hat{\kappa})}{\hat{\kappa} n \left((1 + 4\hat{\kappa}^2)\sinh(2\hat{\kappa}) - 2\hat{\kappa}\cosh(2\hat{\kappa})\right)}\right)^{\frac{1}{6}}, & L = 2, \\ \left(\dfrac{4\pi^{\frac{1}{2}} \mathcal{I}_{\frac{L-1}{2}}(\hat{\kappa})^2}{\hat{\kappa}^{\frac{L+1}{2}}\, n \left(2L\mathcal{I}_{\frac{L+1}{2}}(2\hat{\kappa}) + (2 + L)\hat{\kappa}\, \mathcal{I}_{\frac{L+3}{2}}\right)}\right)^{\frac{1}{4+L}}, & L \geq 3. \end{cases} \tag{17}$$

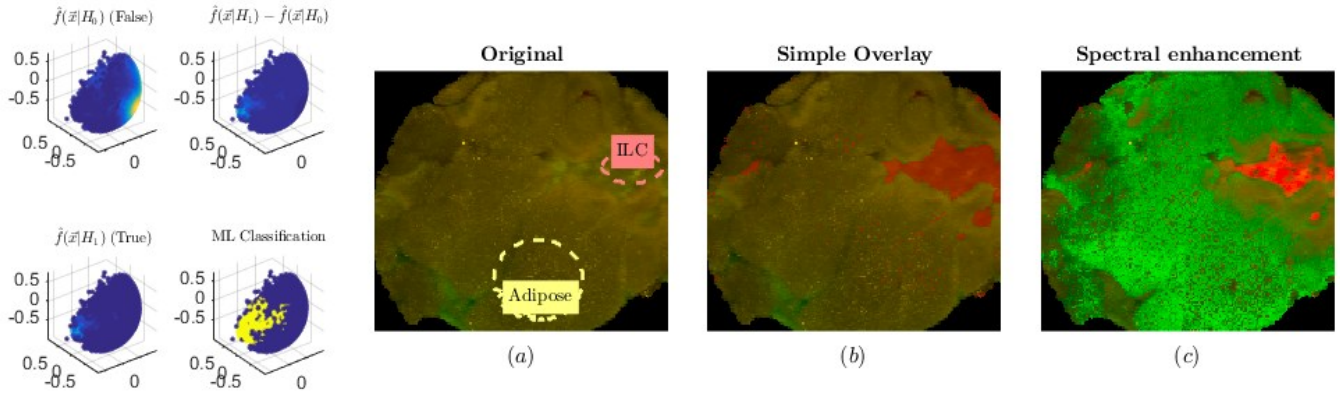

Fig. 2. Sample #23. The surgical margin of an Invasive Lobular Carcinoma (ILC) surrounded by adipose tissue can be seen more clearly after undergoing d-KDE classification.

Here, $\hat{\kappa}$ is an estimate of κ, known as the accumulation parameter. This parameter could be found using a maximum likelihood estimator. Nevertheless, we will employ the following approximation of the ML estimate of κ described by Inderjit Dhillon and Suvrit Sra in 2003, which provides a fair approximation with a negligible error [12]:

$$\hat{\kappa} \approx \frac{\bar{R}L - \bar{R}^3}{1 - \bar{R}^2}, \tag{18}$$

where $\bar{R} \triangleq \frac{\|\Sigma_{i=1}^{n} \boldsymbol{x}_i\|}{n}$ is the norm of the average vector. With these approximations, it is possible to rewrite the PDF estimates for both hypotheses as follows:

$$\hat{f}_{h_{ROT}}(\boldsymbol{x}|H_0) = \frac{1}{n_0}\sum_{i=1}^{n_0} C_L \left(\frac{1}{h_{ROT,0}^2}\right) e^{\frac{\boldsymbol{x}^T \boldsymbol{X}_{i,0}}{h_{ROT,0}^2}}, \tag{19}$$

$$\hat{f}_{h_{ROT}}(\boldsymbol{x}|H_1) = \frac{1}{n_1}\sum_{i=1}^{n_0} C_L \left(\frac{1}{h_{ROT,1}^2}\right) e^{\frac{\boldsymbol{x}^T \boldsymbol{X}_{i,1}}{h_{ROT,1}^2}}. \tag{20}$$

Here, n_0 and n_1 are the number of training vectors used in each estimator, and $h_{ROT,0}$ and $h_{ROT,1}$ are the estimator bandwidths, found by inserting (18) into (17). Sample vectors to use in this estimation are selected on a screen by the experimented surgeon on call, and the estimation will be done for every hyperspectral pixel in the image, thus providing two scores (two PDF values) for every pixel.

5) Maximum Likelihood Classification

Maximum Likelihood (ML) classification, i.e. Maximum-A-Posteriori (MAP) classification assuming all hypotheses are equally probable, has been found to be convenient in this case. When a spectrum scores higher on the PDF associated with malignant tissue than on the PDF associated with healthy tissue, it is classified as malignant, and vice versa. Thus, the rule

$$\Lambda(\boldsymbol{x}) = \frac{\hat{f}(\boldsymbol{x}|H_A)}{\hat{f}(\boldsymbol{x}|H_B)} \underset{H_B}{\overset{H_A}{\gtrless}} \frac{P(H = H_B)}{P(H = H_A)} = \gamma \tag{21}$$

may be used to classify each spectrum/pixel depending on its scores (PDF values). If we take the logarithm of both sides and assume that the likelihood of a pixel being malignant or non-malignant is the same for both cases, we obtain a *Maximum Likelihood* classification rule:

$$\hat{f}(\boldsymbol{x}|H_A) - \hat{f}(\boldsymbol{x}|H_B) \underset{H_B}{\overset{H_A}{\gtrless}} 0 \tag{22}$$

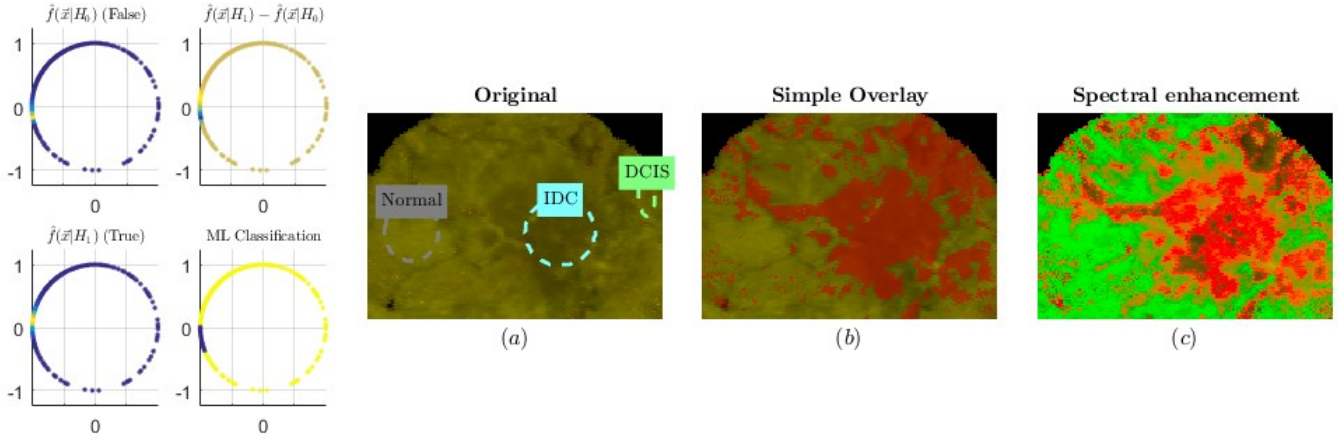

Fig. 3. Sample no. 24. An Invasive Ductal Carcinoma (IDC) and a Ductal Carcinoma In Situ (DCIS) lies within a strip of healthy tissue.

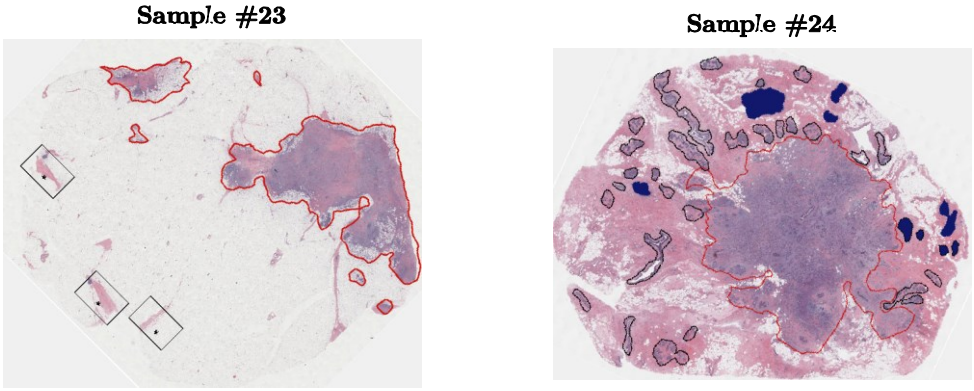

Fig. 4. Histological results of samples 23 and 24. The hidden shape under the surface of the ILC in Sample 23 and the shape of the IDC in Sample 24 are consistently similar in both classification and histological results.

6) Graphical representation of the classification results

The values of the estimated directional PDFs at every pixel using d-KDE and the classification categories they have been assigned to are meant to be shown on a screen, in a way that it helps the surgeon with the assessment of the surgical margin of interest. Two approaches are proposed: (a) an alpha-channel overlay applied over the image, and (b) a multispectral color addition operation that modifies the color of the image according to these results, using the CIE 1931 Color Matching Functions [13] and applying the colorimetric operation

$$\boldsymbol{g}'_k = \boldsymbol{g}_k + K_0\,\hat{f}(\boldsymbol{g}_k|H_0)\boldsymbol{c}_{green} + K_1\,\hat{f}(\boldsymbol{g}_k|H_1)\boldsymbol{c}_{red}, \quad (23)$$

where $\hat{f}(\boldsymbol{g}_k|H_0)$ and $\hat{f}(\boldsymbol{g}_k|H_1)$ are the aforementioned directional PDF estimates, K_0 and K_1 represent tunable non-negative gains, and $\boldsymbol{c}_{green}$ and $\boldsymbol{c}_{red}$ are any two selected spectra – in this case, green and red – chosen to represent the likelihood of a pixel being healthy or malignant tissue, respectively. This operation will add the desired color scheme to the original hyperspectral image upon hyperspectral reconstruction.

III. Results

A. Qualitative results

An empirically appropriate way to present the potential of this method is to show a best-case scenario and its counterpart. Sample 23 (Figure 2) displays an Invasive Lobular Carcinoma (ILC) surrounded by rather distinguishable adipose tissue, while Sample 24 (Figure 3) shows an Invasive Ductal Carcinoma (IDC) and a Ductal Carcinoma In Situ (DCIS) concealed in a layer of healthy tissue. This information is provided by a set of Regions of Interest ascertained in every image by the pathologist in charge, after a thorough histological analysis. These ROIs, digitally marked as filled circular pixel masks, do not represent the exact morphological limits of each tissue type in each image, but rather locations in the hyperspectral image where the tissue type has a certain diagnosis, as given by the pathologist.

Hyperspectral images (Figures 2.a and 3.a) were reconstructed by performing a standardized spectrum-to-RGB transformation, e.g. by integrating reflectance spectra multiplied by the CIE 1931 Color Matching Functions [13]. Given that each pixel only has samples in the 510-785 nm range, the CIE 1931 CMFs lack the first wavelengths of the visible spectrum for each pixel and thus the end result differs from the white-light images, the latter taken with a conventional RGB camera. Using 15% of the ROIs, both PDF estimations for malignant and healthy tissue are generated, and the whole picture is classified afterwards. The result of this procedure can be seen in Figures 2 and 3. These figures are identically composed of two parts each. The first part corresponds to the four scatter plots at the left-hand side, which are representing the pixels (vectors) in the first three dimensions of their L-dimensional space, namely two heat maps of the values of $\hat{f}(\boldsymbol{x}|H_0)$ and $\hat{f}(\boldsymbol{x}|H_1)$ for all pixels in the image (upper and lower left-hand side figures), a heat map

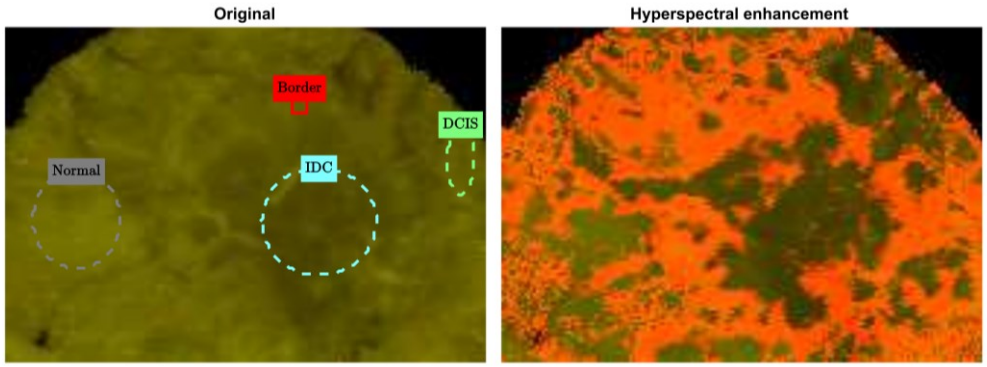

Fig. 5. Prompting d-KDE to highlight a region surrounding a large invasive ductal carcinoma (red square box labeled 'Border'). The other ROIs are not used here and are shown for illustration purposes. The color spectrum for this enhancement is a Gaussian curve with $\mu = 610$ nm, $\sigma = 10$ nm multiplied by $K = 10$.

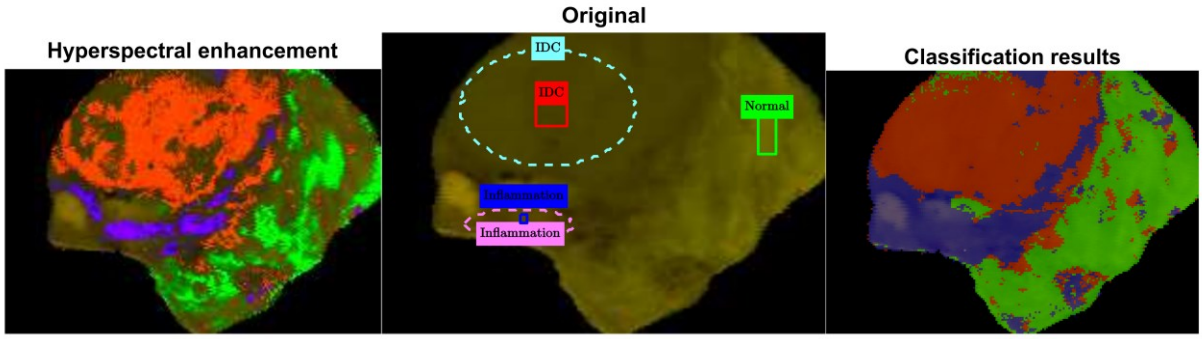

Fig. 6. Multiple regions, namely the square selection boxes labeled 'Normal', 'Inflammation' (in blue) and 'IDC' (in red) in the central image are selected for enhancement (left picture) and classification (right picture). As before, hyperspectral enhancement was performed with (23).

with the value of $\hat{f}(\boldsymbol{x}|H_1) - \hat{f}(\boldsymbol{x}|H_0)$ for each pixel (upper right figure) and finally a color-coded plot with the classification result (lower right figure). Secondly, there are three images labeled (a), (b) and (c), which correspond to (a) a hyperspectral reconstruction from spectral data to sRGB, (b) a semitransparent overlay showing classification results after using the ML rule in (22), and (c) the result of applying the hyperspectral enhancement in (23), using green color for $f(\boldsymbol{x}|H_0)$ and red for $f(\boldsymbol{x}|H_1)$. To compare these results with the traditional methodology, Figure 4 shows the end result of a traditional H&E stain procedure that was performed on a slice of each sample. After analyzing each slice, a trained pathologist establishes the surgical margin and draws the ROIs on the hyperspectral images. There is a fairly strong similarity between the shapes of the tumors in both procedures, which could imply that, given a proper frame of reference, evaluating and/or measuring surgical margins could be performed –and shown– automatically. More research is needed to prove the definitive morphological similarity between these results.

In these examples, only the defined ROIs have been used for estimation and classification/enhancement of margins in a lumpectomy sample. Although this resembles a practical scenario where some regions are clear, the potential capabilities of this algorithm call for further analysis from a qualitative point of view. For example, it would be interesting to see the output of this methodology when an unclear region is selected for enhancement. This is the case of Figure 5, where the border of the tumor is highlighted instead. The selection, in this case, is the red rectangular region labeled '*Border'*. The other ROIs are shown but not used during estimation. As expected, selecting the border of an IDC would result in an enhancement of pixels in the image that have similar spectral signatures (i.e. the borders of the tumor). This, for instance, could be of great use whenever there is uncertainty about what is present in a hyperspectral image.

Another relevant feature of d-KDE is its ability to highlight multiple spectral signatures. Throughout this article, we have focused on binary classification, given that surgical margin assessment requires only the distinction between cancer and the normal tissue surrounding it, but this is not the only scenario where d-KDE could come to good use. In Figure 6, a border of tissue inflammation is highlighted somewhere in between an Invasive Ductal Carcinoma and normal tissue. Again, in this case the square boxes represent the pixels used during estimation, whilst the ROIs provided by the pathologist are not used in any part of the process. In environments where multiple tissue types would be present and classification or delineation was required, this methodology could be used as long as sample spectral signatures are identifiable in the image.

B. Quantitative results

A total of seven categories were created in order to quantify overall classification performance. Each sample is to fall into several categories if category conditions are met. In samples where any type of cancer (DCIS, IDC, ILC) was present accompanied by any non-malignant tissue type (Normal, Benign, Inflammation, Adipose), all malignant tissue ROIs in the image (H_1) were classified against all benign tissue ROIs (H_0). This corresponds to the first column in Tables 1 and 2. The six remaining categories are for samples where a specific tissue type – either normal (healthy), DCIS, IDC, ILC, tissue showing inflammation, or adipose tissue – was found accompanied by at least another tissue type (benign or malignant). In those cases, the tissue type of interest (H_1) was classified against all other tissue types existing in the image (H_0). There were no samples where a benign growth appeared with other kinds of tissue, therefore making it impossible to

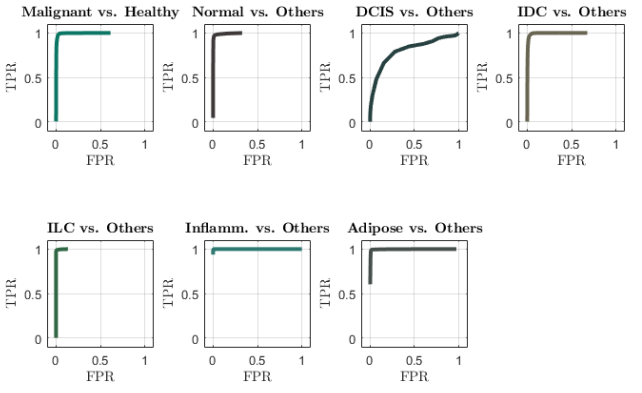

Fig. 7. Receiver Operating Characteristic of d-KDE when selecting $L = 10$.

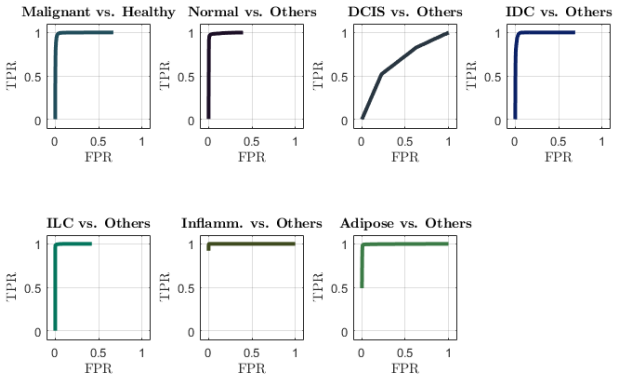

Fig. 8. Receiver Operating Characteristic of d-KDE with L obtained by using (7).

check the ability of the classifier to distinguish benign growths from other materials. In every sample, only 15% of each ROI was randomly picked to train the d-KDE classifier according to the selection of H_0 and H_1, and the remaining 85% was counted to create a confusion matrix and find the sensitivity and specificity for that random pixel selection. The process was repeated 20 times for different random pixel selections, and the average sensitivity and specificity for that sample under those conditions (hypotheses) was found. Finally, the average sensitivity and specificity of a particular category is found by finding the average sensitivity and specificity of all samples in the category. In a similar fashion, the accuracy, Dice-Sorensen metric, positive and negative Likelihood Ratios (LR+, LR-) and the Diagnostic Odds Ratio (DOR) were found 20 times per sample, then averaged as stated before. Tables I and II show the results of following this procedure. The overall sensitivity and specificity of the classifier at discerning malignant spectra from the rest was 98% and 97%, respectively, when selecting L dynamically using (8). This benchmark holds for all other columns as well, with an exception: only one sample (#24) with ductal carcinoma in situ (DCIS) is found in the dataset, and its ROI is notably small in comparison with the other ROIs (194 pixels in the ROI, thus approximately 29 of them are used for estimation) of that region is a training set too small for proper accuracy evaluation, and yet the classifier performs fairly well. Nevertheless, for the majority of samples with the most common invasive cancer growths –invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) [1, 2] – the overall accuracy was equal to or above 97%.

Using the same sample categories explained above, we could change γ in (23) to create the average Receiver Operating Characteristic (ROC) of the classifier and confirm that the sensitivity and specificity is maximal at $\gamma = 1$, which happens to be the case. The shape of the ROC is that of a nearly-ideal classifier, which means that Maximum Likelihood classification is sufficient to assess the type of tissue appearing in each pixel. The ROCs of the classifier for both selection procedures of L are shown in Figures 7 and 8.

IV. Performance Analysis

In this final section, we will evaluate two fundamental factors that could modify the behavior of this methodology, namely (a) the robustness of the algorithm in terms of the amount of pixels used for PDF estimation, and (b) the influence of dimensionality on overall classification performance. First, we study the influence of the ROI percentage used in estimation, given that estimation fidelity of non-parametric estimators grows asymptotically with the number of sample vectors used in the procedure [14]. The relative amount of pixels used will be denoted $R_{\%}$ and, in order to find out its influence, additional simulations have been performed. In this scenario, we evaluated only those lumpectomy samples that show both healthy and malignant ROIs. In this case, $R_{\%}$ percent of their ROIs have been randomly selected for classification, and its corresponding sensitivity, specificity and accuracy has been calculated. This has been performed a total of 10 times per sample, then averaged to find an estimate of the average performance of d-KDE for that sample and relative amount of ROI pixels used. The result of this operation is shown in Figure 9. Although classification performance increases with $R_{\%}$, it is notably constant (average accuracy slope of 0.0001 for $R_{\%}$ in the interval [5,100]), which implies that d-KDE works

TABLE I
D-KDE CLASSIFICATION RESULTS FOR $L = 10$

	Malignant	Normal	DCIS	IDC	ILC	Inflamm.	Adipose
FPR	0.0196	0.0461	0.1862	0.0304	0.0054	0.0051	0.1110
Sensitivity	0.9533	0.9819	0.6359	0.9720	0.9862	0.9967	0.9880
Specificity	0.9804	0.9539	0.8138	0.9696	0.9946	0.9949	0.9889
Accuracy	0.9652	0.9641	0.7979	0.9737	0.9927	0.9951	0.9907
Dice-Sørensen	0.9719	0.9908	0.7303	0.98333	0.9930	0.9983	0.9939
LR+	123.9921	1332.0628	11.5438	106.4292	230.8108	221.2748	174.4786
LR-	0.0482	0.0199	0.4025	0.0292	0.0139	0.0033	0.0120
DOR/10^4 (min)*	2.6817	1.9514	0.0026	2.2217	2.2671	8.7620	2.1880

Results for a constant value of L ($L = 10$). Each column shows the performance of the classifier when distinguishing the tissue specified by the column label from any other tissue type.

(*) For this calculation, trials that returned DOR = ∞ were discarded to obtain a lower bound.

TABLE II
D-KDE CLASSIFICATION RESULTS FOR A DYNAMIC SELECTION OF L

	Malignant	Normal	DCIS	IDC	ILC	Inflamm.	Adipose
FPR	0.0313	0.0174	0.4049	0.0463	0.0071	0.0037	0.0092
Sensitivity	0.9848	0.9720	0.7730	0.9897	0.9872	0.9961	0.9866
Specificity	0.9687	0.9826	0.5951	0.9537	0.9929	0.9963	0.9908
Accuracy	0.9809	0.9785	0.6107	0.9763	0.9916	0.9963	0.9911
Dice-Sørensen	0.9922	0.9854	0.8516	0.9947	0.9935	0.9981	0.9932
LR+	114.5118	122.7030	1.8645	94.3353	182.8337	303.488	326.3379
LR-	0.0165	0.0293	0.3535	0.0118	0.0129	0.0039	0.0135
DOR/10^4 (min)*	3.0297	2.3193	0.0006	2.8021	1.7103	7.5610	2.5518

Results with L selected dynamically with (8), and $\sigma_{contrib} < 0.01$. Each column shows the performance of the classifier when distinguishing the tissue specified by the column label from any other tissue type.

(*) For this calculation, trials that returned DOR = ∞ were discarded to obtain a lower bound.

appropriately even with rough PDF estimates, i.e. generated with a reduced amount of reference pixels. Sample 24 shows d-KDE performing less proficiently, which is related to the fact that DCIS and IDC are equally 'malignant' during this classification, and that DCIS is not only a minority in terms of pixel population, but also has a spectral signature that is very similar to healthy tissue [1,4,5]. In a clinical scenario, after the removal of the IDC, we would proceed into looking for DCIS, which would be highlighted more proficiently as soon as spectral signatures dissimilar to healthy tissue are no longer in the picture.

Secondly, we have evaluated the influence of dimensionality on the estimator. In most cases, it performs fairly well after $L \geq 3$, and does not worsen as L increases, mainly because the first singular vectors will describe most of the general features of most spectra – the rest end up as residuals of apparent minor relevance. In this scenario, Sample 24 again exhibits a singular behavior: its specificity remains constant, whilst its sensitivity decays with L. Again, the fact that DCIS has a similar spectrum to that of healthy tissue will likely set DCIS as a false negative in a scenario where dimensionality reduction is done by methods such as PCA or SVD and there are spectral signatures a lot different in comparison with healthy tissue. Thus, as more information proves DCIS to be similar to Normal tissue, it seems that false negatives spike up and sensitivity goes down, while specificity remains constant.

V. CONCLUSION

The increasing use of BCT/lumpectomy procedures in the treatment of breast cancer should ideally be accompanied by new tools that allow a quick and safe evaluation of the surgical margins of any extracted tumor before closing the intraoperative cavity. In this article, d-KDE classification of breast tissue spectra has been shown to be capable of surpassing the current state-of-the-art margin delimitation procedures in terms of sensitivity and specificity, being the current benchmark (PCA+ICA) with this database at 93% and 95%, respectively [1, 2, 4]. The underlying concept, i.e. *spectral directionality,* shown in algorithms such as Spectral Angle Mapping (SAM) seems to function robustly when combined with dimensionality reduction and nonparametric estimation [15].

The proposed implementation of d-KDE followed a total of five steps. Spectra were corrected by finding the Standard Normal Variate of each spectrum, and their dimensionality was reduced by means of performing the SVD of a matrix whose rows are the corrected spectra. Finally, only the direction is preserved after normalizing them in a lower-dimensional space, and these directions have been proven to be sufficient to achieve a remarkable sensitivity and specificity of 98% and 97% respectively after estimating the directions of all tissue types appearing in the image.

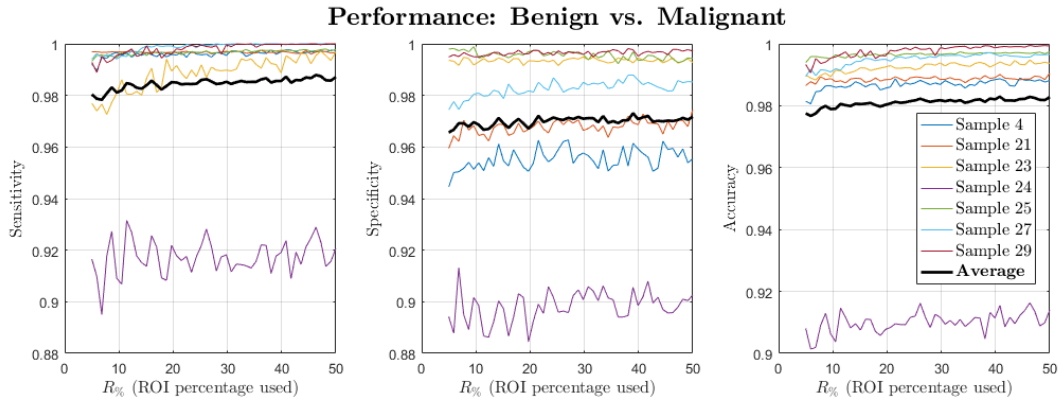

Fig. 9. Average classification performance of d-KDE as a function of the amount of ROI pixels used. In these simulations, d-KDE distinguishes between malignant and non-malignant tissue.

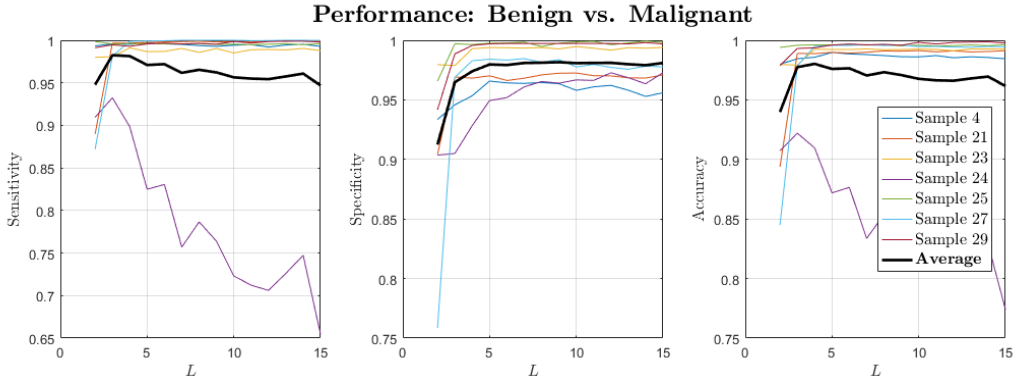

Fig. 10. Average performance of d-KDE when varying L. In these simulations, d-KDE distinguishes between malignant and non-malignant tissue.

Although there are significant morphological similarities between the H&E stain histology pictures and the classification results, more research is needed in order to truly assess the capabilities of the explained methodology. Even with the low penetrative properties of Vis-NIR light, the backscattered light provides information about layers of tissue slightly below the surface, giving the hyperspectral image more information than that of a histological analysis of a single slice of paraffin-embedded tissue (e.g. an ILC below a fat layer in Figure 2). Thus, it would be necessary to either have more slices per sample and superimpose the margins of each transversal cut of the tumor, or possibly perform d-KDE on a slice instead of using the complete tissue sample. Also, it would be crucial to establish an equivalent frame of reference in both the H&E photographs and the hyperspectral pictures, to compensate for scale and rotation differences between pictures and guarantee an absolute margin comparison beyond the defined regions of interest.

From an empirical standpoint, clinical studies need to be completed to prove the experimental validity of this classifier if morphological similarities are proven to be truly exact. In a real clinical setting, the assessment procedure would take place in a way similar to that of frozen section analysis, for instance, but eliminating all time spent in sample preparation: immediately after the growth was extracted, it would be placed inside the imaging system, and the practitioner would only need to indicate by reference (perhaps, aided by a graphical user interface) which tissues to distinguish in the sample. Also, the growing use of far-field spectroscopy imaging technologies in biomedicine [6, 16, 17] could benefit from this new approach, in those cases where tissue differentiability is not evident. In those cases, once the d-KDE approach receives the regions on the L-sphere that represent several tissue types, the device could be pointed inside the intraoperative cavity to classify in real time, highlighting any remaining tissue that would require proper excision.

REFERENCES

[1] A. M. Laughney, P. B. García-Allende, O. M. Conde, W. A. Wells, and K. D. Paulsen, "Automated classification of breast pathology using local measures of broad reflectance", *J. Biomed Opt.*, vol. 15, no. 6, 066019 Nov./Dec. 2010.

[2] A. M. Laughney, V. Krishnaswamy, E. J. Rizzo, M. C. Schwab, R. J. Barth, B. W. Pogue, K. D. Paulsen, and W. A. Wells, "Scatter spectroscopic imaging distinguishes between breast pathologies in tissues to surgical margin assessment", *Clin Cancer Res.* vol. 18, no. 22, pp. 6315-6325, Nov. 2012.

[3] K. von Smitten, "Margin status after breast-conserving treatment of breast cancer: how much free margin is enough?", J Surg Oncol., vol. 98, no. 8, pp. 585-587, Dec. 2008.

[4] P. B. García-Allende, V. Krishnaswamy, P. J. Hoopes, K. S. Samkoe, O. M. Conde, and B. W. Pogue, "Automated identification of tumor microscopic morphology based on macroscopically measured scatter signatures", *J. Biomed Opt..,* vol. 14, no. 3, May-Jun. 2009.

[5] A. Eguizábal, A. M. Laughney, P. B. García-Allende, V. Krishnaswamy, W. A. Wells, K. D. Paulsen, B. W. Pogue, J. M. López-Higuera, and O. M. Conde, "Direct identification of breast cancer pathologies using blind separation of label-free localized reflectance measurements", *Biomed Opt. Express,* vol. 4, no. 7, pp. 1104-1118.

[6] V. Krishnaswamy, P. J. Hooves, S. Samkoe, J. A. O'Hara, T. Hasan, and B. W. Pogue, "Quantitative imaging of scattering changes associated with epithelial proliferation, necrosis and fibrosis in tumors using microsampling reflectance spectroscopy", *J. Biomed Opt.,* vol. 14, no. 1, 014004, Jan-Feb. 2009.

[7] R. J. Barnes, M. S. Dhanoa, S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra", *J. of Applied Spectroscopy*, vol. 43, no. 5, pp. 772-777, May 1989.

[8] Charles E. Miller, "Chemometrics in Process Analytical Technology (PAT)", in *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries* edited by K. A. Bakeev, 2nd ed., John Wiley & Sons, United Kingdom, 2010, pp. 372-374

[9] J. W. Demmel, "Linear Least Squares Problems" in his book *Applied Numerical Linear Algebra*, 1st ed., Society for Industrial and Applied Mathematics, 1997, pp. 109-112.

[10] E. G. Portugués, R. M. Crujeiras, and W. González-Manteiga, "Kernel density estimation for linear-directional data", *J. Multivariate Anal.*, vol. 121, pp. 152-275, Oct. 2013.

[11] E. G. Portugués, "Exact risk improvement of bandwidth selectors for kernel density estimation with directional data", *Electron. J. Stat*., vol. 7, pp. 1655-1685, Jun-Dec. 2014.

[12] I. S. Dhillon and S. Sra, "Modeling data using directional distributions", Technical report, Technical Report TR-03-06, University of Texas at Austin, Austin, TX (2003).

[13] J. Schanda, "Chapter 3: CIE Colorimetry" from his book *Colorimetry: Understanding the CIE System,* John Wiley & Sons, Hoboken, New Jersey. pp. 25-76.

[14] Alan Julian Izenman, "Recent developments in nonparametric density estimation". *Journal of the ASA*, vol. 86, no. 413, pp. 205-224, 1991.

[15] P. Beatriz García-Allende, Olga M. Conde, J. Mirapeix, A. M. Cubillas, J.M López-Higuera, "Data processing method applying principal component analysis and spectral angle mapper for imaging spectroscopic sensors", *IEEE Sensors J.* vol 8, no. 7, pp. 1310-1316, July 2008.

[16] T. D. O'Sullivan, A. E. Cerussi, D. J. Cuccia, and B. J. Tromberg, "Diffuse optical imaging using spatially and temporally modulated light", *J Biomed Opt.* vol. 17, no. 7, July 2012.

[17] D. M. McClatchy III, E. J. Rizzo, W. A. Wells, P. P. Cheney, J. C. Hwang, K. D. Paulsen, Brian W. Pogue, and Stephen C. Kanick. "Wide-field quantitative imaging of tissue microstructure using sub-diffuse spatial frequency domain imaging", *Optica*, vol. 3, no. 6, 2016.