Multivariate Classification of Fourier Transform Infrared Hyperspectral Images of Skin Cancer Cells

Francisco Peñaranda*, Valery Naranjo*, Lena Kastl†, Björn Kemper†, Gavin R. Lloyd‡,
Jayakrupakar Nallala§, Nicholas Stone§ and Jürgen Schnekenburger†
*Instituto de Investigación e Innovación en Bioingeniería (I3B),
Universitat Politècnica de València, Camino de Vera s/n, 46022 Valncia, Spain
Email: {frapeago,vnaranjo}@upv.es
†Biomedical Technology Center, University of Münster, Münster, Germany
‡Biophotonics Research Unit, Gloucestershire Hospitals NHS Foundation Trust, Gloucester, UK
§Biomedical Physics, School of Physics, University of Exeter, Exeter, UK

Abstract—**A multilevel framework for the multiclass classification of spectra extracted from Fourier transform infrared images is described. This learning structure was employed to discriminate the spectra extracted from hyperspectral images of two batches of four different skin cultured cells (two normal and two tumor), where the cells of one batch had been stained with fluorescence live cell dyes. Different options were explored in each stage of the framework, specifically in the spectral pre-processing and the employed classification algorithm. Special care was taken to optimize the learning models and to objectively estimate the generalization performance by means of cross-validation. A very high discriminative performance was obtained for all the unstained skin cell types. However, the presence of the stains introduces spectral artifacts that worsen the class separation, as has been demonstrated in several classification experiments.**

I. Introduction

Fourier transform infrared (FTIR) microspectroscopy is an emerging technology that has demonstrated a great potential for the diagnosis of different kinds of cancer [1], [2]. This technique provides spectra with several hundreds or even thousands of absorbance values, registered at different wavenumbers (inverse of wavelengths) in the near- and mid-infrared region. During the last decade, the capability of FTIR devices has evolved from the measurement of single-point spectra to the acquisition of hyperspectral images, what allows to record a greater amount of data within a shorter time [3].

In FTIR images, each pixel has an associated spectrum that informs of the biochemical structure of a microscopic region of the space. These spectra can be processed through multivariate analysis to perform an objective characterization of diverse biological materials [4]. One of the specific uses is the categorization of fixed cytological samples from different types of diseases [5]. Due to the novelty of this technique, few cytological studies exist where an objective classification with a quantitative evaluation has been performed and, to our knowledge, no one has focused on skin cancer cells [6].

This study presents a framework for the multiclass classification of four types of skin cells (two normal and two cancerous) based on their FTIR spectra, giving special attention to the processing and handling of these signals. These cells were cultured under standard cell culture and properly treated in order to acquire FTIR images. Besides, each cell line was divided into two batches and one of them received a staining treatment with different fluorescence live cell dyes. These stains enable to create images by fluorescence microscopy that can be used as a ground truth in a possible mixture of different cell types. Such a mixture of skin cells pretends to be a proof of concept in a future advanced demonstration of the diagnosis properties of FTIR spectra in this type of cancer. However, the possible influences of the dyes must be previously evaluated.

The main goal is to find out if FTIR spectroscopy provides enough information to accurately characterize each skin cell line. An additional objective is to assess if the presence of the fluorescence dyes can worsen this classification.

This paper is organized as follows: Sec. II describes the characteristics of the used dataset of skin cells and details the classification framework that has been applied and assessed systematically in four different experiments; Sec. III presents the quantitative and qualitative results of the four classification experiments, which are discussed in Sec. IV; finally, the main conclusions of the study are summarized in Sec. V.

II. Materials and Methods

A. Dataset

Four different skin cell lines, two non-tumor (NIH-3T3 fibroblasts, HaCaT keratinocytes) and two tumor cells (A-375, SK-MEL-28 melanoma cells), were separately cultured and divided into two different batches of samples. The samples of one batch were stained with different fluorescence live cell dyes for each cell type and the other batch of samples remained unstained. The fluorescence dyes were chosen in such a way that an overlap of fluorescence emission spectra was prevented meanwhile the cell viability was preserved. Finally, all the samples followed a standardized preparation protocol suitable for FTIR spectroscopic measurements [7].

A hyperspectral FTIR image was acquired from each one of the eight samples by using a FTIR Agilent imaging system with a focal plane array (FPA) detector of 128×128 pixels. Each pixel had an equivalent size of 5.5×5.5 μm^2 in the image, which is a good compromise between separation of cells and coverage of sufficient cellular regions to obtain

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

average fingerprints. The spectra were acquired in transmission mode in the interval of wavenumbers 1000-3800 cm^{-1} with a spectral resolution of 4 cm^{-1}. In order to improve the signal-to-noise ratio, 128 scans of the same region were co-added. Additionally, a background FTIR image of the substrate without any sample was acquired as a reference to convert the recorded intensity spectra to absorbance magnitude [2]. As a result, the dataset comprises 8 hyperspectral FTIR images of 128×128 pixels corresponding to each cell line and staining option.

B. *Experimental setup*

Four multiclass classification experiments with different training and test datasets are proposed to evaluate the hypotheses exposed in Sec. I:

- *Experiment 1:* the training and test datasets are composed of the extracted spectra from the unstained batch of samples, which are classified into the four cell lines through nested *cross-validation* (CV). This is the most important experiment because it assesses the discrimination capability of FTIR spectra without introducing any staining artifact. Therefore, its performance is taken as a reference for the rest of experiments.
- *Experiment 2:* is analogous to *Experiment 1* but using the stained samples for nested CV. It serves to evaluate the possible changes in the classification performance due to the addition of the fluorescence stains.
- *Experiment 3:* the classification algorithms are trained by CV with the stained spectra and tested in the whole unstained dataset. In this experiment two effects are jointly evaluated: the generalization capability between spectra from different images and, again, the possible influence of the staining.
- *Experiment 4:* is equivalent to *Experiment 3* but taking the unstained samples for training by CV and the stained spectra for testing. The same effects are explored, too.

C. *Classification framework*

The general classification pipeline followed in all the experiments described in Sec. II-B is presented in Fig. 1. As stated by the *No Free Lunch Theorem*, there are no context- or problem-independent reasons to favor one learning method over another, each dataset may have a different structure that will require a different solution [8], [9]. Therefore, the choices at each stage of the classification pipeline should be based on the performance of the classifier on an independent test dataset. The IRootLab [10] and the LibSVM [11] toolboxes were linked with our in-house MATLAB algorithms to carry out the whole pipeline. The different explored alternatives and the aims of each step of the workflow are detailed below.

1) Spectra Extraction: The main purpose of this stage is to separate the spectra associated with cells from the substrate, whose presence may lead to misclassification. To that end, a binary mask must be obtained from each hyperspectral data cube. These masks mark the pixels whose spectra are retained. In order to illustrate the problem and as a reference for comparison, Fig. 2a shows a white light image (taken in the visible spectrum range) of a region of the stained HaCaT sample where the cells have a darker intensity than the substrate. This white light image was not available for all the samples and only covered a partial region of them.

Spectra were previously smoothed by a Savitzky-Golay (SG) filter, cropped to the *fingerprint region* and corrected by Rubberband baseline correction (see Sec. II-C2). No normalization step was applied in order to reveal tissue structures primarily based on absorbance intensity [4] and, thus, highlighting the cells (high absorbance) over the substrate (low absorbance). The standard deviation of each pixel's spectrum was computed as an integral absorbance measurement, getting a value for each pixel. The computed values were transformed to grayscale in each sample to obtain a gray image. The gray image of the stained HaCaT sample is presented in Fig. 2b, whose intensities have been inverted for comparison with Fig. 2a. Eventually, the popular Otsu's algorithm was used to compute automatic thresholds and convert each gray image to binary (Fig. 2c).

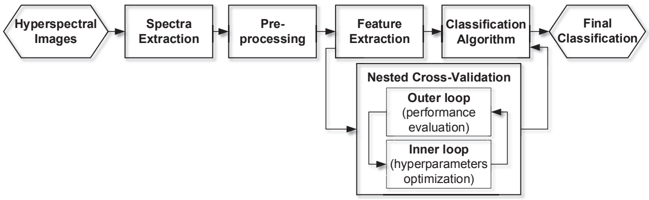


Fig. 1. Diagram of the classification pipeline.

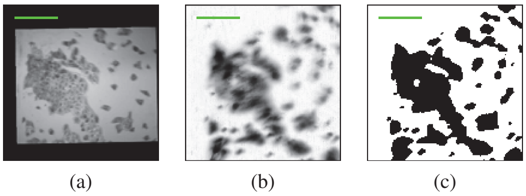


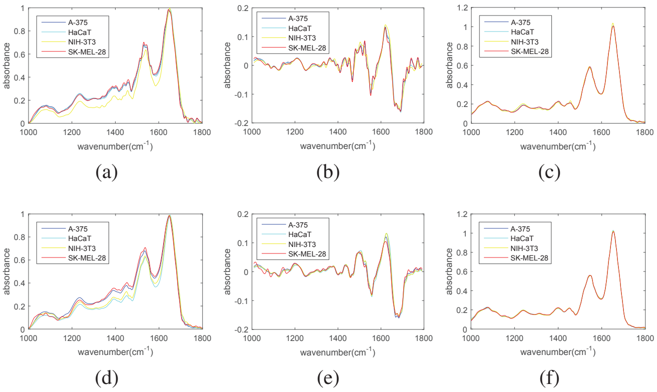
Fig. 2. Spectra extraction in stained HaCaT sample: (a) image taken in the visible spectrum range; (b) gray image extracted from the hyperspectral FTIR image; (c) final mask with relevant pixels for classification in black and discarded pixels (substrate) in white. Green scale bars represent 200 µm.

TABLE I
NUMBER OF EXTRACTED SPECTRA FOR EACH SKIN CELL LINE AND STAINING OPTION, SUITABLE FOR CLASSIFICATION.

	A-375	HaCaT	NIH-3T3	SK-MEL-28
Unstained	7818	2854	7758	7872
Stained	7931	5026	5194	7432

As a result, a different number of spectra were extracted for each cell line and staining option due to the distinct growth properties of the cell types (Tab. I).

2) Pre-processing: Pre-processing is the most important step in the workflow. High knowledge of spectroscopy theory is required to compensate the different physical phenomena that occur during the acquisition of FTIR data and that can worsen the classification. This is still an open field with



(a) (b) (c)

(d) (e) (f)

Fig. 3. Mean spectra of each skin cell line in (a)–(c) unstained and (d)–(f) stained samples for {(a),(d)} *rubber*, {(b),(e)} *diffsg1* and {(c),(f)} *rmiesc* pre-processing.

no standard solutions, the only way to check if one specific pre-processing is better is to compare its classification performance. As a guideline, different combinations of the methods suggested in [4], [9] were used in this work. Deeper description of the methods applied in each stage can be found in [12].

Two fixed steps were sequentially applied: *Savitzky-Golay smoothing,* to reduce the random noise of the spectra, whose relevant parameters were a window of 9 samples and a 2nd order fitting polynomial; *crop to band (1000-1800 cm^{-1}),* which is the so-called *fingerprint region* where the vibration modes of the most relevant biomolecules are located [2].

After these common steps, three different alternatives were explored. The objectives of these three options are the same: firstly, to reduce baseline artifacts that result from scattering; secondly, to normalize the spectra to account for confounding factors such as varying thickness and to highlight differences in biochemical structure [4]. The three options have been defined by a short word:

- **rubber:** rubberband baseline correction followed by Amide I peak normalization ($\sim$1650 cm^{-1}).
- **diffsg1:** first derivative transformation with a SG filter (window of 9 samples and a 2nd order fitting polynomial) and vector normalization.
- **rmiesc:** resonant Mie scattering correction with *matrigel* reference spectrum [13], which implicitly performs a normalization.

Fig. 3 shows the mean spectra of each cell line for the two staining options with the three pre-processing alternatives.

3) Feature Extraction: Principal Component Analysis (PCA) has been employed to reduce the dimensionality of the spectra. The objectives are to decrease the computational cost and to improve the classification by keeping the most relevant features. The number of principal components (PCs) to retain is one of the parameters that was subject to optimization by *cross-validation*. The studied values of PCs ranged from the first 10 to 100 in steps of 10.

4) Classification: In this step, a multiclass classification model is trained with a set of spectra and used to categorize independent test spectra as one of the four cell types. Thorough explanations of the used methods can be found in [8], [14].

The four employed algorithms are: *k-Nearest-Neighbor* (**KNN**) classifier, where the number of neighbors k was explored from 1 to 9; *Naive Bayes* classifier (**NB**); *Linear Discriminant Analysis* (**LDA**); *Support Vector Machine* (**SVM**) with a linear kernel, where the cost parameter C was varied in the interval $[2^{-5}, 2^{9}]$ in steps of 2^2. In SVM, the *one-against-one* approach for multiclass classification was applied.

Another important component of the classification is the metric for its evaluation. A little imbalance exists between the four classes of the dataset (Tab. I), reaching almost a 3:1 ratio in unstained samples. To avoid favoring the majority classes, the *Balanced Accuracy* (BA) was used to select the optimal models and to assess the final multiclass classifications. BA is the mean of the accuracies for each class and is defined as:

$$BA(\%) = \frac{1}{N}\sum_{i=1}^{N}\frac{c_{ii}}{\sum_{j=1}^{N} c_{ij}} \cdot 100, \tag{1}$$

where N is the number of classes (4 in this case) and c_{ij} is the number of spectra of class i classified as class j.

5) Cross-Validation: Special care was taken to estimate the generalization performance of the learning framework on independent test data and to select the complexity of the models while minimizing possible overfitting [14].

A nested CV was performed in *Experimets 1* and *2* with two loops, named *outer* and *inner loop*. In the *outer loop*, spectra were split into five folds by dividing each image in five vertical strips with equal number of extracted spectra in order to maintain the original class proportions. For each iteration of the *outer loop*, one of the folds was considered as the test set and left out of the *inner loop*. In the *inner loop*, the spectra of the four remaining folds were again split into five folds by considering their spatial proximity in the original images. In each iteration of the *inner loop*, one of the last five folds was considered as the *validation* set and was used to evaluate the performance of the model that was trained with the spectra of the other four folds. The main objective of the *inner loop* is to find the model with higher mean BA by making an exhaustive search of the hyperparameters (no. of PCs, k in KNN and C in SVM). Finally, the optimal hyperparameters were used to train a model with all the spectra of the four original folds of the *outer loop* and its performance was assessed on the left test spectra. As a result, five values of BA were obtained, one for each iteration of the *outer loop*.

The evaluation is less complex in *Experiments 3* and *4*. In those trials, one of the batches of spectra (unstained or stained) is considered as the independent test set. The other batch is used to perform an optimization process equivalent to the *inner loop* of the previously described nested CV. The whole training batch of spectra is used to train the final model with the optimal hyperparameters, which is evaluated in the test set to get a single value of BA.

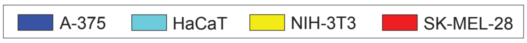


Fig. 4. Color code to represent the predicted labels for pixels in the images of qualitative results. White pixels represent substrate.

TABLE II
BALANCED ACCURACY (MEAN ± STD) IN NESTED CV OF THE UNSTAINED SAMPLES (EXPERIMENT 1)

	rubber	diffsg1	rmiesc
KNN	78.6 ± 5.9	84.1 ± 4.4	86.3 ± 5.1
NB	78.7 ± 2.9	81.6 ± 4.9	76.8 ± 5.0
LDA	93.4 ± 3.9	93.6 ± 3.7	90.3 ± 5.8
SVM	**93.8** ± **3.6**	93.8 ± 3.8	92.4 ± 4.5

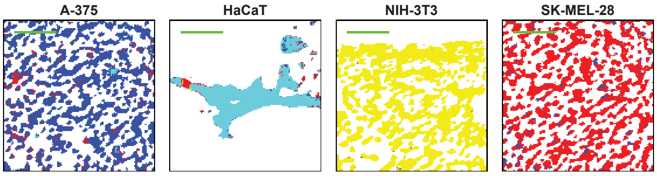


Fig. 5. Qualitative results in nested CV of the unstained samples (Experiment 1) for the combination rubber+SVM. Green scale bars represent 200 µm.

TABLE III
BALANCED ACCURACY (MEAN ± STD) IN NESTED CV OF THE STAINED SAMPLES (EXPERIMENT 2)

	rubber	diffsg1	rmiesc
KNN	68.4 ± 5.7	72.3 ± 6.1	74.7 ± 5.6
NB	56.9 ± 5.1	58.6 ± 7.4	59.2 ± 4.3
LDA	79.1 ± 3.7	80.5 ± 2.9	77.1 ± 4.9
SVM	82.4 ± 3.4	**83.6** ± **3.7**	79.9 ± 4.0

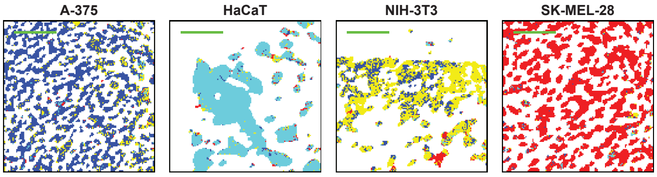


Fig. 6. Qualitative results in nested CV of the stained samples (Experiment 2) for the combination diffsg1+SVM. Green scale bars represent 200 µm.

III. RESULTS

The performance of the proposed classification framework was assessed in the available dataset through four different experiments, such as was detailed in Sec. II. For each experiment, the overall quantitative results in terms of BA are presented for every combination of pre-processing (Sec. II-C2) and classification algorithm (Sec. II-C4). The predicted labels of the best combinations were selected to construct pseudocolor images by using the color code shown in Fig. 4. The aim of these images is to obtain deeper information about why the misclassification may occur even in the best scenario.

In *Experiments 1* and *2*, the five values of BA obtained by nested CV were used to compute their mean and standard deviation (std), as can be seen respectively in Tabs. II and III. The most optimistic combinations of pre-processing and classification algorithm (highest *mean-std*) were selected to represent the qualitative results. To that end, the predicted labels in the test sets of the nested CV were joined to create pseudocolor labelled images for each sample of the batches (Figs. 5 and 6).

TABLE IV
BALANCED ACCURACY IN CLASSIFICATION OF THE UNSTAINED SAMPLES, USING THE STAINED SAMPLES FOR TRAINING (EXPERIMENT 3)

	rubber	diffsg1	rmiesc
KNN	47.0	49.4	31.7
NB	41.8	41.4	46.0
LDA	62.2	56.3	61.9
SVM	**66.9**	60.5	60.6

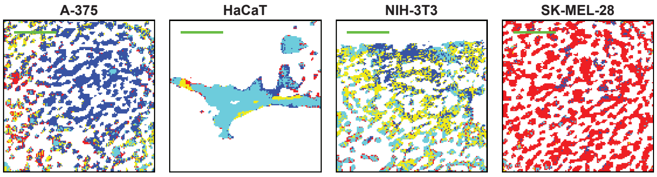


Fig. 7. Qualitative results in the classification of the unstained samples, using the stained samples for training (Experiment 3), for the combination rubber+SVM. Green scale bars represent 200 µm.

TABLE V
BALANCED ACCURACY IN CLASSIFICATION OF THE STAINED SAMPLES, USING THE UNSTAINED SAMPLES FOR TRAINING (EXPERIMENT 4)

	rubber	diffsg1	rmiesc
KNN	50.4	40.9	42.9
NB	51.3	49.0	34.5
LDA	58.8	58.6	**59.3**
SVM	48.0	49.8	44.7

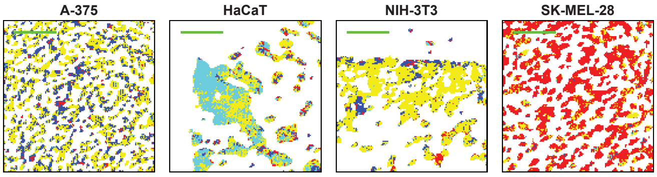


Fig. 8. Qualitative results in the classification of the stained samples, using the unstained samples for training (Experiment 4), for the combination rmiesc+LDA. Green scale bars represent 200 µm.

In *Experiments 3* and *4*, a single value of BA was computed for each combination of pre-processing and classification algorithm with the corresponding test set. These values are presented in Tabs. IV and V. The best combination (highest BA) was selected in each experiment to construct the pseudocolor images shown in Figs. 7 and 8.

IV. DISCUSSION

Several observations may be inferred from the results presented in Sec. III. As can be seen in Tab. II, the mean BA in the *Experiment 1* is higher than 90% with a std around 4% for all the pre-processing options in LDA and SVM classifiers. These values demonstrate the good discrimination capability of FTIR spectra from unstained skin cells even independently of the applied pre-processing. The qualitative results for the most optimistic combination (Fig. 5) inform of a good overall classification in all the samples. Besides, the misclassifications take place mainly in isolated regions or in borders of clustered cells. These mistakes may be due to a suboptimal removal of spectra associated to substrate before the classification and also to an imperfect correction of scattering artifacts.

The overall results are a bit worse in *Experiment 2* (Tab. III), being the highest mean BA over 80% with a std around 4% again in LDA and SVM classifiers. The qualitative results of the most optimistic combination (Fig. 6) show some minor errors, which may be due to the same suboptimal processes as in the unstained samples. However, significant mistakes happen in NIH-3T3 sample, where a considerable amount of spectra (26% specifically) are wrongly classified as A-375. Therefore, adding the fluorescence dyes slightly reduces the discrimination of the spectra but seems to introduce some confounding artifacts, specially in NIH-3T3 and A-375 cells.

The classification performance significantly decreases in *Experiments 3* and *4* (Tabs. IV and V), where the BA barely reaches 60% in the test sets for LDA and SVM and is only over 65% in one combination of *Experiment 3*. In *Experiment 3* (Fig. 7), a poorer performance is again observed in the discrimination of NIH-3T3 cells, where more than 30% of the extracted spectra are wrongly classified as HaCaT and around 25% are classified as A-375. However, the situation is inverted in *Experiment 4* (Fig. 8), where the worst classifications are performed in cells A-375 and HaCaT, whose spectra are considerably misclassified as NIH-3T3. It is difficult to deduce the main cause of these misclassifications because two effects are involved in these experiments: the generalization capability of spectra extracted from different images and the influence of the dyes on the spectral classification. Nevertheless, as was observed in *Experiment 2*, the presence of the stains seems to have a negative impact on the discriminative properties of the analysed skin cell spectra.

V. Conclusion

A multilevel framework for pixel-wise discrimination of hyperspectral FTIR images from different types of skin cells has been detailed. Different options of pre-processing and classification algorithms have been proposed to estimate the effectiveness of their possible combinations. Special care has been taken in the selection of the optimal learning models and their generalization assessment through cross-validation.

A dataset of FTIR images from four different types of cultured skin cells (two normal and two tumor) have been studied by means of the proposed framework. These cells were divided into two batches and the cells of one batch were stained with distinct fluorescence live cell dyes. The presence of these stains can be used to create ground truth images that inform of the position of the cells in possible mixed celular cultures. However, before creating those mixture experiments it was necessary to evaluate if the stains can affect the discrimination of the skin cell spectra.

High discriminative capabilities of the unstained cell spectra were demonstrated even independently of the pre-processing method if relatively complex learning algorithms (LDA and SVM) are used. Some minor mistakes were identified that may be caused by a suboptimal previous separation of the substrate and by scattering effects in the borders of the cells. On the other hand, the classification experiments with the stained cell spectra revealed a detrimental influence of the added dyes.

In the future, the main efforts will be focused on reducing the observed scattering artifacts and trying to remove the spectral fingerprint of the dyes. To that end, more complex pre-processing methods will be studied, such as advanced versions of Resonant Mie Scattering correction [13] with the incorporation of cell size information or Extended Multiplicative Signal Correction [15] with reference spectra from measurements of the pure dyes. Moreover, more cell samples will be cultured in order to obtain more FTIR images. Thus, more sound results and conclusions will be obtained about the general capabilities of FTIR spectra to discriminate skin cancer cells.

Acknowledgment

This research has been supported by the European Commission under the Seventh Framework Programme (FP7), Project MINERVA (317803; www.minerva-project.eu).

References

[1] C. Kendall *et al.*, “Vibrational spectroscopy: a clinical tool for cancer diagnostics,” *Analyst*, vol. 134, pp. 1029–1045, 2009.

[2] G. Bellisola *et al.*, “Infrared spectroscopy and microscopy in cancer research and diagnosis,” *American Journal of Cancer Research*, vol. 2, no. 1, pp. 1–21, 2012.

[3] R. Bhargava, “Infrared spectroscopic imaging: The next generation,” *Applied Spectroscopy*, vol. 66, no. 10, pp. 1091–1120, 2012.

[4] M. J. Baker *et al.*, “Using Fourier transform IR spectroscopy to analyze biological materials,” *Nature Protocols*, vol. 9, no. 8, pp. 1771–1791, Aug. 2014.

[5] F. L. Martin *et al.*, “Distinguishing cell types or populations based on the computational analysis of their infrared spectra,” *Nat. Protocols*, vol. 5, no. 11, pp. 1748–1760, Nov. 2010.

[6] G. Clemens *et al.*, “Vibrational spectroscopic methods for cytology and cellular research,” *Analyst*, vol. 139, pp. 4411–4444, 2014.

[7] L. Kastl *et al.*, “Standardized cell samples for midIR technology development,” in *Proc. SPIE 9315, Design and Quality for Biomedical Technologies VIII*. SPIE-Intl Soc Optical Eng, 2015.

[8] R. O. Duda *et al.*, *Pattern Classification*. John Wiley & Sons, 2000.

[9] J. Trevisan *et al.*, “Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives,” *Analyst*, vol. 137, pp. 3202–3215, 2012.

[10] ——, “IRootLab: a free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis,” *Bioinformatics*, vol. 29, no. 8, pp. 1095–1097, 2013.

[11] C.-C. Chang *et al.*, “LIBSVM: A Library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.

[12] P. Lasch, “Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging,” *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 100 – 114, 2012.

[13] P. Bassan *et al.*, “Resonant mie scattering (rmies) correction of infrared spectra from highly scattering biological samples,” *Analyst*, vol. 135, pp. 268–277, 2010.

[14] T. Hastie *et al.*, *The Elements of Statistical Learning*. Springer New York, 2009.

[15] H. Martens *et al.*, “Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures,” *Analytical Chemistry*, vol. 75, no. 3, pp. 394–404, 2003.