Discrimination of skin cancer cells using Fourier transform infrared spectroscopy

Francisco Peñaranda[a], Valery Naranjo[a,], Gavin R. Lloyd[b,1], Lena Kastl[c], Björn Kemper[c], Jürgen Schnekenburger[c], Jayakrupakar Nallala[d], Nicholas Stone[d]

[a] *Instituto de Investigación e Innovación en Bioingeniería (I3B), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain*
[b] *Biophotonics Research Unit, Gloucestershire Hospitals NHS Foundation Trust, Gloucester, United Kingdom*
[c] *Biomedical Technology Center, University of Münster, Münster, Germany.*
[d] *Biomedical Physics, School of Physics, University of Exeter, Exeter, United Kingdom*

Abstract

Fourier transform infrared (FTIR) spectroscopy is a highly versatile tool for cell and tissue analysis. Modern commercial FTIR microspectroscopes allow the acquisition of good-quality hyperspectral images from cytopathological samples within relatively short times. This study aims at assessing the abilities of FTIR spectra to discriminate different types of cultured skin cell lines by different computer analysis technologies. In particular, 22700 single skin cells, belonging to two non-tumoral and two tumoral cell lines, were analysed. These cells were prepared in three different batches that included each cell type. Different spectral preprocessing and classification strategies were considered, including the current standard approaches to reduce Mie scattering artifacts. Special care was taken for the optimisation, training and evaluation of the learning models in order to avoid possible overfitting. Excellent classification performance (balanced accuracy between 0.85 and 0.95) was achieved when the algorithms were trained and tested with the cells from the same batch. When cells from different batches were used for training and testing the balanced accuracy reached values between 0.35 and 0.6, demonstrating the strong influence of sample preparation on the results and comparability of cell FTIR spectra. A deep study of the most optimistic results was performed in order to identify perturbations that influenced the final classification.

Keywords: Machine learning, Multivariate analysis, Cancer diagnosis, Cytopathology, Fourier transform infrared spectroscopy

1. Introduction

Infrared spectroscopy is a highly promising optical technology in the combined identification and localization of pathophysiological cell and tissue alterations [1]. A limiting factor for the use of Fourier Transform Infrared (FTIR) spectroscopes in biomedical problems has been the sensitivity of the systems, which is related to the relatively high acquisition times needed to obtain measurements of sufficient quality. This problem is of greater concern for single layers and/or individual cells, which contain less biological material than tissues and provide lower signals. Therefore, to obtain spectra from cells with signal-to-noise ratio comparable to tissues, the acquisition time is normally on the order of two or even four times higher for cells than for tissue. Moreover, the measurement of cells has normally been a very tedious task because they are frequently spread out in the sample preparations. These facts have hampered the proper analysis of cell spectra and prevented the systematic assessment of their discriminative power.

Recently, modern FTIR microspectroscopes have increased their acquisition speed mainly thanks to

Email address: vnaranjo@upv.es (Valery Naranjo)
[1] Present address: *Phenome Centre Birmingham, School of Biosciences, University of Birmingham, Birmingham, United Kingdom.*

the development of larger and more sensitive imaging sensors. The present work constitutes a proof-of-concept to assess if a modern benchtop FTIR microspectroscope, together with the existing protocols of sample preparation and spectral analysis, are ready to provide a reliable diagnostic system using cytological samples. In particular, the ability of FTIR spectra to differentiate between cells from cultures of four different skin cell lines, including two melanoma cell lines with malignant phenotypes, was studied.

1.1. Related work

Because of the aforementioned difficulties, the application of FTIR microspectroscopy to cytopathological problems is considerably less than the large number of studies and significant advances accomplished in histopathology. Nevertheless, several research groups around the world have pushed, and still push, forward with its application and development.

Until the advent of FTIR microscopes, cytological studies analysed average FTIR spectra recorded from large samples of *cell pellets*, which lacked enough spatial resolution to accurately distinguish cell subpopulations [2]. The first relevant studies of cells by combining FTIR spectroscopes and optical microscopy were performed with synchrotron sources around the beginning of this century [3, 4]. In fact, these kinds of facilities were used in pioneering studies that revealed one of the main problems in FTIR cytology: Mie scattering and Resonant Mie Scattering (RMieS) [5–7]. Improvements in optical components and sensor sensitivities extended the use of FTIR microspectrometers with thermal sources to laboratories. In addition, the introduction of Focal Plane Array (FPA) sensors enabled faster measurements of cell preparations [8].

The most relevant bibliography involving FTIR cell analyses can be found in recent reviews [9, 10]. Regarding the discrimination of cells from the diagnostic point of view, commonly referred to as Spectral Cytopathology (SCP), the most important works using FTIR microscopy are mainly related to Diem's collaborations [11–18]. All these studies are focused on smear cells directly extracted from different parts of the patients rather than in cell cultures. They have covered different types of cancer pathologies: urine [14], cervix [15] or upper respiratory and digestive tract [16–18]. In most of them, cells were deposited on low-e slides and measured in transflection. Due to the preparation of the extracted samples, cells were relatively spread and isolated. A patented method called PapMap [8], which computes a mean spectrum per isolated cell and discards clumped cells, is used in those studies. It should be noted that there is a fair amount of debate regarding the viability of using transflection with samples of varying thickness, due to interference effects between illumination and reflection photons [19–21].

As stated above, most existing studies related to cancer discrimination analyse cells directly extracted from the patient. In principle, this is the ideal methodology to develop decision support systems to assist the clinician in the diagnosis of cytopathological samples. Nevertheless, one of the problems of using these kinds of samples is the need for an explicit labelling by expert pathologists, which is considered the *ground truth*, or *gold standard*. This process is very time-consuming for the experts and is also subject to their interpretation. In addition, most of the extracted cell samples contain debris and undesirable heterogeneities that may diminish the quality and reliability of the cell datasets. These problems, added to the intrinsic difficulties when recording FTIR spectra of cells, have resulted in most existing studies being based on reduced datasets of spectra (e.g., less than 1000 cell spectra). As a consequence, their results are generally founded on just qualitative assessments (e.g., by means of Principal Component Analysis (PCA) scores plots) or quantitative classifications where no clear separations of the training and test sets are stated. Hence, their findings offer limited statistical significance and generalisation capabilities of the models are very difficult to guarantee, especially given the high-dimensional nature of FTIR spectra.

Finally, a preliminary study to discriminate skin cell lines by using a reduced dataset of individual FTIR spectra was carried out in a previous publication [22]. Nevertheless, the relevance of the obtained results was limited, since the number of analysed cells was much lower than in this work. In addition, it did not incorporate the variability due to the measurement and preparation, as it only employed a single hyperspectral image per cell line.

1.2. Objective of the study

The main objective of this study is to assess the capabilities of FTIR spectra to discriminate different skin cell lines, comprising two non-tumoral and

two tumoral types. These cell lines were cultured and fixed in a controlled environment before being measured with a modern state-of-the-art commercial FTIR microspectrometer. The selected cellular models are rather stable and have significant differences in tumorigenicity. Cultured cells belong to catalogued cell lines with approximately *constant* characteristics within the same populations, which reduces the uncertainty linked to the establishment of a reliable *ground truth*. Thus, they provide more homogeneous genotypic and morphologic characteristics during a series of experiments than cell or tissue samples from different patients. In addition, cell cultures provide higher spatial densities of cells, which increases the efficiency for recording larger amounts of data. Thus, these cultured cells should potentially constitute a suitable standard model for the evaluation of the processing and classification of FTIR cell spectra with higher levels of statistical significance than current related studies.

Nevertheless, some technical problems and additional difficulties arise when dealing with cell cultures. Some of these particularities introduce confounding artefacts that may mislead the discrimination. The aim of this work is to apply different data analysis methodologies that diminish those critical biases and promote a discrimination based on the genuine biochemical information of the cell lines. Special efforts will be made to evaluate the generalisation capabilities and robustness of those methodologies against possible fluctuations in experimental conditions. This variability, introduced both during sample preparation and spectral acquisition, will be also considered during the discrimination analysis. Therefore, another goal of this work is to identify the possible limitations of the current measurement and sample preparation protocols.

2. Materials and methods

2.1. Discrimination pipeline

Fig. 1 shows the flow diagram with the main steps that were followed to discriminate skin cell lines, including melanoma and non-melanoma cells, based on their FTIR spectra. This process starts with the preparation of cultured samples of catalogued skin cell lines and their measurement with a modern FTIR microspectroscope in order to obtain the hyperspectral images. From these images, the pixels associated with cells were separated from the non-cell pixels so that only useful spectra were retained for subsequent steps.

FTIR spectra extracted from the retained pixels were individually preprocessed by different techniques to normalise their values and remove unwanted variations which may mislead the posterior quantitative analysis. A mean spectrum was computed for each cell in order to reduce the complexity of the dataset and try to mitigate remaining undesirable effects in the preprocessed spectra of individual pixels, such as random noise. The anomalous or extreme values in the dataset of mean cell spectra were filtered out to promote stability and reduce bias in the dimensionality reduction and classification algorithms.

An exploratory analysis was performed in the resulting dataset in order to study the main trends within the data. Finally, the kept mean cell spectra went through a process of feature extraction and supervised classification where different alternatives were explored, too. This final process of classification was subject to a methodology called *nested Cross-Validation (CV)* consisting of two *loops*, which separates the training and testing subsets of spectra in order to avoid over-fitting and give a reliable measurement of the performance of the constructed classification algorithms. All of these steps will be described in detail in the following sections.

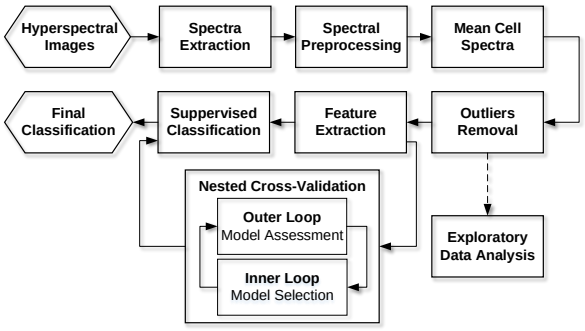


Figure 1: Flow diagram of the main steps applied for the discrimination of skin cells.

2.2. Hyperspectral images

2.2.1. Cell culture and sample preparation

Tab. 1 summarises the cell lines used in this study. With the use of these cell lines, the two major cellular skin constituents, keratinocytes (HaCaT) and fibroblasts (NIH-3T3), together with two skin cancer cell types (A-375, SK-MEL-28) are represented. All cell lines were cultured in DMEM (Sigma-Aldrich) supplemented with 10% FBS (PAN-Biotech), 2 mM

Table 1: Information about the cell lines used in this study.

Cell line	Species	Origin	Type	Provider	Catalogue number
A-375	Homo sapiens	Skin	Malignant melanoma	CLS	300110
HaCaT	Homo sapiens	Skin	Keratinocytes	CLS	300493
NIH-3T3	Mus musculus	Embryo	Fibroblasts	DSMZ	ACC 59
SK-MEL-28	Homo sapiens	Skin	Malignant melanoma	CLS	300337

L-Glutamine (Lonza) and 1 mM sodium pyruvate (Lonza) at 37 °C and 5% CO_2. All cell lines tested negative for mycoplasma. For FTIR analysis, cells were grown on CaF_2 windows (grade VUV, $12.5 \times 12.5 \times 1.5$ mm^3, CRYSTAL GmbH) for 24 hours using silicone micro inserts (4 well, ibidi). These inserts enable to separate cultivation of up to four cell lines on the same window. After removing the inserts, CaF_2 windows were rinsed carefully with PBS to remove loose cells. Remaining cell layers were fixed with glutaraldehyde (1% in PBS) for 30 min, dehydrated in an ascending ethanol series, and airdried. More detailed information about the cell culture and the sample preparation can be found in previous publications [23–25].

2.2.2. FTIR measurements

The measurement system consisted of an Agilent 620 FTIR microscope coupled to an Agilent 670 FTIR spectrometer with a Globar® light source and a liquid nitrogen cooled FPA detector with 128×128 pixels. IR light was transmitted through a 15× Cassegrain reflective condenser and objective (NA = 0.62) giving an effective pixel size of 5.5×5.5 µm^2 and a corresponding field of view (FOV) of 704×704 µm^2 per frame. Six frames per sample were taken and combined, covering a horizontal and vertical area of around 2.1×1.4 mm^2 that included the whole cell culture. The measurements were carried out in transmission mode and absorption spectra were acquired between 1000-3900 cm^{-1} with a wavenumber interval of 4 cm^{-1}. The images of the reference backgrounds (taken from regions of empty substrate) and the images of the cell samples were created by co-adding 256 and 128 scans, respectively.

2.2.3. Batches definition

Cells from each cell line were independently cultured and later seeded and fixed separately in specific regions of a CaF_2 window forming a *cell sample*. The separation of these regions were delimited by silicone inserts or *moulds* with four rectangular holes that were attached to the CaF_2 window before cell seeding and were removed after fixation. Fig. 2 shows the representative grayscale images obtained from the FTIR hyperspectral images of all the cell samples that were prepared and measured for this study. As can be observed, three samples per cell line were measured giving a total of twelve cell samples.

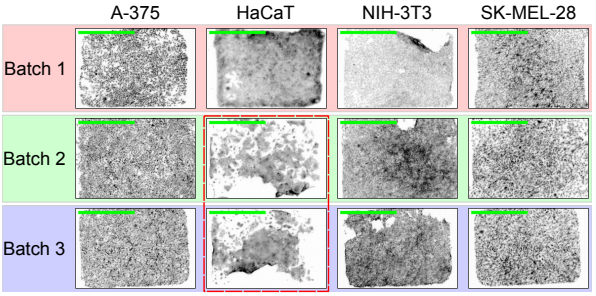


Figure 2: FTIR grayscale images of the measured samples of skin cultured cell lines. Images from different cell lines are arranged by columns and from different batches by rows. The red dotted line reminds that images of HaCaT cells from *Batch 2* and *Batch 3* were cultured in a separate CaF_2 window. Green scale bars represent 1 mm.

These samples have been arranged in different *batches* according to preparation and measurement criteria:

- The samples within *Batch 1* were cultured at the same time and later seeded and fixed in the same CaF_2 window. They were also measured during the same day.
- The samples within *Batch 2* and *Batch 3*, with the exception of the samples of HaCaT cell line, were cultured at the same time (around one year and a half later than cells from *Batch 1*) and were also seeded and fixed in two separate CaF_2 windows. HaCaT samples of *Batch 2* and *Batch 3* were later cultured (3 months later) and seeded and fixed in a separate CaF_2 window due to problems during the prepara-

tion of that cell line in the original CaF_2 windows of those batches. Finally, all the samples from *Batch 2* and *Batch 3* (including HaCaT cells) were measured during the same day in the FTIR microspectrometer, around nine months after acquiring the images of *Batch 1*. This time delay was due to the complexity of preparing the cell cultures and several difficulties during the measurement logistics (e.g., cell cultures were prepared and measured in laboratories of different countries).

Both preparation and measurement factors introduce variations in the FTIR spectra, which determine the strategies of the quantitative analysis and will also have influence in the obtained results.

2.3. Spectra extraction

The first step in the analysis of the FTIR hyperspectral images is to separate pixels containing cell structures from those containing only substrate. This separation was obtained by a binary mask, which was automatically calculated by applying Otsu's method [26] to a representative grayscale image computed from the FTIR data cube. Further details of this procedure can be found in the Electronic Supplementary Information (ESI).

2.4. Spectral preprocessing

Spectral preprocessing is normally recognised as a key step in quantitative analysis of FTIR spectra, especially for cytological studies [27–29]. In order to better understand the need of spectral preprocessing, Fig. 3 shows some raw spectra corresponding to pixels from different positions of a FTIR hyperspectral image of NIH-3T3 cell line. Those pixels, whose effective size is 5.5×5.5 μm^2, have been overlaid on independent white light images of higher resolution aligned with the FTIR hyperspectral images. Additional examples of raw spectra corresponding to pixels from different regions of the four studied cell lines can be found in the ESI.

As can be observed, even spectra from nearby pixels in the same image can be highly variable. Most of the *gross* variations are mainly determined by differences in concentrations and optical path lengths (thickness), which commonly result in offsets and multiplicative factors. Nevertheless, more complex distortions can be observed in some cases, showing the pattern distinctive of Mie scattering and RMieS [7], e.g., complex baselines, lower ratio between the Amide I peak (∼1600-1700 cm^{-1},

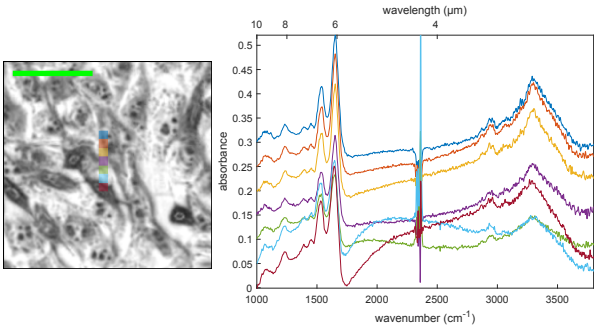


Figure 3: Example of raw spectra extracted from a FTIR hyperspectral image of NIH-3T3 cell line. Left plot: white light image with a vertical line of selected pixels overlaid in different colors. Green scale bar represents 50 μm. Right plot: absorbance spectra corresponding to each pixel with the same color. A proper offset has been added to each spectrum for clarity.

with maximum around 1650 cm^{-1}) and the Amide II peak (∼1500-1600 cm^{-1}, with maximum around 1545 cm^{-1}), or derivative-like depressions beyond the Amide I peak. By looking at the white light images, those complex spectra can be associated with regions where the cells are more compact and rounded. Hence, they have characteristics that produce Mie scattering phenomena [5].

This morphological heterogeneity, which gives rise to critical variations in the FTIR spectra, can be related to different stages of the cell cycle, as was demonstrated in previous studies [4, 30–32]. Those studies showed that cells in stages *G2* and close to mitotic phase *M* have more compact structures that increase the presence of RMieS artefacts. They also demonstrated that FTIR spectra from cells of the same cell type could be discriminated attending to their stage in the cell cycle. In addition, some other cells may have entered into apoptosis, due to the high levels of cell stress that can appear locally in the cell culture, acquiring compact and rounded configurations too.

The main aim of spectral preprocessing is to normalise and standardise the spectra so that they can be categorised based on biochemical properties, rather than unwanted physical properties (thickness, concentrations, morphology, etc) that increase the spectral variability and presumably can confound the discrimination. The raw spectra together with four different preprocessing methods have been considered in order to study their possible influence in the discrimination of cell lines:

- *Raw*: original spectra without preprocessing.

- *Min-Max*: Min-Max normalisation. Each spectrum is scaled to between 0 and 1.
- *SNV*: Standard Normal Variate [33].
- *DiffSG1*: 1st order differentiation by Savitzky-Golay filter (2nd order polynomial and 19 fitting points) and vector normalisation.
- *RMieS-EMSC*: *Resonant Mie Scattering-Extended Multiplicative Signal Correction* algorithm [34–36] after 20 iterations. It was found empirically that 20 iterations is a conservative number that generally gives a stable correction.

Traditionally, the first three preprocessing methods (Min-Max, SNV and DiffSG1) have been mainly applied in FTIR spectroscopic measurements of tissues, although they have been also employed for cell spectra [37]. Nevertheless, since the characterisation of RMieS artefacts [7] and the publication of RMieS-EMSC algorithm [34], it has generally become the default preprocessing method in cytological studies.

The spectral range was cropped to the *fingerprint region* (1000-1800 cm^{-1}). This restriction was applied to the raw spectra and before preprocessing except for the case of RMieS-EMSC, whose spectra were cropped after preprocessing so that the Mie scattering baselines could be modelled and corrected more accurately with the information of higher wavenumbers [6]. The RMieS-EMSC algorithm also down-weighted the 2200-2550 cm^{-1} region associated with CO_2 perturbations.

2.5. Mean cell spectra

The vast majority of cytological studies based on FTIR hyperspectral images compute a mean or average spectrum per cell in order to reduce complexity and increase the robustness of later analyses. These mean spectra are computed after (instead of before) preprocessing the raw spectra of individual pixels independently in order to compensate for uneven spectral distortions due to the heterogeneous spatial properties of cells [31].

This approach is in line with *object-based* or *object-oriented* classification approaches, which are very popular in the analysis of hyperspectral images in remote sensing [38]. Object-based approaches incorporate a certain level of spatial information that generally improves the performance of *pixel-wise* or *pixel-by-pixel* classification.

In order to compute mean cell spectra, it is essential to delimit the regions of pixels belonging to each individual cell. The segmentation of cells was performed with a methodology similar to the one proposed by Filik *et al.* [39], which applies the marker-controlled watershed transformation [40–42], with some modifications, as further described in the ESI. Finally, a mean cell spectrum per segmented region was computed by averaging the preprocessed absorbance values of all the pixels' spectra of that segmented region along the wavenumber dimension.

2.6. Outliers removal

Outliers can seriously bias and deteriorate the performance of the learning framework. The correct way of proceeding with outliers is attempting to determine and address their causes in order to detect possible failures in the followed methodologies [43]. In this problem, there are many sources that may cause anomalous samples, starting from the cell culture preparation, the acquisition of spectra and finally in the previous steps of the classification pipeline (e.g., spectral artefacts introduced or not corrected during preprocessing). Due to the complexity of the problem, the limited technological maturity of FTIR microspectroscopy and the lack of standardised and universally-recognised analysing protocols, it is reasonable to remove all possible perturbations that may be caused by outliers and leave the (difficult) task of analysing outliers for future studies.

The whole dataset of mean spectra per segmented cell region was studied to detect outliers. The preprocessing option RMieS-EMSC was taken as a reference to define those cell regions whose mean spectra are anomalous. This process was performed by using PCA decomposition (retaining the first principal components (PCs) that accounted for 99% of total variance) and Mahalanobis distance (see ESI for further details). The cell regions with mean spectra preprocessed by RMieS-EMSC and detected as outliers were discarded in all the preprocessing alternatives described in Sec. 2.4 in order to keep the same set of cells for the final analysis.

2.7. Feature extraction

Feature extraction techniques aim to reduce the number of variables included in the classification models, to decrease their complexity and increase their stability. The problem of high dimensionality in hyperspectral images is commonly known as

the *Hughes phenomenon* [44] in the remote sensing field [45, 46]. In summary, this phenomenon states that as the number of dimensions increases, the effectiveness of the classifier decreases. The main reason is that the number of parameters involved in the classification model increases with the number of dimensions and, for a fixed sample size, the uncertainty in the estimation of those parameters become wider [45]. PCA [47, 48] and Partial Least Squares (PLS) [43, 49] were the methods used to reduce the dimensionality of FTIR spectra and try to prevent this problem. The number of retained components K is a hyperparameter that was optimised both for PCA and PLS in the inner loop of the *nested CV* (see Sec. 2.9 and the ESI).

2.8. Supervised classification

Supervised classification is the task of *pattern recognition* [50, 51] or *machine learning* [52] whose goal is to create models capable of predicting or inferring the *labels*, *classes* or *groups* of the samples from a dataset, taking as inputs a set of measured *features*. These models must be constructed or *trained* with a set of samples whose labels or classes are known. In this problem, the *features* are the *scores* of the components retained in the feature extraction step. These scores are used to predict the *label*, i.e., the *cell line*, of each sample or cell. As there are four different cell lines, this study can be categorised as a multiclass classification problem. Two main aspects must be defined in the supervised classification step, namely the specific classification algorithm and the measurement used to assess its performance.

2.8.1. Classification algorithms

Three classification algorithms or techniques, very popular in the chemometrics field [49, 53], were explored: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Partial Least Squares Discriminant Analysis (PLS-DA). The main difference is in the definition of the class boundary; hence, these three alternatives have been considered to diminish their possible influence in the final discrimination. LDA and QDA were preceded by PCA as a feature extraction step. PCA was not required for PLS-DA as feature extraction is an inherent part of the algorithm already.

2.8.2. Assessment metric

Another important component in classification is the metric for its evaluation. As the problem in hand is a multiclass problem, a suitable assessment metric is the *overall accuracy* [45, 54], which condenses the global classification performance of all classes. As will be presented later (Tab. 2), a small imbalance exists between the four classes of the dataset, reaching almost a 3:1 ratio in some cases (e.g., NIH-3T3/HaCaT in *Batch 2*, see Tab. 2). To avoid favouring the larger classes [55, 56], the *Balanced Accuracy* (BA) was used instead of the overall accuracy 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:

$$\mathrm{BA} = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{c_{ii}}{\sum_{j=1}^{N_c} c_{ij}} \tag{1}$$

where N_c is the number of classes (4 in this case) and c_{ij} is the number of spectra of class i classified as class j. Hence, c_{ii} is the number of cell spectra correctly classified for the class i.

2.9. Nested cross-validation

One of the main concerns in classification is to obtain a reliable measure of the performance of the developed models. In order to avoid, or at least diminish, over-fitting and assess generalization ability of the classifier, the test and training sets of spectra must be very well defined and separated during the whole process of training, optimisation and assessment of the learning framework [49, 52]. This task was accomplished by applying a hierarchical CV approach, called *nested CV*, which consists of two *loops*: an *outer loop* for assessing the constructed classification models and an *inner loop* for training and optimising those learning models. More details about these two loops can be found in the ESI. Following this structure, two cross-validations alternatives were applied with the aim of checking the dependency of the discrimination capabilities of FTIR spectra on sample preparation and measurement conditions.

2.9.1. Cross-validation alternatives

Once the CV method has been chosen, it is important to decide how the test, training and validation sets will be created from the whole dataset of mean cell spectra that were retained after the removal of outliers. The prediction capabilities of the classification models are determined by the similarities between the spectra from the test sets and the training sets. In order to assess those similarities,

two approaches were followed to separate the mean cell spectra of each kind of set:

- *One-Batch-Out CV*: consecutively, the spectra extracted from the images of one batch (Fig. 2) are considered as the test set in the *outer loop* and the classification algorithms are trained with the spectra of the other batches in the *inner loop*. The process is repeated so that each batch is considered once as the test set. The results for each test batch are combined to provide the final performance measurement (a single value of BA). The idea behind this approach is to assess the uniformity of cell lines between batches and to check if there may be some critical factors for correct discrimination (e.g., the sample preparation procedure or the measurement conditions).

- *In-Batch CV*: as shown in Fig. 4 for *Batch 1*, each image is split in 5 vertical stripes with an equal number of segmented cells. Each vertical stripe of the same color is considered as the test set and the remaining stripes form the training set. This is repeated until all stripes have been in the test set once. The main reason to group cells in vertical stripes, instead of randomly choosing them for the test sets, is to try to construct synthetic subimages where the spatial variability within the same cell culture is also assessed. When this process is finished for one batch, it is repeated for the remaining batches independently. The results for each test stripe and batch are finally combined to provide the final performance measurement (a single value of BA). The aim of this approach is to assess the discrimination between cell lines inside each batch and to compare the performance with the *One-Batch-Out CV* approach.

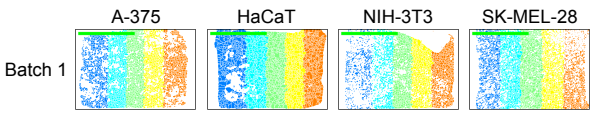


Figure 4: Sketch of the *In-Batch CV* approach. Each image is split in 5 vertical stripes with an equal number of segmented cells; CV is performed independently within each batch by alternately considering one of the stripes from each cell line as the test set and the rest of stripes as the training set. Green scale bars represent 1 mm.

3. Results

3.1. Exploratory data analysis

Before presenting the final classification results, it is very useful and even advisable to perform an exploratory analysis of the data in search for trends or patterns that may justify the outputs of the classification. This analysis is centred on the inputs of the final classification module, that is, the mean cell spectra after outliers removal (Fig. 1). This task consists of visualising the main outputs of descriptive statistics and PCA.

3.1.1. Descriptive statistics

Morphological information. The information about the final number of segmented cell regions and their corresponding number of pixels retained after outliers removal for each cell line and batch are presented in Tab. 2. A total of 22700 cells were retained for the final analysis. As can be observed, the number of cells varies between cell cultures.

The information of the total number of retained cells and pixels can be complemented by the histograms of pixels per cell shown in Fig. 5. The majority of cells cover less than 20 pixels but there are slight differences between samples and cell lines. HaCaT cells tend to form more homogeneous monolayers and adopt flatter configurations than the other cell lines, which is reflected in more uniform histograms. On the contrary, malignant cell lines (A-375 and SK-MEL-28) and in a less extent NIH-3T3 cells normally grow in a more proliferative and disordered way. As a result, those cell lines present higher proportions of smaller cells, which also tend to be more rounded and compact and, hence, more liable to produce scattering artefacts.

Spectral information. The mean spectra for each cell line and preprocessing alternative are presented in Fig. 6. As can be observed, the differences between cell lines change depending on the preprocessing. In fact, the order of presentation of the preprocessing alternatives corresponds with the level of differences observed. This order also agrees with the complexity of each preprocessing method and the level of transformation that individual spectra experiment. In this sense, Min-Max alternative broadly removes the offsets and applies a basic normalisation. Therefore, with this preprocessing the differences observed between cell lines are likely due to Mie scattering artefacts and, by extension, to cell morphology as evidenced by the uneven ratios of

Table 2: Information about the retained cells after outliers removal. Number of segmented cellular regions (first number) and number of pixels inside them (second number) for each cell line (columns) and batch (rows). The last row and column present the corresponding total marginal values.

	A-375	**HaCaT**	**NIH-3T3**	**SK-MEL-28**	**Total**
Batch 1	2247 / 35473	1638 / 41525	1750 / 31524	2495 / 34610	8130 / 143132
Batch 2	2438 / 31314	966 / 23367	2656 / 45348	1931 / 35875	7991 / 135904
Batch 3	1867 / 27906	875 / 20609	2318 / 45377	1519 / 32022	6579 / 125914
Total	6552 / 94693	3479 / 85501	6724 / 122249	5945 / 102507	22700 / 404950

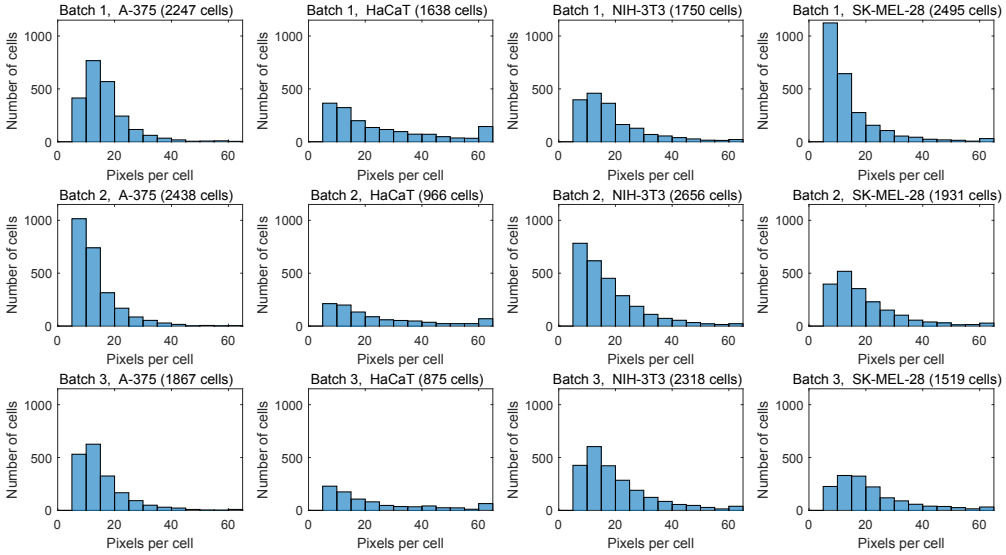


Figure 5: Histograms of pixels per cell in the final retained cells of each sample.

Amide I-Amide II peaks and remaining baselines. Either way, it must be demonstrated if the reduction of undesirable artefacts with the most complex preprocessing alternatives is necessary and effective enough to discriminate, or improve the discrimination of, cell lines. More figures showing the central tendency and dispersion of spectra from each batch can be found in the ESI.

3.1.2. Principal component analysis

The capabilities of PCA for dimensionality reduction can be used to visualise the intrinsic structures of hyperspectral datasets. Because PCA is an unsupervised method (no information of the cell class is used), their plots provide useful details about the *proximity* or *similarity* between spectra and the possible *natural* groups that they may form. Following this line, PCA can be used to extract preliminary information from the different datasets which each cross-validation alternative will have to deal with.

In the case of *One-Batch-Out CV*, it is interesting to explore the whole dataset of mean cell spectra in order to study if there are relevant differences between the batches. The most important plots after applying PCA to the whole dataset of retained mean cell spectra are shown in Fig. 7. Again, RMieS-EMSC has been chosen as the reference preprocessing because it is the most advanced alternative. In Fig. 7a, 100 mean cell spectra from each cell line and batch have been randomly selected to construct the score plots of the 2 first PCs. In that plot, points from the same cell line have been drawn with the same color and same symbols have been used to identify points from each batch. As can be observed, some subgroups of the same cell lines seem to appear mainly from non-tumoral HaCaT and NIH-3T3, which spread less than melanoma cell lines A-375 and SK-MEL-28. Despite those local subgroups, a global separation between cell lines is not so evident.

In Fig. 7b the points belonging to each cell line

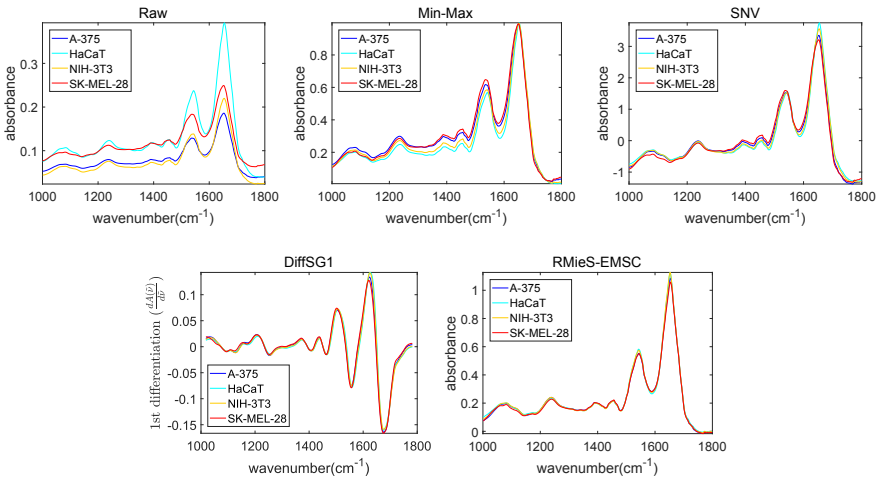


Figure 6: Mean spectra computed with the datasets of average spectra per cell region that were retained after outliers removal for each preprocessing alternative (each subfigure) and cell line (see legends).

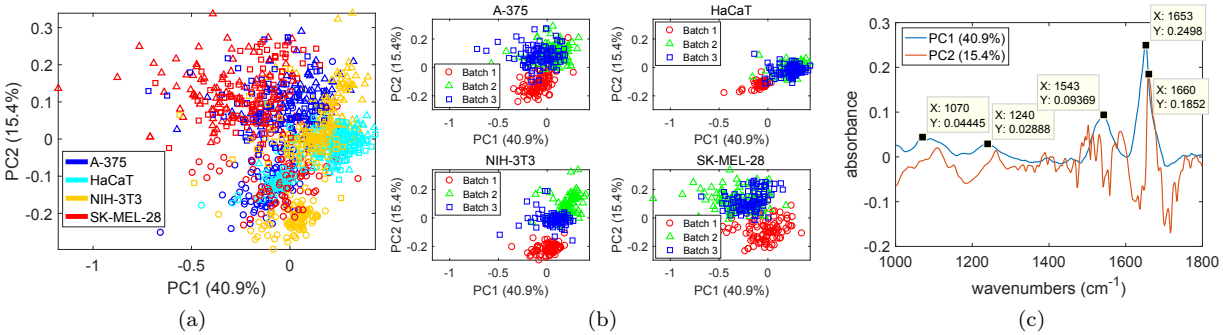


Figure 7: Principal component analysis of the whole dataset of retained mean cell spectra preprocessed by RMieS-EMSC. (a) Score plots of the two first PCs with the corresponding percentage of explained variance in parentheses where 100 mean cell spectra for each cell line (depicted with different colors) and batch (depicted with different symbols) have been randomly chosen from the whole dataset. (b) Same score plots of subfigure (a) but showing only the randomly selected spectra of each cell line; symbols used to denote the batch are the same as in subfigure (a) but colors have been changed to better identify differences between batches. (c) Loadings of the 2 first PCs with the corresponding percentage of explained variance in parentheses.

have been isolated in distinct subplots in order to better distinguish differences between batches within each cell line. In those subplots, batches have been denoted by the same symbols used in Fig. 7a but assigned different colors. As can be seen, spectra from *Batch 1* clearly tend to cluster separately from the rest of batches in all cell lines. However, points from *Batch 2* and *3*, with the exception of NIH-3T3, are much more overlapping. If these subplots are compared with Fig. 7a, now spectra from *Batch 1* can be better identified in the lower part of the graphs. In that batch, NIH-3T3 spectra seem more isolated than the other cell lines. In addition, HaCaT cells from *Batches 2* and *3* can be better distinguished in the right central region more or less mixed with NIH-3T3 spectra. Finally, spectra from melanoma cell lines A-375 and SK-MEL-28 of *Batches 2* and *3* are much more tangled in the upper-central region of the chart.

Fig. 7c displays the corresponding loading vectors of the 2 first PCs. As observed, the highest weights of the first PC (PC1), which explains 40.9% of the total variance, are located around the Amide I peak (∼1600-1700 cm^{-1}, with maximum at 1653 cm^{-1}) and Amide II peak (∼1500-1600 cm^{-1}, with maximum at 1543 cm^{-1}). Other stronger weights occur around 1070 cm^{-1} and 1240 cm^{-1}. The second PC (PC2) explains 15.4% of the total variance and it gives important clues about the presence of two spectral artefacts. The first artefact is RMieS,

which can cause shifts in the maxima's position of the strongest absorption peaks and may be responsible for the difference in the position of the maximum weight (1660 cm^{-1}) with respect to PC1. The second artefact is likely to be water vapour, which may be responsible for the *high-frequency* fluctuations throughout the range 1300-1800 cm^{-1}.

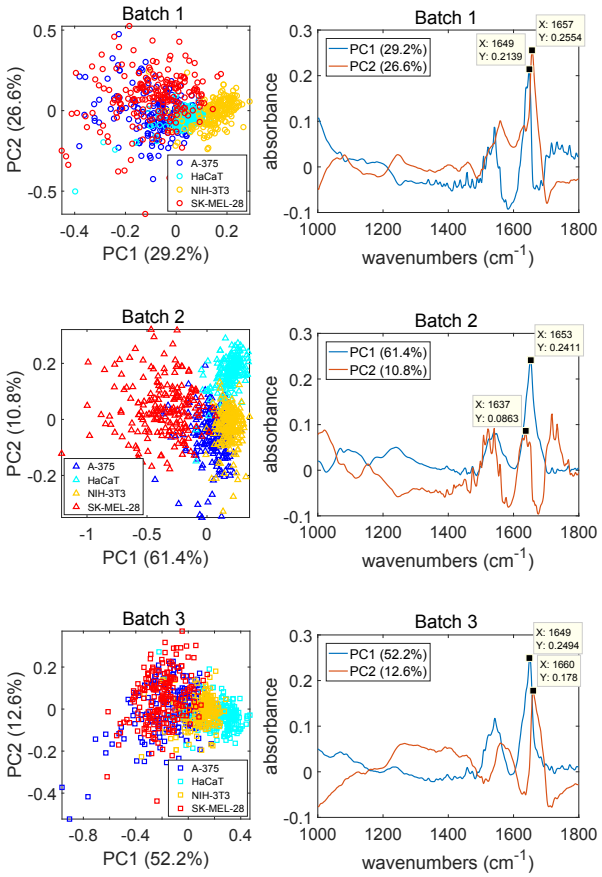


Figure 8: Principal component analysis of the subsets of retained mean cell spectra from each batch (rows) preprocessed by RMieS-EMSC. First column: Score plots of the two first PCs, with the corresponding percentage of explained variance in parentheses, where 200 mean cell spectra for each cell line (depicted with different colors) have been randomly chosen from the corresponding batch's subset. Second column: Loadings of the 2 first PCs with the corresponding percentage of explained variance in parentheses.

In the case of *In-Batch CV*, the similarities between cell lines within each batch may be inferred from the described study of the whole dataset. However, it may be more convenient to analyse each batch independently in order to reduce the interferences of other batches during the computation of the main directions of variation. Fig. 8 shows the most relevant plots when applying PCA to the subsets of retained mean cell spectra preprocessed by RMieS-EMSC from each batch.

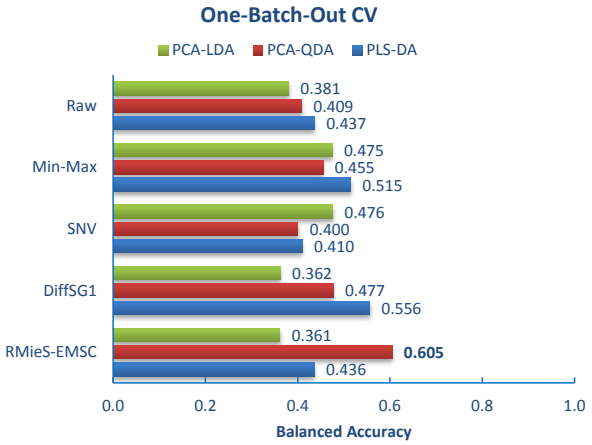
The score plots of the 2 first PCs of each batch (first column of Fig. 8) reveal congruent information with the study of the whole dataset about the relative dispersion of each cell line: HaCaT and NIH-3T3 tend to form more compact subgroups, meanwhile A-375 and SK-MEL-28 are relatively more spread. Concerning the separation between cell lines, slight differences can be observed within each batch: in *Batch 1*, NIH-3T3 seems to be the most isolated cell line followed by HaCaT, which has partial overlaps with the other cell lines; in *Batch 2*, HaCaT and SK-MEL-28 spectra seem well separated but there is a significant overlap between A-375 and NIH-3T3; in *Batch 3*, HaCaT and NIH-3T3 seem to overlap only partially with the rest of cell lines, meanwhile A-375 and SK-MEL-28 spectra appear more mixed, as in *Batch 1*.

The loading vectors of the 2 first PCs (second column of Fig. 8) again highlight the problems of RMieS, which may be responsible for the shifts in the maximum weights around the Amide I peak, and water vapour. In particular, fluctuations likely due to water vapour seem stronger in *Batch 2*, clearly distorting the maximum weights. Finally, although in *Batch 1* the first PC is less relevant than in the rest of batches, more than 80% and 95% of the total variance can be explained by respectively retaining 5 and 15 PCs in all batches.

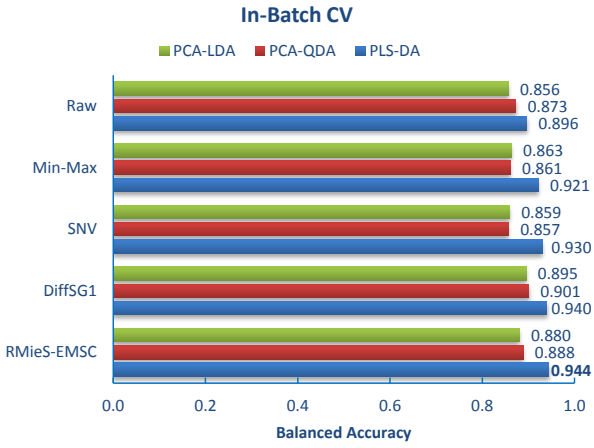
3.2. Classification results

The preliminary descriptive information of the most relevant trends are based on two principal components and must be contrasted with objectively-assessed classification models that consider the whole dataset of cells and a larger number of components. Fig. 9 shows graphically the final classification results in terms of Balanced Accuracy (BA) for the combinations of explored preprocessing options (Sec. 2.4), classification algorithms (Sec. 2.8) and cross-validation alternatives (Sec. 2.9). Each final BA was computed by combining the predicted labels of the corresponding test sets of mean cell spectra.

At first glance, there is an evident difference between the ranges of BA values obtained by the two CV approaches. Meanwhile in *One-Batch-Out CV* BA values range from around 0.35 to approximately 0.6 (combination RMieS-EMSC and PCA-QDA), in *In-Batch CV* all the combinations provide a final BA between 0.85 and 0.95 (combina-



(a)



(b)

Figure 9: Classification results in terms of Balanced Accuracy for the different preprocessing options (grouped by rows), classification algorithms (colors) and cross-validation alternatives: (a) *One-Batch-Out CV* and (b) *In-Batch CV*.

tion RMieS-EMSC and PLS-DA). Regarding preprocessing, there is no clearly prevailing option and the final performance also depends on the employed classification algorithm. Although the maximum BA values in both CV alternatives have been reached by a combination including RMieS-EMSC.

If now we focus on the best combinations of each CV alternative, further details may be obtained about the underlying reasons to get those final BA values. The aim is to gain deeper information about why the misclassification may occur even in the best scenario.

Firstly, it is worth studying the optimisation curves of the number of retained components, computed with the corresponding validation sets as was

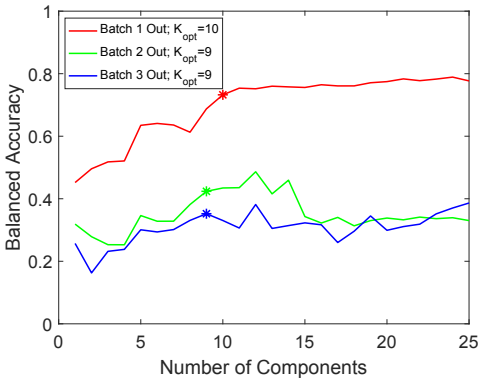


Figure 10: Optimisation curves for the best combination (RMieS-EMSC and PCA-QDA) of *One-Batch-Out CV*.

described in Sec. 2.9. Figures 10 and 11 show the optimisation curves for the best combinations of *One-Batch-Out CV* and *In-Batch CV*, respectively. In *One-Batch-Out CV*, a remarkably higher performance (around double BA) is reached when the *Batch 1* is left out as test set and algorithms are trained and validated only with the spectra from *Batch 2* and *3* than when one of those batches is left out and the other one forms the initial training set with *Batch 1*. This fact confirms that there are more similarities between spectra from *Batch 2* and *3* than with those from *Batch 1*, which is in line with the preliminary studies by PCA (Fig. 7). On the other hand, the optimisation curves of *In-Batch CV*, individually constructed for each batch by leaving out vertical stripes of cells in the hyperspectral images, present much more similar characteristics between them. Even so, *Batch 2* provides a slightly lower performance than the other batches but in all cases BA is above 0.9. This suggests a high degree of similarity between the spectra from the same hyperspectral image and, hence, from the same cell line, as well as relevant differences with other cell lines within the same batch.

Finally, again for the best combinations (highest BA) of each CV alternative, the predicted cell labels in the test sets were joined to create pseudo-color images by using the color code shown in Fig. 12a. Figures 12b and 12c respectively present those images for *One-Batch-Out CV* and *In-Batch CV* with the same distribution of batches (rows) and cell lines (columns) as in Fig. 2. Note that cells within each image should be coloured with the color specified outside each column for a correct classification. These images provide more detailed information about the position of the misclassified spectra.

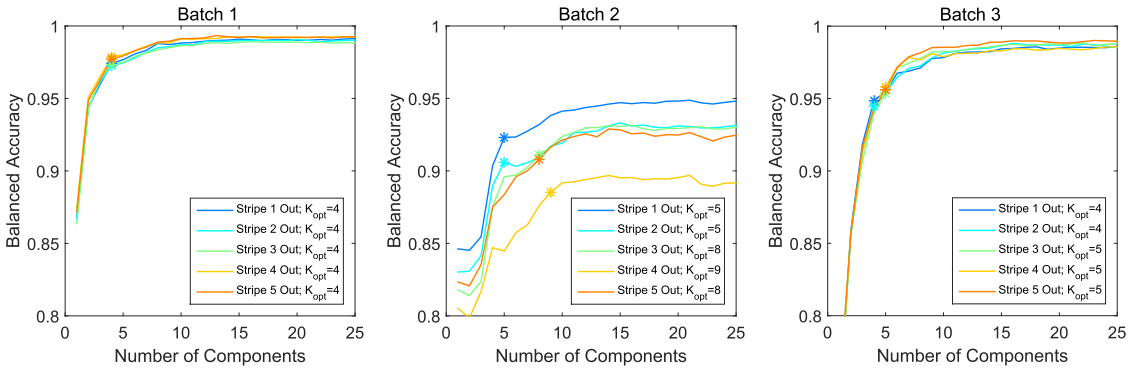


Figure 11: Optimisation curves for the best combination (RMieS-EMSC and PLS-DA) of *In-Batch CV*.

When analysing Fig. 12b, some observations can be stated about the classification performed by the best combination of *One-Batch-Out CV*:

- *Batch 1*: it presents the highest rate of misclassification. A significant amount of malignant A-375 cells are incorrectly labelled as the benign NIH-3T3. The worst case is the non-tumoral HaCaT cell line, whose cells are mostly *identified* as the tumoral SK-MEL-28. A small number of NIH-3T3 cells are confounded with cancerous A-375 cells. Lastly, almost half of SK-MEL-28 cells are wrongly classified as A-375.
- *Batch 2*: the malignant A-375 cell line of this batch is another case with extreme misclassification, again with the benign NIH-3T3. However, the rest of cell lines are very well categorised.
- *Batch 3*: A-375 cells are also mistaken with NIH-3T3 but in a lower degree than in Batch 2. The rest of cell lines present high rates of correct classification, although some cells from NIH-3T3 and SK-MEL-28 are mainly confused with the same cell types as in *Batch 1*.
- *Overall*: all batches classified A-375 and NIH-3T3 either correct or as the corresponding mesenchymal cell line. A-375 melanoma cells are highly dedifferentiated and have a mesenchymal phenotype like fibroblasts (see extended discussion in the ESI). The *One-Batch-Out CV* has therefore a strong performance in the classification of cell types: mesenchymal (A-375, NIH-3T3), epithelial (HaCaT) and melanocyte (SK-MEL-28) rather than in the identification of specific cell lines.

Finally, the images of the best combination of *In-Batch CV* (Fig. 12c) confirm the good general classification of most cells. The only remarkable inaccuracies are mainly located in the malignant A-375 and the benign NIH-3T3 cell lines of *Batch 2*, whose cells are mutually confused with the other cell line.

4. Discussion

The excellent classification results obtained in the *In-Batch CV* suggest a high potential of FTIR spectra to discriminate the skin cell lines, at least inside each batch. Nevertheless, those high rates of success are obtained independently of the preprocessing technique and even for the raw spectra. This fact warns that the results of the *In-Batch CV* may be too optimistic and some level of over-fitting may be playing a critical role. Those suspicions are supported by the less favourable results obtained in the *One-Batch-Out CV* alternative, which better assesses the generalisation capabilities of the discrimination framework.

Indeed, a deeper study of the most optimistic discriminative option reveals that critical differences exist, especially between cells from *Batch 1* and the other batches. These findings are also supported by the unsupervised PCA score plots of the exploratory analyses. Taking into account that cells from *Batch 1* were prepared and measured at time points different from *Batch 2* and *3* (Sec. 2.2) and, although similar sample preparation and measurement protocols were followed for all batches, they are likely to introduce confounding factors in FTIR spectra which are critical for the correct discrimination of cell lines. Nonetheless, when the experimental conditions are more comparable, as in

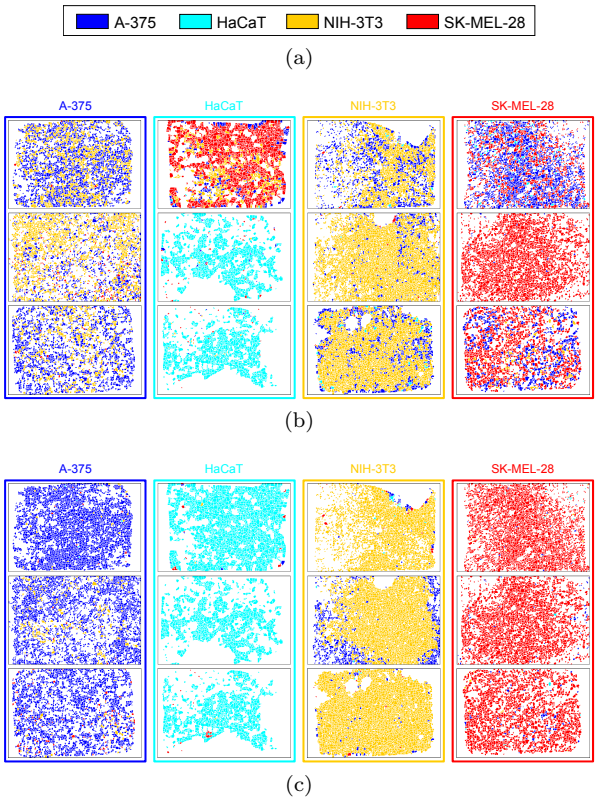


Figure 12: Qualitative results. (a) Color code for the predicted labels of cells in the images of qualitative results. (b) Qualitative results for the best combination (RMieS-EMSC and PCA-QDA) of *One-Batch-Out CV*. (c) Qualitative results for the best combination (RMieS-EMSC and PLS-DA) of *In-Batch CV*.

Batch 2 and *3*, the extrapolation of the discrimination capabilities is more satisfactory.

These findings raise the question of whether the discrimination of cells in this problem is critically biased by the experimental factors. These factors are mixed with the rest of physicochemical information in the final hyperspectral images and discovering the exact sources of these perturbations is not easy. Despite this, the exploratory analysis of preprocessed spectra has revealed the likely undesirable presence of water vapour effects even in the loadings of first PCs. These water vapour fluctuations, which highly depend on the environmental conditions during the spectral acquisition, could not be avoided even when purging the sample area with dry air. Those interferences, whose relative importance is higher in spectra with lower absorbance, increase their impact in the global dataset due to the spectral normalisation.

The exploratory analyses also offered interesting information about the efficiency and consequences of the RMieS-EMSC algorithm, which is currently the most advanced preprocessing method available with a practical implementation. In this study, 20 iterations of the algorithm were used, which took around 4 weeks to process the whole dataset of extracted individual spectra (around 400 thousand in total). However, remaining shifts in the most relevant weights of the PCA loadings around the Amide I peak suggest a suboptimal correction of RMieS artefacts. These artefacts, associated with morphological differences, are accentuated by cell culture conditions, where cells have sustained nutrients, support and enough space to proliferate, favouring the presence of mitotic stages. Differences in growing and proliferation properties between cell lines are responsible for uneven sizes and shapes, with tumoral cell lines more prone to being smaller and more compact. Therefore, Mie and RMieS artefacts may have also played determinant confounding roles in the discrimination of cell lines. For example, this factor may be responsible for the misidentification of A-375 cells as NIH-3T3 and vice versa, although this misclassification may not be an artefact of cell preparation or classification methods but rather be caused by the high similarity of these mesenchymal cell types (see extended discussion in the ESI).

The identification of cell types is a crucial effort towards an identification and localisation of skin components. The structure of cell type distribution is the relevant information for skin pathology, similar to the state of the art hematoxylin and eosin staining. The misclassification of specific cell lines (e.g., A-375 cells classified as NIH-3T3) is critical for a cell biological application where an identification of cell lines would be desirable. From a diagnostic point of view, the detection of cell type distribution and the much higher information content of tissue samples may be sufficient for the development of a digital pathology system. Hence, the current technologies and methodologies has the potential for a transfer to the clinical practice but needs more training and the reduction of preparation related informations.

5. Conclusions

The potential application of the current state-of-the-art FTIR technology to cytopathological diagnosis has been assessed. A dataset of approxi-

mately 22700 cultured cells, derived from two tumoral and two non-tumoral skin cell lines, has been analysed to evaluate the discrimination capabilities of current FTIR technology in cytopathological problems. This number of analysed cells is relatively high when compared with existing studies. The analysed cells were distributed across three different batches, which were prepared and measured at different time points by following similar protocols to maximise uniformity between batches. However, these controlled protocols are limited by current technological and procedural restrictions, e.g., the lack of an isolation chamber for the measured samples. Therefore, external perturbations were inevitably introduced both during sample preparation and spectral acquisition.

Different methodologies and approaches have been explored to process FTIR hyperspectral images and correctly analyse cell spectra throughout the whole discrimination pipeline. Excellent discrimination results are obtained when the algorithms are trained and tested with cell spectra from the same batch. However, those results for a specific cell line identification are not confirmed when cells from different batches are mixed to construct the algorithms and they are finally applied to a different batch. This fact questions the capabilities of the discriminative properties of FTIR spectra of low content cell biological samples, which seems to be critically influenced by the differential factors between batches, such as the sample preparation protocol or the measurement conditions (e.g., different water vapour content).

Experimental design of future studies should assess if the good results observed between cell lines of the same batches are driven by genuine biochemical differences rather than artificial perturbations introduced by sample preparation and measurement protocols. In any case, the high rates or misclassification for some cell lines in different batches warn about the need of a better standardisation of the aforementioned protocols, together with the specific analysis methodologies. This is fundamental to develop and establish a reliable technology for cell biological studies based on FTIR microspectroscopy in the near future.

Further development of standardisation protocols is needed, which will also have to incorporate and adapt to the existing routine procedures of biomedical laboratories. Additional improvements can be introduced in the different steps of the proposed discrimination pipeline apart from the experimental improvements. Regarding spectral analysis, it is expected that the most relevant factors may be the preprocessing and the supervised classification framework. In the case of preprocessing, more complex techniques for the removal of interferences such as water vapour should be explored in case that measurements could not be performed in isolated environments. Authors would like to point out that different trials were made to remove or at least reduce the water vapour interferences by spectral processing. However, unsuccessful results were obtained due to the pixel variability in water vapour signals introduced by FPA detectors as well as their high sensitivity to the environmental conditions. In addition, new advances in algorithms for the correction of RMieS, with higher computing speed, more stability and robustness against perturbations (e.g., again water vapour), should be incorporated.

Finally, the biomedical relevance and the extrapolation of the results obtained with cultured cells are limited because they are grown in artificial conditions and have a rather low information content. Therefore, the development of a reliable diagnostic decision support system will necessarily have to use FTIR spectra of patient cells and tissues with a reference *ground truth* provided by pathologists, such as in histopathological studies. The here presented data demonstrate the current capabilities, limitations and future possibilities of FTIR spectroscopy for cell biological and pathological applications. A solution of the detected problems in sample preparation and data classification would open a wide field of biomedical applications in the identification of cellular and pathophysiological processes.

Conflicts of interest

None Declared.

Acknowledgements

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

Author contributions

FP acquired the FTIR images, performed the image and data analysis, and wrote the first draft of the manuscript. VN contributed to the image and

data analysis, the writing of the first draft of the manuscript and the conception of the idea of the study. GRL contributed to the image and data analysis and to the correction of the first draft of the manuscript. LK and BK created the cell cultures and all the supporting documentation. JS contributed to the conception of the idea of the study. JN supervised and collaborated in the acquisition of the FTIR images. NS contributed to the conception of the idea of the study. All authors were involved in discussions about the detailed design of the methodology and experiments. All authors read, corrected and approved the final manuscript.

References

[1] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, F. L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, Nature Protocols 9 (8) (2014) 1771–1791. doi:10.1038/nprot.2014.110.

[2] M. A. Cohenford, B. Rigas, Cytologically normal cells from neoplastic cervical samples display extensive structural abnormalities on IR spectroscopy: Implications for tumor biology, Proceedings of the National Academy of Sciences 95 (26) (1998) 15327–15332. doi:10.1073/pnas.95.26.15327.

[3] N. Jamin, P. Dumas, J. Moncuit, W.-H. Fridman, J.-L. Teillaud, G. L. Carr, G. P. Williams, Highly resolved chemical imaging of living cells by using synchrotron infrared microspectrometry, Proceedings of the National Academy of Sciences 95 (9) (1998) 4837–4840.

[4] H.-Y. N. Holman, M. C. Martin, E. A. Blakely, K. Bjornstad, W. R. Mckinney, IR spectroscopic characteristics of cell cycle and cell death probed by synchrotron radiation based Fourier transform IR spectromicroscopy, Biopolymers 57 (6) (2000) 329–335. doi:10.1002/1097-0282(2000)57:6<329::AID-BIP20>3.0.CO;2-2.

[5] B. Mohlenhoff, M. Romeo, M. Diem, B. R. Wood, Mie-Type Scattering and Non-Beer-Lambert Absorption Behavior of Human Cells in Infrared Microspectroscopy, Biophysical Journal 88 (5) (2005) 3635 – 3640. doi:http://dx.doi.org/10.1529/biophysj.104.057950.

[6] A. Kohler, J. Sulé-Suso, G. D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. G. van Pittius, G. Parkes, H. Martens, Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction, Applied Spectroscopy 62 (3) (2008) 259–266. doi:10.1366/000370208783759669.

[7] P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, P. Gardner, Resonant Mie scattering in infrared spectroscopy of biological materials - understanding the 'dispersion artefact', Analyst 134 (2009) 1586–1593. doi:10.1039/B904808A.

[8] J. M. Schubert, A. I. Mazur, B. Bird, M. Miljković, M. Diem, Single point vs. mapping approach for spectral cytopathology (SCP), Journal of Biophotonics 3 (8-9) (2010) 588–596. doi:10.1002/jbio.201000023.

[9] G. Clemens, J. R. Hands, K. M. Dorling, M. J. Baker, Vibrational spectroscopic methods for cytology and cellular research, Analyst 139 (18) (2014) 4411–4444. doi:10.1039/C4AN00636D.

[10] J. Doherty, G. Cinque, P. Gardner, Single-cell analysis using Fourier transform infrared microspectroscopy, Applied Spectroscopy Reviews 52 (6) (2017) 560–587. doi:10.1080/05704928.2016.1250214.

[11] M. Diem, M. Miljković, B. Bird, T. Chernenko, J. Schubert, E. Marcsisin, A. Mazur, E. Kingston, E. Zuser, K. Papamarkakis, et al., Applications of infrared and Raman microspectroscopy of cells and tissue in medical diagnostics: Present status and future promises, Journal of Spectroscopy 27 (5-6) (2012) 463–496. doi:10.1155/2012/848360.

[12] M. Miljković, B. Bird, K. Lenau, A. I. Mazur, M. Diem, Spectral cytopathology: new aspects of data collection, manipulation and confounding effects, Analyst 138 (14) (2013) 3975–3982.

[13] M. Diem, Infrared microspectroscopy of cells and tissue in medical diagnostics, in: Modern Vibrational Spectroscopy and Micro-Spectroscopy, Wiley Online Library, 2015, Ch. 13, pp. 283–338.

[14] B. Bird, M. J. Romeo, M. Diem, K. Bedrossian, N. Laver, S. Naber, Cytology by infrared microspectroscopy: Automatic distinction of cell types in urinary cytology, Vibrational Spectroscopy 48 (1) (2008) 101–106. doi:10.1016/j.vibspec.2008.03.006.

[15] J. M. Schubert, B. Bird, K. Papamarkakis, M. Miljković, K. Bedrossian, N. Laver, M. Diem, Spectral cytopathology of cervical samples: detecting cellular abnormalities in cytologically normal cells, Laboratory investigation 90 (7) (2010) 1068–1077. doi:10.1038/labinvest.2010.72.

[16] K. Papamarkakis, B. Bird, J. M. Schubert, M. Miljković, R. Wein, K. Bedrossian, N. Laver, M. Diem, Cytopathology by optical methods: spectral cytopathology of the oral mucosa, Laboratory Investigation 90 (4) (2010) 589–598. doi:10.1038/labinvest.2010.1.

[17] D. Townsend, M. Miljković, B. Bird, K. Lenau, O. Old, M. Almond, C. Kendall, G. Lloyd, N. Shepherd, H. Barr, et al., Infrared micro-spectroscopy for cytopathological classification of esophageal cells, Analyst 140 (7) (2015) 2215–2223. doi:10.1039/c4an01884b.

[18] M. Diem, M. Miljković, B. Bird, A. I. Mazur, J. M. Schubert, D. Townsend, N. Laver, M. Almond, O. Old, Cancer screening via infrared spectral cytopathology (scp): results for the upper respiratory and digestive tracts, Analyst 141 (2) (2016) 416–428. doi:10.1039/C5AN01751C.

[19] J. Lee, E. Gazi, J. Dwyer, M. D. Brown, N. W. Clarke, J. M. Nicholson, P. Gardner, Optical artefacts in transflection mode ftir microspectroscopic images of single cells on a biological support: the effect of backscattering into collection optics, Analyst 132 (2007) 750–755. doi:10.1039/B702064C.

[20] P. Bassan, H. J. Byrne, J. Lee, F. Bonnier, C. Clarke, P. Dumas, E. Gazi, M. D. Brown, N. W. Clarke,

P. Gardner, Reflection contributions to the dispersion artefact in FTIR spectra of single biological cells, Analyst 134 (2009) 1171–1175. doi:10.1039/B821349F.

[21] J. Filik, M. D. Frogley, J. K. Pijanka, K. Wehbe, G. Cinque, Electric field standing wave artefacts in ftir micro-spectroscopy of biological materials, Analyst 137 (2012) 853–861. doi:10.1039/C2AN15995C.

[22] F. Peñaranda, V. Naranjo, L. Kastl, B. Kemper, G. R. Lloyd, J. Nallala, N. Stone, J. Schnekenburger, Multivariate Classification of Fourier Transform Infrared Hyperspectral Images of Skin Cancer Cells, in: 2016 24th European Signal Processing Conference (EUSIPCO), 2016, pp. 1328–1332. doi:10.1109/EUSIPCO.2016.7760464.

[23] L. Kastl, C. E. Rommel, B. Kemper, J. Schnekenburger, Standardized cell samples for midIR technology development, in: Proc. SPIE, Vol. 9315, 2015, p. 931507. doi:10.1117/12.2080251.

[24] L. Kastl, B. Kemper, G. R. Lloyd, J. Nallala, N. Stone, V. Naranjo, F. Peñaranda, J. Schnekenburger, Potential of mid IR spectroscopy in the rapid label free identification of skin malignancies, in: Proc. SPIE, Vol. 9703, 2016, p. 970307. doi:10.1117/12.2208897.

[25] L. Kastl, B. Kemper, G. R. Lloyd, J. Nallala, N. Stone, V. Naranjo, F. Peñaranda, J. Schnekenburger, Performance of mid infrared spectroscopy in skin cancer cell type identification, in: Proc. SPIE, Vol. 10060, 2017, p. 1006006. doi:10.1117/12.2253314.

[26] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66. doi:10.1109/TSMC.1979.4310076.

[27] P. Lasch, Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging, Chemometrics and Intelligent Laboratory Systems 117 (2012) 100 – 114. doi:10.1016/j.chemolab.2012.03.011.

[28] R. Gautam, S. Vanga, F. Ariese, S. Umapathy, Review of multidimensional data processing approaches for raman and infrared spectroscopy, EPJ Techniques and Instrumentation 2 (1) (2015) 8. doi:10.1140/epjti/s40485-015-0018-6.

[29] H. J. Byrne, P. Knief, M. E. Keating, F. Bonnier, Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells, Chemical Society Reviews 45 (2016) 1865–1878. doi:10.1039/C5CS00440C.

[30] C. Hughes, M. D. Brown, F. J. Ball, G. Monjardez, N. W. Clarke, K. R. Flower, P. Gardner, Highlighting a need to distinguish cell cycle signatures from cellular responses to chemotherapeutics in SR-FTIR spectroscopy, Analyst 137 (2012) 5736–5742. doi:10.1039/C2AN35633C.

[31] M. Jimenez-Hernandez, C. Hughes, P. Bassan, F. Ball, M. D. Brown, N. W. Clarke, P. Gardner, Exploring the spectroscopic differences of Caki-2 cells progressing through the cell cycle while proliferating in vitro, Analyst 138 (2013) 3957–3966. doi:10.1039/C3AN00507K.

[32] D. R. Whelan, K. R. Bambery, L. Puskar, D. McNaughton, B. R. Wood, Synchrotron Fourier transform infrared (FTIR) analysis of single living cells progressing through the cell cycle, Analyst 138 (2013) 3891–3899. doi:10.1039/C3AN00316G.

[33] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, Appl. Spectrosc. 43 (5) (1989) 772–777.

[34] P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, P. Gardner, Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples, Analyst 135 (2010) 268–277. doi:10.1039/B921056C.

[35] P. Bassan, A. Kohler, H. Martens, J. Lee, E. Jackson, N. Lockyer, P. Dumas, M. Brown, N. Clarke, P. Gardner, RMieS-EMSC correction for infrared spectra of biological cells: Extension using full Mie theory and GPU computing, Journal of Biophotonics 3 (8-9) (2010) 609–620. doi:10.1002/jbio.201000036.

[36] P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J. H. Shanks, M. Brown, N. W. Clarke, P. Gardner, FTIR microscopy of biological cells and tissue: data analysis using resonant Mie scattering (RMieS) EMSC algorithm, Analyst 137 (2012) 1370–1377. doi:10.1039/C2AN16088A.

[37] J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott, F. L. Martin, Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives, Analyst 137 (2012) 3202–3215. doi:10.1039/C2AN16300D.

[38] T. Blaschke, M. Kelly, H. Merschdorf, Object-Based Image Analysis: Evolution, History, State of the Art, and Future Vision, in: Remotely Sensed Data Characterization, Classification, and Accuracies, Remote Sensing Handbook, CRC Press, 2015, pp. 277–294. doi:10.1201/b19294-20.

[39] J. Filik, A. V. Rutter, J. Sule-Suso, G. Cinque, Morphological analysis of vibrational hyperspectral imaging data, Analyst 137 (2012) 5723–5729. doi:10.1039/C2AN35914F.

[40] F. Meyer, S. Beucher, Morphological segmentation, Journal of Visual Communication and Image Representation 1 (1) (1990) 21–46. doi:http://dx.doi.org/10.1016/1047-3203(90)90014-M.

[41] P. Soille, Morphological Image Analysis, Springer Nature, 2004. doi:10.1007/978-3-662-05088-0.

[42] R. C. Gonzalez, R. E. Woods, Digital Image Processing (3rd Edition), Pearson, 2007.

[43] M. J. Adams, Chemometrics in Analytical Spectroscopy, 2nd Edition, RSC Analytical Spectroscopy Series, The Royal Society of Chemistry, 2004. doi:10.1039/9781847550484.

[44] G. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE Transactions on Information Theory 14 (1) (1968) 55–63. doi:10.1109/TIT.1968.1054102.

[45] P. Mather, B. Tso, Classification Methods for Remotely Sensed Data, 2nd Edition, CRC Press, 2009.

[46] J. Li, A. Plaza, Hyperspectral Image Processing: Methods and Approaches, in: Remotely Sensed Data Characterization, Classification, and Accuracies, Remote Sensing Handbook, CRC Press, 2015, pp. 247–258. doi:10.1201/b19294-18.

[47] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems 2 (1) (1987) 37 – 52. doi:10.1016/0169-7439(87)80084-9.

[48] R. Bro, A. K. Smilde, Principal component analysis, Analytical Methods 6 (9) (2014) 2812–2831.

[49] R. Brereton, Chemometrics for pattern recognition, Wiley, 2009.

[50] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd Edition, Academic Press, Boston, 1990. doi:10.1016/B978-0-08-047865-4.50001-6.

[51] C. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag New York, Inc, 2006.

[52] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd Edition, Springer-Verlag New York, Inc, 2009. doi:10.1007/978-0-387-84858-7.

[53] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Boca Raton, FL, 2009.

[54] R. G. Congalton, K. Green, Assessing the Accuracy of Remotely Sensed Data, 2nd Edition, CRC Press, 2008.

[55] H. He, E. A. Garcia, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284. doi:10.1109/TKDE.2008.239.

[56] S. Wang, X. Yao, Multiclass Imbalance Problems: Analysis and Potential Solutions, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42 (4) (2012) 1119–1130. doi:10.1109/TSMCB.2012.2187280.