Multispectral imaging via feature selection: a frugal innovation approach for pathogen identification

Denis Leroux1, Manuel Petit1, Corinne Davenas2, and Corine Fulchiron2

1Technology Research on Infectious Diseases | Open Innovation & Partnerships Dept., bioMérieux, Marcy l’Etoile, France
2R&D Microbiology, bioMérieux, La Balme Les Grottes, France

Abstract. In order to develop an affordable clinical diagnostic instrument for use in more decentralized settings, we have assessed the feasibility to move from hyperspectral to multispectral imaging via parsimonious feature selection. The targeted application is the label-free identification at the species-level of uropathogens from images of bacterial colonies on their growth support. We show that the number of predictors (i.e., discrete spectral channels), can be dramatically reduced from 240 to less than 10 channels with limited performance loss. The impact of bandwidth is also investigated to consider the high degree of redundancy of raster images obtained by diffuse reflectance and propose a suitable design for a simple filter wheel based solution. Targeting the 8 most prevalent bacterial species responsible for > 80% of urinary tract infections, up to 94% of correct identification rate was reached using only 4 spectral windows.

1 Introduction

The analysis of diffuse reflectance via hyperspectral imaging is known to be very efficient in identifying microbial colonies at the species or at the Gram level [1-3]. This work aims to develop a low-cost system to identify bacterial species causing Urinary Tract Infections (UTI) from images of bacterial colonies, directly on the culture medium. UTI, beside its clinical importance, is a great case study as only a few uropathogens species are responsible for most infections. This work focuses on assessing the performance of multispectral imaging by performing a parsimonious analysis from hyperspectral data to reduce the number of predictors (aka ‘channels’) to a minimum while keeping the performance level high enough for a Minimum Viable Product (MVP). To this extent, we have defined a minimum Classification Identification Rate of 94%, (CIR = no. correct ID / total no. ID). Simply the question is “can we reach a performance of identification high enough for In-Vitro-Diagnostics applications while relying on a minimum number of predictors?”.

Biological samples. Bacteria originated from a bioMérieux strain collection, referred as ‘pure strain’. Eight species constituted our biological model with 3 strains cultivated in triplicates on COS (Columbia agar + 5% sheep blood, bioMérieux) for 18 to 24h of incubation. Included per class: Enterococcus faecalis (SP1), Escherichia coli (SP2), Klebsiella pneumoniae (SP3), Proteus mirabilis (SP4), Proteus vulgaris (SP5), Pseudomonas aeruginosa (SP6), Staphylococcus aureus (SP7), and Streptococcus agalactiae (SP8). A total of 1745 colonies were analyzed.

1.1 Materials and Methods

2 Materials & methods

Hardware. The hyperspectral imaging hardware system is constituted of a hyperspectral camera (Pika II, Resonon, USA) equipped with a CMOS type detector sensitive in the 394 to 893 nm VNIR region with a 2.1 nm spectral resolution (240 spectral channels total). Hyperspectral data were acquired by scanning the culture plate on a moving stage in line scan mode.

2.1 Materials and Methods

Multivariate analysis and feature selection. Radiance spectra averaged over individual colonies are extracted from hypercubes of a culture plate (Fig. 1) Data preprocessing of shown data consists in applying a moving average of order 4, on reflectance data, thus translating in an 8.4 nm bandwidth per channel. Only raw data are reported here. Support Vector Machine with a radial kernel (C = 100, G = 0.01) was used as classification algorithm, in a one-vs-one mode. Data partition ratios for calibration/test sets and the train/validation sets were of 2/1 and 10/1 respectively. Parsimonious analysis was conducted using (i) the full VNIR spectral range from 394 to 893 nm and (ii) the visible range VIS only from 394 to 750 nm. Only data acquired in the VIS range is discussed here. The chosen FSSVM feature selection strategy consists in evaluating iteratively subsets with a growing number of predictors, in a combinatorial way, with the stepwise selection

* Corresponding author: denis.leroux@biomerieux.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
criteria being to maximize the classification performance in a 10-fold cross validation radial SVM.

![Fig. 1. Agar plate with selected colonies in red (left), typical raw radiance spectra per species (right).](image)

### 3 Results & discussion

**Feature selection.** Surprisingly, only 4 iterations (Fig. 2) were sufficient to reach an error target below our MVP target of 6%; while hyperspectral performance (1% error) is reached after only 7 iterations. Starting at the 6th iteration, multiple predictors provide an identical incremental increase of performance, in the range of our model precision (1%).

![Fig. 2. ‘Cumulated ER’: a 5.3% error is measured for a sparse matrix made of 4 predictors at 747, 619,595 and 630 nm.](image)

**Bandwidth impact.** We define here bandwidth as any band (i.e., combination of continuous channels) of a given width that carry the same amount of information. Binning was used to estimate the impact of reducing the number of predictors and simulate the impact of using filters of different bandwidths using the full spectrum range (i.e., hyperspectral modality). Our motivation is to check for data redundancy and estimate what maximal bandwidth could be used in a filter-based system, to minimize bandpass filter costs and allow more light per band to pass through. Reducing the number of predictors from 220 to 55 by computational binning (4-fold binning) did not reduce performance (Fig. 3), and only a small performance drop was observed up to a 12-fold binning. Thus, an acceptable bandwidth was calculated (no. of bins x 2.1 nm) to be in the [8, 25 nm] interval.

![Fig. 3. Binning vs. error rate: optimal moving average between 4 and 12 translates into a [8nm, 25nm] bandwidth interval](image)

**Retrospective exploratory analysis** was conducted a posteriori by plotting boxplots of raw reflectance at the 3 prominent selected features values, with the goal of eyeballing thresholds values capable of separating the 8 species classes in a hierarchical decision tree. Indeed, using only 6 reflectance thresholds in a tree of depth 3 seems to allow class separation (Fig. 4). Note that this exercise is not meant to replace our flat SVM approach but only to perform a ‘sanity check’ from the original data. It is also an indication of low overfitting risk.

![Fig. 4. Reflectance boxplots at 3 channels, a-f lines depict possible thresholds to separate classes in a hierarchical tree.](image)

### 4 Conclusions

This parsimonious analysis showed that a dramatic dimension reduction was observed, suggesting that only 4 bands (possibly up to 25 nm wide) could be used to reach MVP performance, and efficiently separate the 8 most prevalent uropathogens. There was no benefit in using the NIR region extension in the 770-893 nm, but 750 nm was found to be the first selected feature in FSSVM analysis. The perspective of developing a low-cost multispectral system based equipped with bandpass filters or based on narrow emission using LEDs seem a reasonable alternative to an expensive hyperspectral system.

### References