Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 290 (2023) 122270

Contents lists available at ScienceDirect

Spectrochimica Acta Part A:
Molecular and Biomolecular Spectroscopy

journal homepage: www.journals.elsevier.com/spectrochimica-acta-part-a-molecular-and-biomolecular-spectroscopy

Automatic classification of *Candida* species using Raman spectroscopy and machine learning

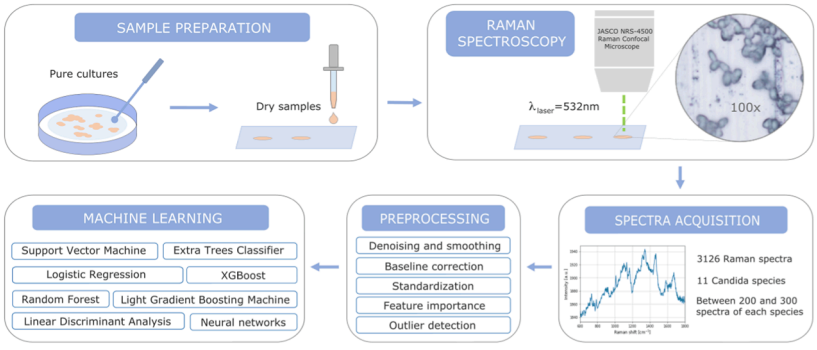
María Gabriela Fernández-Manteca [a,1,*], Alain A. Ocampo-Sosa [a,b,1], Carlos Ruiz de Alegría-Puig [a,b,c], María Pía Roiz [a,b], Jorge Rodríguez-Grande [a,b], Fidel Madrazo [a], Jorge Calvo [a,b,c], Luis Rodríguez-Cobo [a,d,e], José Miguel López-Higuera [a,d,e], María Carmen Fariñas [a,c,f,g], Adolfo Cobo [a,d,e,*]

[a] *Instituto de Investigación Sanitaria Valdecilla (IDIVAL), Santander, Spain*
[b] *Servicio de Microbiología, Hospital Universitario Marqués de Valdecilla, Santander, Spain*
[c] *CIBER de Enfermedades Infecciosas (CIBERINFEC), Instituto de Salud Carlos III, Madrid, Spain*
[d] *Photonics Engineering Group, Universidad de Cantabria, Santander, Spain*
[e] *CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Instituto de Salud Carlos III, Madrid, Spain*
[f] *Servicio de Enfermedades Infecciosas, Hospital Universitario Marqués de Valdecilla, Santander, Spain*
[g] *Departamento de Medicina y Psiquiatría, Universidad de Cantabria, Santander, Spain*

HIGHLIGHTS

- Traditional methods of microbiological identification are complex and time-consuming.
- Raman spectroscopy solves these problems as it is a fast and cheap technique that does not require sample preparation.
- Raman spectroscopy combined with machine learning algorithms has great potential for identifying and classifying pathogenic microorganisms.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Raman spectroscopy
Candida identification
Machine learning
Convolutional neural network
Overfitting

ABSTRACT

One of the problems that most affect hospitals is infections by pathogenic microorganisms. Rapid identification and adequate, timely treatment can avoid fatal consequences and the development of antibiotic resistance, so it is crucial to use fast, reliable, and not too laborious techniques to obtain quick results. Raman spectroscopy has proven to be a powerful tool for molecular analysis, meeting these requirements better than traditional techniques. In this work, we have used Raman spectroscopy combined with machine learning algorithms to explore the automatic identification of eleven species of the genus *Candida*, the most common cause of fungal infections worldwide. The Raman spectra were obtained from more than 220 different measurements of dried drops from pure cultures of each *Candida* species using a Raman Confocal Microscope with a 532 nm laser excitation source.

* Corresponding authors.
E-mail addresses: ma-gabriela.fernandez@alumnos.unican.es (M.G. Fernández-Manteca), adolfo.cobo@unican.es (A. Cobo).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.saa.2022.122270
Received 1 October 2022; Received in revised form 29 November 2022; Accepted 20 December 2022
Available online 22 December 2022

After developing a spectral preprocessing methodology, a study of the quality and variability of the measured spectra at the isolate and species level, and the spectral features contributing to inter-class variations, showed the potential to discriminate between those pathogenic yeasts. Several machine learning and deep learning algorithms were trained using hyperparameter optimization techniques to find the best possible classifier for this spectral data, in terms of accuracy and lowest possible overfitting. We found that a one-dimensional Convolutional Neural Network (1-D CNN) could achieve above 80 % overall accuracy for the eleven classes spectral dataset, with good generalization capabilities.

1. Introduction

Candidiasis is one of the most important fungal infections in hospital settings [1]. This infectious disease may be caused by at least 20 different yeast species of the genus *Candida*: *Candida albicans*, *Candida glabrata*, *Candida tropicalis*, *Candida parapsilosis*, *Candida metapsilosis, Candida orthopsilosis, Candida krusei*, *Candida guilliermondii, Candida fermentati, Candida carpophila, Candida lusitaniae*, *Candida dubliniensis Candida pelliculosa*, Candida *kefyr*, *Candida lipolytica*, *Candida famata*, *Candida inconspicua*, *Candida rugosa*, *Candida norvegensis and Candida auris* [1,2]. Although isolation ratios may vary, *C. albicans* remains the most common etiologic agent, followed by C. *glabrata*, *C. tropicalis*, *C. parapsilosis* and *C. krusei*, accounting for more than 95 % of candidiasis in the last three decades [3]. Candidiasis encompasses a number of infections ranging from skin or mucosal infections to more severe conditions such as invasive candidiasis, including bloodstream infection (candidemia), endocarditis, central nervous system, urinary tract candidiasis, and chronic disseminated candidiasis [4]. Invasive candidiasis is associated with high morbidity and mortality rates, especially in immunocompromised and critically ill patients [5].

On the other hand, resistance to antifungal treatments is another problem reported in the studies of fungal infections [6,7]. The inappropriate and excessive use of antibiotics is one of the leading causes that have generated antimicrobial resistance due to the lack of time and precision in the identification of the pathogenic microorganism causing the infection [7]. Indiscriminate or prolonged antibiotic therapy is a major factor in developing candidiasis [8], and growth in culture is still the gold standard for *Candida* spp. its identification and diagnosis. However, culture-based methods have several drawbacks, such as species growth ratio, risk of inadequate sampling, and possible suppression by antifungal therapy [9].

Alternative techniques not based on cultures and widely used nowadays are the polymerase chain reaction (PCR) or the enzyme-linked immunosorbent assay (ELISA) [10]. Although these techniques are well established and reliable, they require complex sample preparation, high cost, and highly specific professional expertise [11].

Given the growing need for rapid and accurate identification of microorganisms [12], more efficient methodologies have been developed in order to prevent the spread of infection and antibiotic resistance as well as the adverse outcomes due to treatment delay. One example is the MALDI-TOF mass spectrometry technique [13], which is highly successful in identifying clinical samples such as bacteria [14]. This technique provides high reproducibility and labor-free measurements, but has the disadvantage of being time-consuming and expensive.

To face these problems, Raman spectroscopy (RS) has been developed to analyze biological samples by identifying chemical markers. This technique uses the energy of a laser source to irradiate the samples and generate inelastically scattered light. The captured signal contains unique information about the vibrational modes of the excited molecular bonds. It allows the generation of a spectroscopic fingerprint of the sample, providing quantitative and qualitative information to characterize samples in real-time, which would facilitate solving daily clinical problems [15]. In addition, it allows discrimination and classification of microorganisms in samples containing a mixture of species [16] and at the level of a single cell [17].

RS has shown to play an essential role in the field of chemical analysis in all types of samples, highlighting the biomedical field [18–20]. It is a non-destructive, low-cost technique that requires minimal sample preparation. Another of its main advantages is the immediacy of the measurement, around seconds, so it can replace other more laborious techniques conventionally used. Its use could mean skipping the cultivation step, which would be a significant advance in many fields when it comes to quickly find the right therapeutic treatments for patients with severe infections [21].

Currently, machine learning has acquired a great interest in information extraction in sample classification and identification problems and, specifically, applied to Raman spectra [22]. However, as with other spectroscopic techniques, delicate spectral preprocessing is necessary before performing complex analysis. Raman spectra without this pretreatment contain many artifacts, noise, and non-relevant information [23]. This preprocessing includes steps such as normalization, smoothing, baseline correction, or a study of contributing spectral features.

To date, many studies have been published that apply Raman spectroscopy and machine learning to problems of identification and classification of bacterial species. However, studies dealing with fungal infections are limited, but all of them have shown high-impact results. For instance, In Chouthai et al. [24] used Raman spectroscopy to identify and classify-five *Candida* species using Principal Component Analysis (PCA) and Differential Functional Analysis (DFA) with 100 % accuracy. Samek et al. [25] acquired a significant number of Raman spectra of several isolates of *C. parapsilosis* separated by given time intervals to verify the reproducibility of the technique. Pezzotti et al. [26] studied the identification and resistance to antifungal drugs of the *C. auris* species using Raman imaging of the living yeast cells. Witkowska et al. [27] demonstrated that the Surface-enhanced Raman spectroscopy (SERS) technique combined with PCA allows for distinguishing many common fungal pathogens in several minutes. Although these studies have shown the benefits of Raman spectroscopy in fungal infections when combined with machine learning algorithms, the number of species in each study was limited. Further research is still needed to explore the challenge of identifying the many *Candida* species with clinical interest.

In this work, we report the application of Raman spectroscopy combined with machine learning algorithms as a method of rapid identification and classification of different *Candida* species, with an emphasis on the analysis of the quality of the acquired spectra and the generalization capabilities of the models. From 3126 spectra of 11 *Candida* species, we analyze their quality in terms of intra- and inter-class variability at species and isolate level, propose a preprocessing scheme with feature selection, data augmentation, and outliers removal, and train several machine learning and deep learning algorithms to obtain their classification performance figures, in terms of overall accuracy and generalization ability.

2. Materials and methods

2.1. Candida isolates

A total of 67 clinical isolates from 11 different *Candida* species were included in this study. Most isolates were recovered from blood cultures. Other sources of isolation were urine culture, vaginal swab, bronchoalveolar lavage, surgical wound, cerebrospinal fluid, intravascular

Table 1
Candida species analyzed in this study. BC: Blood culture, CSF: Cerebrospinal fluid, ICT: Intravascular catheter tip, BAL: Bronchoalveolar lavage, PJI: Prosthetic joint infection, VS: Vaginal swab, UC: Urine culture.

Candida species	Number of isolates	Isolate code	Source
C. albicans	5	Calb-1	BC
		Calb-2	CSF
		Calb-3	ICT
		Calb-4	VS
		Calb-5	VS
C. dubliniensis	4	Cdub-1	BC
		Cdub-2	BC
		Cdub-3	BC
		Cdub-4	BC
C. glabrata	8	Cgla-1	BC
		Cgla-2	BC
		Cgla-3	BC
		Cgla-4	BC
		Cgla-5	BC
		Cgla-6	BC
		Cgla-7	BC
		Cgla-8	BC
C. guilliermondii	6	Cgui-1	BC
		Cgui-2	BC
		Cgui-3	BC
		Cgui-4	BC
		Cgui-5	BC
		Cgui-6	BC
C. inconspicua	2	Cinc-1	BC
		Cinc-2	BC
C. krusei	7	Ckru-1	BC
		Ckru-2	BC
		Ckru-3	BC
		Ckru-4	BC
		Ckru-5	BC
		Ckru-6	BC
		Ckru-7	BC
C. lusitaniae	5	Clus-1	PJI
		Clus-2	BC
		Clus-3	BC
		Clus-4	BC
		Clus-5	BC
C. metapsilosis	6	Cmet-1	UC
		Cmet-2	UC
		Cmet-3	UC
		Cmet-4	UC
		Cmet-5	BC
		Cmet-6	BC
C. orthopsilosis	4	Cort-1	BC
		Cort-2	BC
		Cort-3	BC
		Cort-4	BC
C. parapsilosis	12	Cpar-1	BC
		Cpar-2	BC
		Cpar-3	BC
		Cpar-4	BC
		Cpar-5	BC
		Cpar-6	VS
		Cpar-7	BC
		Cpar-8	BC
		Cpar-9	VS
		Cpar-10	BC
		Cpar-11	BC
		Cpar-12	BC
C. tropicalis	8	Ctro-1	BC
		Ctro-2	VS
		Ctro-3	BC
		Ctro-4	UC
		Ctro-5	UC
		Ctro-6	UC
		Ctro-7	BC
		Ctro-8	BC

catheter tip and prosthetic joint infection (Table 1). S pecies identification was confirmed by theMALDI-TOF MS system Vitek MS with the SARAMIS (Spectral Archive and Microbial Identification System) database (version 2.0, bioMérieux), and in the case of isolates from the *C. parapsilosis* complex (*C. parapsilosis* sensu stricto, *C. metapsilosis* and *C. orthopsilosis*), identification was also confirmed by ITS-rDNA sequencing [29].

2.2. Sample preparation

On the day before Raman measurement, a single colony of each *Candida* spp. isolate was re-streaked onto Columbia blood agar medium (Oxoid ltd, England). Colonies were incubated overnight at 37 °C under aerobic conditions. Samples for measurement were prepared by taking a full inoculation loop of yeast cells and suspending it in 90 μL of sterile water. Three drops containing 30 μL of each yeast suspension were deposited on a microscope slide wrapped with aluminum foil. Samples were allowed to dry in a Class II biosafety cabinet for approximately 30 min before measurement. Experiments were repeated at least 3 times, meaning that 9 technical and 3 biological replicates, respectively, were used for measurement.

2.3. Raman instrumentation

Spectra were acquired using a Confocal Raman Microscope NRS-4500 (JASCO Inc.) equipped with a 532 nm laser excitation source with a total power of 18 mW. Photodegradation of the sample was avoided using a neutral density filter, with which 25 % of the total power (4.5 mW) was used. The laser excitation source is followed by a Rayleigh scattering rejection filter and a 900 ln/mm grating spectrometer coupled to a high-resolution EMCCD detector (Newton EMCCD, 1600x400 pixels). The detector, with a thermoelectric cooling system (-70 °C), achieves low dark noise and a high signal-to-noise ratio with a spectral resolution of approximately 2 cm^{-1} and a better pixel resolution that was interpolated to 0.5 cm^{-1} with a cubic function. Confocality was achieved by passing the Raman scattered signal through a 17 μm diameter pinhole.

A 100x objective (Olympus MPlan N, 100x/0.9) was used, which produced a laser spot diameter close to 1 μm. The focus was obtained through brightfield viewing of the sample surface by moving the stage until the best focus was achieved. Each spectrum consisted of an average of 5 measurements with an exposure time of 5 s (total time of 25 s) in the range 600–1800 cm^{-1}. The instrument was calibrated for wavelength and intensity using a silicon wafer as a reference.

The spectra were acquired and stored using the JASCO Spectra Manager™ software, which controls the spectral acquisition parameters and stores them for later analysis and interpretation.

2.4. Spectra acquisition details

We used an aluminum foil coating as a substrate to minimize the fluorescence signal in the Raman spectra. We deposited 30 μL of the pure cultures on the substrates and focused the laser beam on the surface of individual yeast cells using the brightfield display provided by the detector. Between 220 and 350 spectra of each species were acquired in random areas of the drops to assess reproducibility and avoid possible bias. Each spectrum was measured at a different spatial point but always on the surface of the yeasts. As the candidas cell bodies have round or oval shapes with the size of several microns, on the order of the laser spot diameter at the focal point, and due to the high numerical aperture of the laser beam, the captured Raman spectra can be considered an integration of the internal cell structures molecular fingerprints over the volume of the cell. We have obtained a dataset of 3126 Raman spectra (each one in the wavenumber range from 600 to 1800 cm^{-1} in 0.5 cm^{-1} steps) of the *Candida* species considered in this work. Fig. 1 shows a schematic view of the acquisition process.

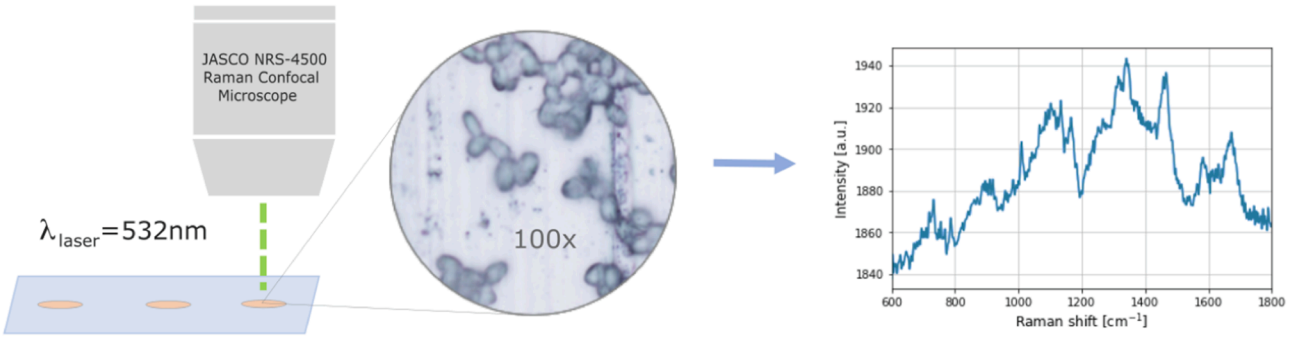


Fig. 1. Schematic view of the *Candida* spectra acquisition process.

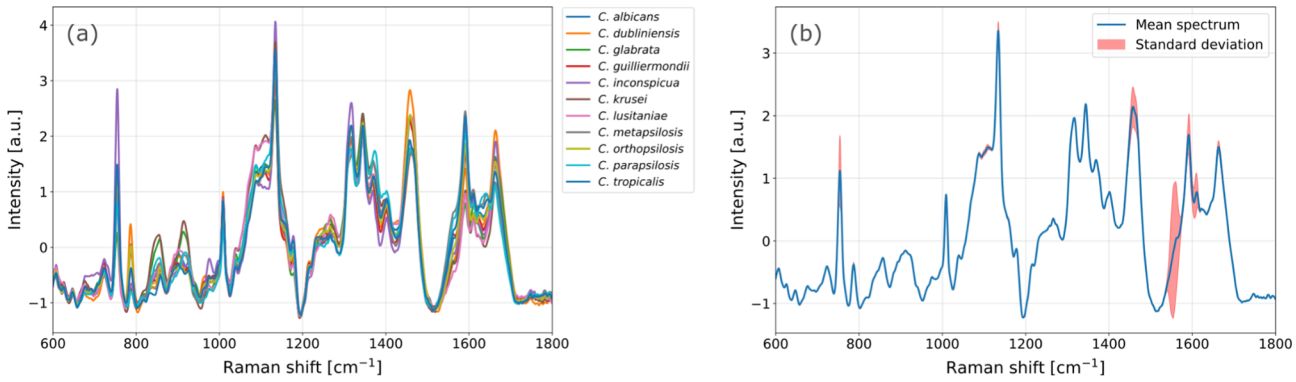


Fig. 2. Mean spectrum of each species (a) and the average of the mean spectrum of each species (b), with variance bands in red (standard deviation value has been multiplied x5 to better highlight variable bands).

2.5. Data preprocessing

The raw spectra comprising 2400 wavenumbers were preprocessed as follows: first, each spectrum was normalized by its energy (dividing by the total optical intensity) to account for variations in the captured Raman scattering. Secondly, a Savitzsky-Golay smoothing filter (2nd order) was applied to reduce noise, using a window parameter of 21 that considers the small wavenumber step of 0.5 cm^{-1} and the spectral resolution. Next, the baseline was estimated and subtracted using the Asymmetric Least Square (ALS) algorithm based on the Whittaker smoother [28] and, finally, a Standard Normal Variate (SNV) transformation (zero mean and standard deviation of one) was applied on a per-spectrum basis.

Fig. 2 shows the mean spectrum of the complete preprocessed dataset with variance bands in red (+/- standard deviation), from the average of the mean spectrum of each species, to accentuate the inter-class variability. This mean spectrum was used to manually extract features, selecting all the apparent peaks in the average spectrum (24 features), the central wavenumber of high variance spectral zones (5 features), and relative minima of the spectrum with no significant variance (14 features), for a total of 43 selected features. Applying feature selection to datasets reduces the risk of overfitting and training time, and the effect of our proposed feature selection on the classification performance is analyzed in Section 3.

The possibility of removing outliers from the dataset was also considered. The isolation forest algorithm [30] was applied to the dataset. This algorithm creates unsupervised ensembles of binary classification trees and scores each spectrum based on its average path length, which is shorter for anomalies. A variable percentage of spectra with higher scores (potential outliers) was removed from the dataset, and the effect on the classification performance was analyzed.

Finally, data augmentation techniques were explored to improve the generalization ability of the models. For this kind of spectral data, different possibilities have been suggested [31]: a linear combination of several spectra, adding noise, or spectral shifting, that could help with long-term instrumental drift. We have evaluated the effect on the classification performance and generalization of the linear combination of three spectra with random weights, and of adding gaussian noise with several amplitudes.

2.6. Machine learning algorithms

This kind of spectral dataset (with hundreds or a few thousand homogeneous numerical features) lies somehow between tabular data (typically with dozens of heterogeneous -continuous numerical and categorical- features) and very high dimensionality data, such as images and videos. For the latter, deep learning algorithms (in particular, convolutional neural networks [32]) have shown a clear advantage in performance in the last years; for tabular data, classical machine learning algorithms such as those based on decision trees (Random Forest, XGBoost…) still can outperform most recent deep learning models [33]. Spectral data poses another challenge: the high number of correlated features, i.e., the problem of multicollinearity, hinders classification performance. As there is not a clear best approach, in this work, we have trained a representative set of machine learning and deep learning algorithms, and their classification performance was evaluated. We used custom code in python and the models included in the SciKitLearn [34] and Keras/Tensorflow [35] libraries. At the same time, the training, hyperparameter optimization and performance estimation were automated using the PyCaret [36] and Optuna [37] libraries.

The tested algorithms were: Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Random Forest (RF), Extra Trees Classifier (ET), XGBoost (XGB), Light Gradient Boosting Machine (LGBM), Multilayer Perceptron (MLP) and a one-dimensional Convolutional Neural Network (1-D CNN).

The complete dataset was randomly split as follows: 20 % of spectra

Table 2
Major Raman bands found in spectra obtained for the identification and classification of *Candida* species and their tentative vibrational assignment based on the provided references.

Raman shift (cm^{-1})	Band assignment	Component	Reference
607	C—C twisting mode of phenylalanine	Proteins	[38]
626	C—C twisting mode of phenylalanine	Proteins	[38]
650	Amino acids	Proteins	[24]
725	Adenine: ring breathing modes of purines	DNA	[38]
787	O—P—O str. thymine, cytosine, uracil	DNA	[25]
855	Ring breathing tyrosine	Proteins	[39]
915	C—O—C str. glucose	Sugars	[24]
1009	C—C skeletal str. aromatic ring phenylalanine	Proteins	[38]
1090	C—N str.	Proteins	[40]
1136	C—O and C—O skeletal str.	Carbohydrates	[41]
1178	C—H wagging tyrosine	Proteins	[42]
1217	Amide III; adenine; polyadenine	DNA/Proteins	[39]
1268	Amide III	Proteins	[25]
1317	Guanine, adenine	DNA	[42]
1345	C—N str. of tryptophan, adenine, guanine	DNA/Proteins	[42]
1374	CH2 wagging; b-1,3 glucans	Lipids/Sugars	[41]
1402	COO^- symmetric str.	Proteins	[39]
1458	C—H deformation	Lipids	[43]
1550	Exopolysaccharides	Carbohydrates	[44]
1592	Cytochrome	Proteins	[45]
1612	C—C phenylalanine, tyrosine, tryptophan	DNA/Proteins	[42]
1642	C═O str. thymine	DNA	[42]
1663	C—O str. amide I	Proteins	[39]
1748	C═O str. esters	Lipids	[46]

as "unseen" to assess the classification and generalization performance, and the remaining 80 % of spectra for training with stratified cross-validation with 10 folds of training and validation data.

3. Results and discussion

3.1. Raman features

From the averaged spectra in Fig. 2, it can be seen a consistency in the spectral information of all the species, with some unambiguous Raman peaks. A tentative identification of the Raman features is shown in Table 2. A visual inspection of the spectra does not suggest apparent distinct features that vary among the species and could help in its classification; only the 787 cm^{-1} (DNA), 1612 cm^{-1} (DNA/proteins), 1642 cm^{-1} (DNA) and 1550 cm^{-1} (*exo*-polysaccharides) peaks show significant inter-species differences.

3.2. Feature importance

The importance of features on several models was studied to determine if any of the Raman bands identified in Table 2 contribute significantly to the classification metrics. The scores of each model were obtained after training with cross-validation and hyperparameters optimization from the spectra with manual feature selection applied. It was found that most of the models with better accuracy values had similar scores. Fig. 3 shows the mean score values of the tested models (LR, LDA, RF, ET and LGBM) with their corresponding standard deviations to get an overview of the influence of these features on the classification performance.

The Raman band with the most significant contribution is centered at 786.5 cm^{-1}, associated with genetic material (see Table 2). This contribution is more than twice the rest of the features considered, all of them having similar scores and diverse associated compositions, such as proteins (855 cm^{-1} band) or sugars (915 cm^{-1} band). The discriminative nature of DNA bands for pathogen identification is in agreement with other works [43,47].

3.3. Quality of spectra and class separability

Firstly, the preprocessed spectra of the eleven species were analyzed to estimate their quality based on their intra-class and inter-class variability, which has a major impact on the achievable classification performance of a given dataset. Intra-class (i.e. intra-species) variance could be attributed to changing experimental conditions, noise, or sample heterogeneity, and has a negative impact on class separability

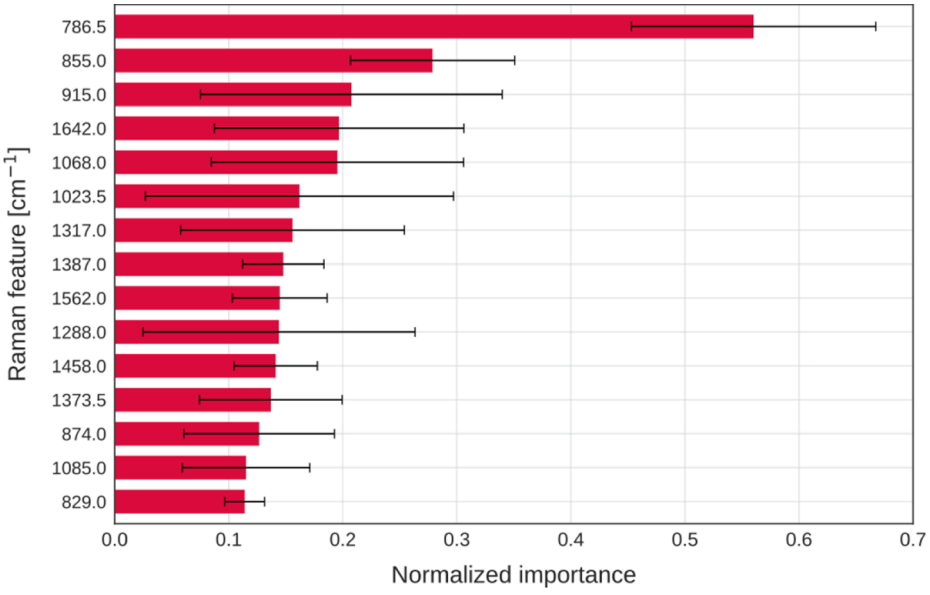


Fig. 3. Mean scores and standard deviations (error bars) of the most important features for several models (LR, LDA, RF, ET and LGBM) trained with feature selection.

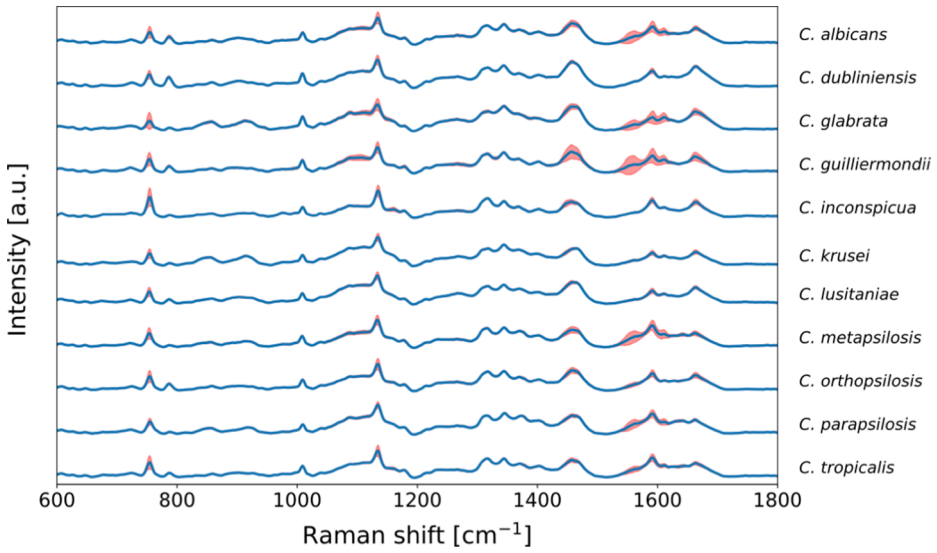


Fig. 4. Average spectrum (with variance overlay) of each *Candida* species.

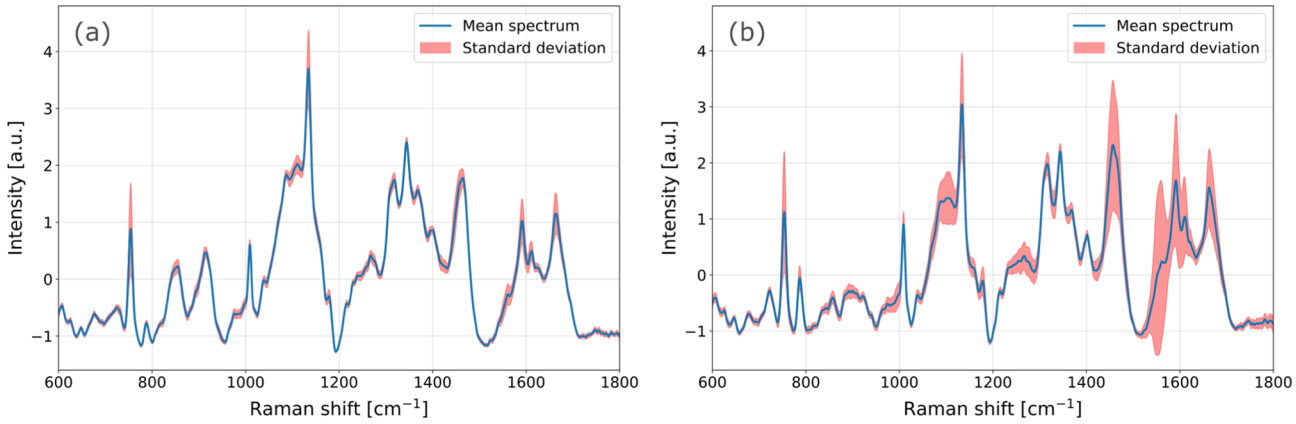


Fig. 5. Average spectrum of *C. krusei* (a) and *C. guilliermondii* (b), which show different levels of intra-class variance.

[48]. Inter-species differences, on the other hand, are needed for proper classification. In Fig. 4, we show the average spectrum of the eleven species (between 220 and 350 spectra each). The variance at each wavenumber is shown as a red overlay (band extension is +/- standard deviation).

It can be seen from Fig. 4 that some species have a low variability of measured spectra (notably, *C. krusei*) while others, in particular, *C. guilliermondii*, exhibit a high variance in several Raman features. The spectrum of those two species is shown in detail in Fig. 5.

The low variance at certain spectral bands in all species suggests repetitive experimental conditions. Furthermore, the highest variability in all species seems to be located in two spectral bands, the peak at 1458 cm^{-1} associated with lipids (see Subsection 3.1 for a tentative assignment of Raman features), and, interestingly, a wide band around 1550 cm^{-1} with no clear peaks or features. This fact could be attributed to a higher spatial heterogeneity in the molecular fingerprints of the samples, or to inter-isolates differences. Different isolates are known to come from different patients, but there is no a priori information about genetic or gene expression differences among them. For example, *Candida* spp. are known to secret substances under certain conditions [49] that can introduce spectral changes in different isolates or species under the same experimental conditions, thus affecting the performance of any classifier.

To obtain more information about the origin of this variability (spatial heterogeneity or inter-isolates differences), we have performed the following procedure: for each species, the spectra of each of their isolates were averaged, and the resulting mean spectrum was plotted with variance bands. Then, the spectra were randomly assigned to groups in the same number as the isolates of each species, and the spectra of each group were averaged and plotted. The result is shown in Fig. 6, for two representative species, *C. lusitaniae* (low inter-isolates variance) and *C. guilliermondii* (higher inter-isolate variance). For all species, the variance is significantly reduced (right part of Fig. 6) when the spectra are grouped randomly instead of isolates-based. This suggests that different isolates have different molecular fingerprints, and spectral variability is not due to molecular composition heterogeneity.

This inter-isolates variability could negatively impact the ability to classify at the species level. To assess class (inter-species) separability, we applied the Linear Discriminant Analysis (LDA) to the dataset. This supervised algorithm reduces the feature space to a lower number of dimensions, looking for a linear combination of the features that maximize the inter-class separation (of the class centroids or means) and minimizes the intra-class variance [50]. Fig. 7 shows the LDA projection of the entire dataset to a 2-dimensional space (only the centroids of each class are shown). This figure gives an idea of the separability between the classes. Four major clusters seem to appear: *C. inconspicua*; *C. krusei* and *C. glabrata*; *C. metapsilosis* and *C. parapsilosis*; and a fourth cluster with the remaining species. These distances are expected to be translated to the performance of any classification algorithms, as shown in Subsection 3.5.

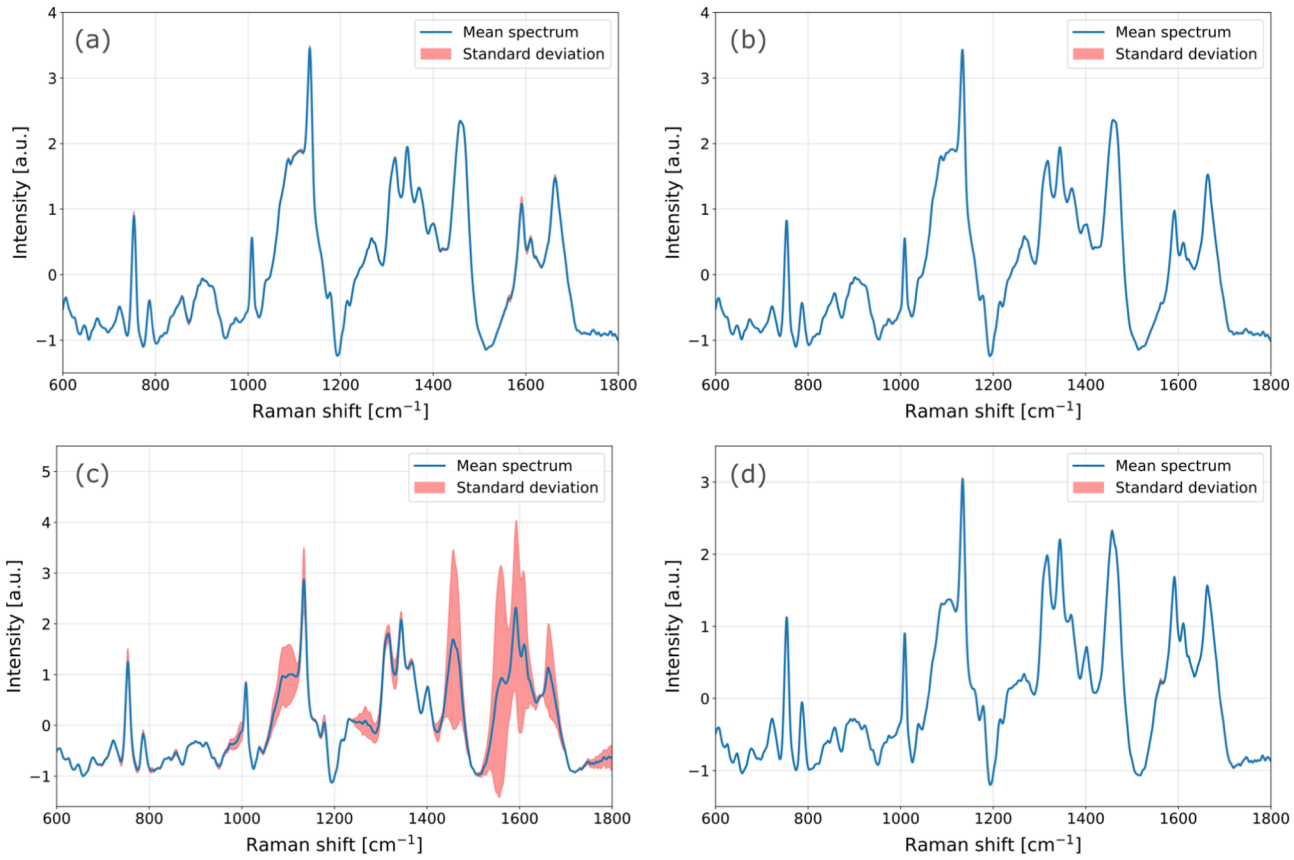


Fig. 6. *C. lusitaniae*: averaged spectrum of independent isolates (a) and averaged spectrum of randomly-split data (b). *C. guilliermondii*: averaged spectrum of independent isolates (c) and averaged spectrum of randomly-split data (d).

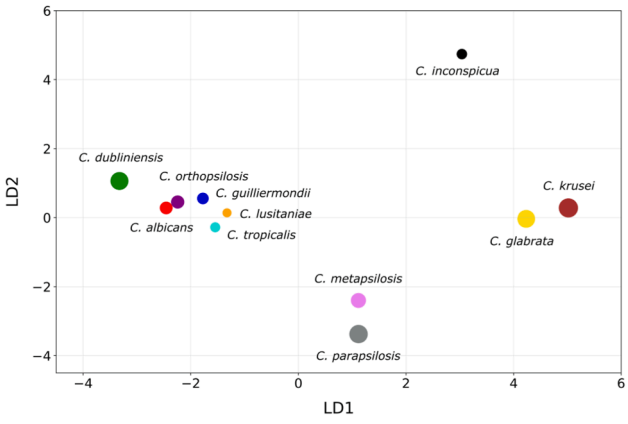


Fig. 7. Centroids of the *Candida* species clusters on a 2D projection of the dataset using LDA.

3.4. Outlier spectra

The dataset was analyzed for outlier spectra that could have originated from any experimental issue. The Isolation Forest algorithm was applied to the entire dataset to calculate anomaly scores for each observation, which are a measure of how difficult it is to classify each spectrum using a binary tree [51]. It was found that all spectra of a single isolate of *C. guilliermondii* (Cgui-5) have the highest scores. Its outlier nature was confirmed by means of the LDA algorithm previously discussed in Subsection 3.3, but trained at the isolate level, as shown in Fig. 8(a). It is clearly apparent that this single isolate is projected far away from the rest of the isolates, thus indicating distinct features that make it easily (but wrongly) distinguishable from the others.

In Fig. 8(b), this isolate's mean spectrum (from 9 acquired spectra) is shown, compared with several other randomly chosen isolates from the dataset. There are clear differences between this isolate spectrum and the others, in particular, Raman features at 1550 cm^{-1} (*exo*-polysaccharides) and 1592 cm^{-1} (cytochrome) have higher intensity, while others, such as the DNA fingerprint at 787 cm^{-1} or 1612 cm^{-1}, are not present at all. Interestingly, the features with increased intensity in this isolate correspond with extracellular substances that are part of the *Candida* spp.' secretome [52]. This suggests that extracellular substances could greatly affect Raman spectra of *Candida* spp., even if great care is taken to focus the laser spot on the cellular bodies.

As those clear outlier spectra could be detected a priori in new measurements, we have removed this isolate from the dataset. Additionally, we have still applied the Isolation Forest outlier removal algorithm with different numbers of spectra with the highest anomaly score removed to verify the presence of not-so-evident outliers and their impact on the classification performance. The impact of outlier removal is discussed in Subsection 3.5 and depicted in Fig. 9.

3.5. Classifiers' performance

In this work, we have paid special attention to reducing and evaluating the degree of overfitting in the trained models, as there is an increasing concern in the spectroscopists community about the publication of heavily overfitted models with impressive metrics that, actually, have poor generalization capabilities and therefore are not valid for new measurements, let alone for different instrumentation or experimental conditions [48]. We have applied several techniques in the design and training of the models to reduce overfitting, such as cross-validation, feature selection and data augmentation. For the 1-D CNN

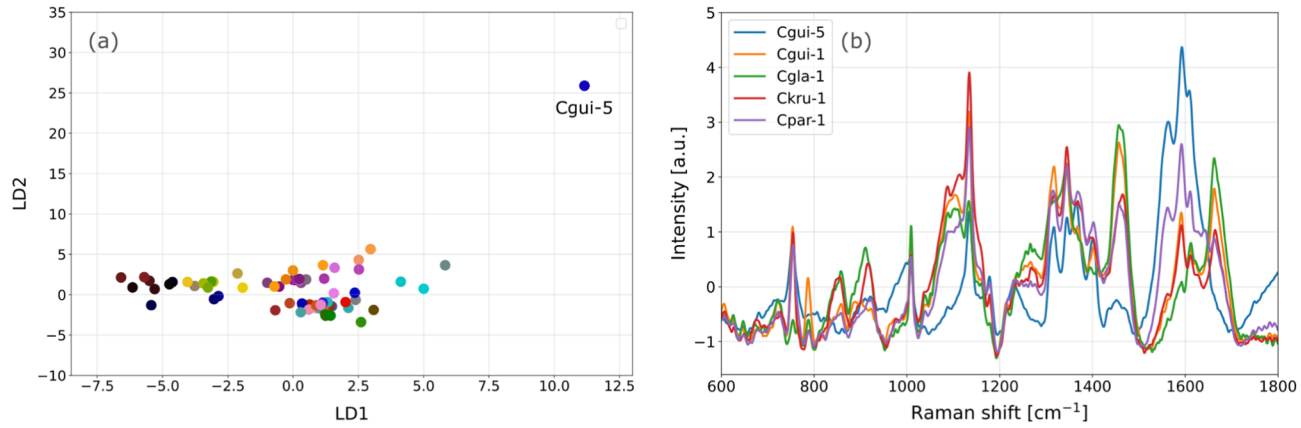


Fig. 8. (a) LDA bi-dimensional projection at the isolate level (only centroids of isolate clusters are shown). The outlier isolate of *C. guilliermondii* (Cgui-5) is located at the top-right corner. (b) average spectrum of Cgui-5, compared with several other spectra randomly chosen.

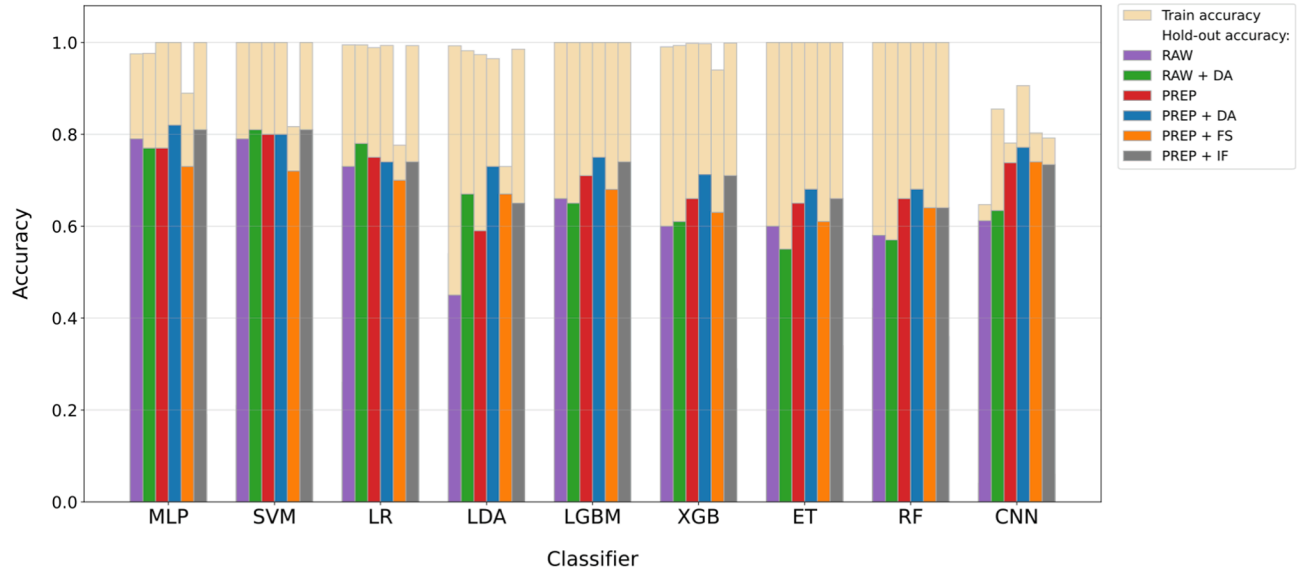


Fig. 9. Comparison between the classification models (MLP, SVM, LR, LDA, LGBM, XGB, ET, RF and CCN) trained with several spectral treatments (RAW, RAW + DA, PREP, PREP + DA, PREP + FS and PREP + IF). The train accuracies are shown with yellow bars, and the accuracies of the hold-out (unseen) subset are shown with coloured bars.

deep learning model, its architecture includes dropout layers and a constrained model size. In addition, the training was performed with regularization and early stopping, as those techniques also help to reduce overfitting [53].

The classification algorithms (MLP, SVM, LR, LDA, LGBM, XGB, ET, RF and CCN) were trained using 6 different spectral treatments: spectra only standardized (RAW), only standardized with data augmentation (RAW + DA), preprocessed with normalization, baseline removal and standardization as discussed above (PREP), preprocessed with data augmentation (PREP + DA), preprocessed with feature selection (PREP + FS) and preprocessed with outlier removal (PREP + IF). For all cases, cross-validation was used with 10 splits from the set reserved for training. All models except 1-D CNN were trained with Pycaret and its automatic hyperparameter optimization. In the case of 1-D CNN, Keras and optimization of hyperparameters with the Optuna library were used.

As an evaluation method of the degree of overfitting, the accuracies of the unseen subset were compared with those obtained with the train subset. If the discrepancies between both values were high, the model would not be able to properly generalize to new unknown data.

Fig. 9 compares the performance obtained by the models when trained with the different preprocessing schemes. The same hyperparameters were set for all the models in order to evaluate the effect of each spectral treatment added to the original dataset. We have selected the hyperparameters that, in general, allowed us to obtain the highest accuracy of the unseen data while maintaining a low discrepancy with the train accuracy.

From Fig. 9, it can be seen that all models based on decision trees (LGBM, XGB, ET, RF) have high overfitting (train accuracy = 1.0 in most cases), so they have not been considered suitable for this study. The rest of the models trained with Pycaret (MLP, SVM, LR and LDA) also showed a poor ability to adapt to new data properly. For these models, better results are achieved with the preprocessed with feature selection (PREP + FS) dataset. However, the best model, achieving low overfitting for all datasets while maintaining the accuracy of the hold-out subset, is the 1-D CNN. From the figure, it can be seen that applying the data augmentation produces a slight increase in the hold-out accuracies and very high training accuracy values, so the degree of overfitting has not improved with this spectral treatment.

On the other hand, hardly any changes were observed when applying feature selection or outliers removal with the isolation algorithm. Interestingly, the accuracy of the two 'RAW' subsets is notably lower,

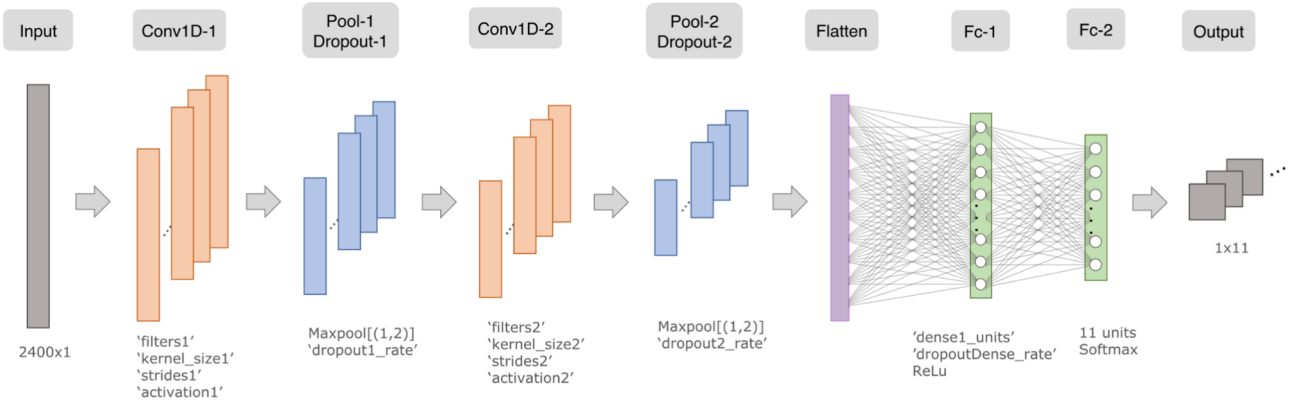


Fig. 10. 1-D CNN architecture selected for *Candida* species classification.

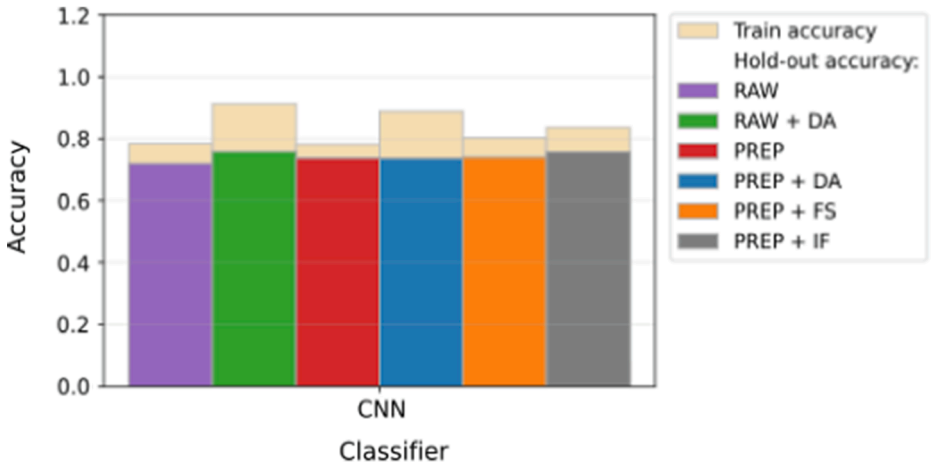


Fig. 11. Comparison between the accuracies obtained by training different 1-D CNNs with hyperparameter optimization (HPO) for each dataset.

Table 3
List of 1-D CNN hyperparameters tuned using the Optuna library.

Layer	Hyperparameter	Search range	Selected value
Conv1D-1	'filters1'	[8,16,32]	16
Conv1D-1	'kernel_size1'	[9,17,31]	9
Conv1D-1	'strides1'	[1,8,16]	8
Conv1D-1	'activation1'	['relu', 'linear']	'relu'
Dropout-1	'dropout1_rate'	(0.1, 0.5)	0.21
Conv1D-2	'filters2'	[8,16]	16
Conv1D-2	'kernel_size2'	[5,7,9]	7
Conv1D-2	'strides2'	[1–2,4]	1
Conv1D-2	'activation2'	['relu', 'linear']	'relu'
Conv1D-2	'dropout2_rate'	(0.1, 0.5)	0.11
Fc-1	'dense1_units'	[32,64,128]	128
Fc-1	'dropoutDense_rate'	(0.1, 0.5)	0.41
	learning rate	(1e-5, 1e-2)	2.29e-4
	number of epochs	[5,10,15,25,30]	15

but this is because the same hyperparameters were used for the comparison. After a hyperparameter optimization for each particular dataset preprocessing scheme, we found a similar achievable accuracy of 0.75–0.8, irrespective of the applied transformation to the spectral dataset. This is in agreement with other works that highlight the ability of deep neural networks, and in particular, Convolutional Neuronal Networks, to obtain good performance with minimal preprocessing [54]. Fig. 10 depicts the 1-D CNN architecture selected for training with the different datasets. The resulting accuracies after hyperparameter optimization of the 1-D CNN model for the different preprocessing schemes are shown in Fig. 11, showing the minimal impact of the preprocessing scheme on the achievable overall accuracy.

Table 3 shows the hyperparameters considered in the search for the optimal architecture and the selected values, from which the accuracies and degree of overfitting of Fig. 9 were obtained.

According to Table 3, the optimal 1-D CNN architecture has a convolutional 1-D layer with 16 filters, kernel size 9, stride size 8 and ReLu activation, followed by a dropout layer with a rate of 0.21, and a max-pooling layer with pool size 2 and stride size of 2. It is followed by a second convolutional 1-D layer with 16 filters, kernel size 7, stride size 1 and ReLu activation followed by a dropout layer with a rate of 0.11 and a max-pooling layer with pool size 2 and stride size of 2. The next layer is a flatten layer that converts the outputs of the convolutional filters to 1-D data for the fully connected layers. The first fully connected layer has 128 units with a dropout rate of 0.41 and the second fully connected layer has 11 units with softmax activations. In addition, the trainings were performed by setting the learning rate to 2.3e-4 and the number of epochs to 15.

To assess the contribution of each class to the overall accuracy, Fig. 12 shows the confusion matrix obtained for the 1-D CNN model trained with the preprocessed dataset (PREP), which could be considered the best performing model in our work due to the high accuracy and the best generalization capability to unknown data.

As expected, the model classifies some *Candida* species better than others. There are species that classify worse and it is not due to confusion with another specific one. We can affirm that all analyzed *Candida* species present spectral differences that allow their classification with high accuracy, while for specific species it is more difficult to find these differences. This could be explained by the high similarities among some

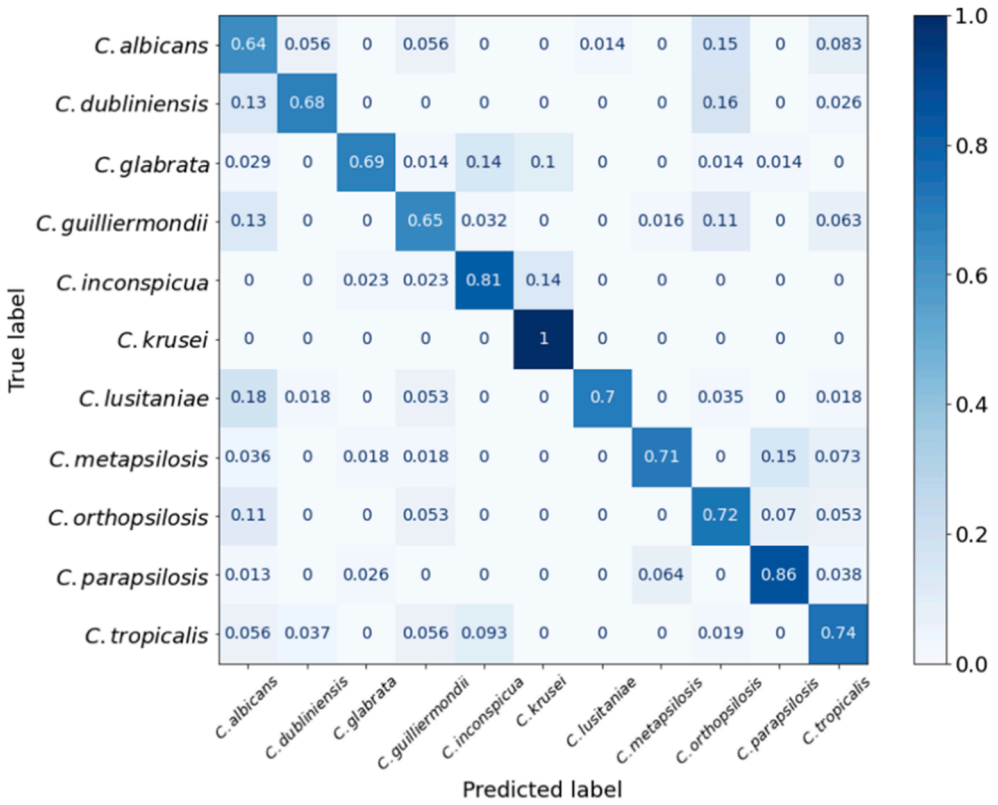


Fig. 12. Confusion matrix for unseen data obtained from 1-D CNN training with the preprocessed *Candida* species dataset.

Table 4
Performance metrics for 1-D CNN trained with the preprocessed *Candida* species dataset.

	Accuracy	Recall	Precision	Specificity	F1-score	AUC
C. albicans	0.90	0.64	0.55	0.93	0.59	0.79
C. dubliniensis	0.97	0.68	0.79	0.99	0.73	0.84
C. glabrata	0.96	0.69	0.92	0.99	0.79	0.84
C. guilliermondii	0.94	0.65	0.72	0.97	0.68	0.81
C. inconspicua	0.96	0.81	0.67	0.97	0.74	0.89
C. krusei	0.98	1.00	0.75	0.98	0.86	0.99
C. lusitaniae	0.97	0.70	0.98	1.00	0.82	0.85
C. metapsilosis	0.96	0.71	0.87	0.99	0.78	0.85
C. orthopsilosis	0.93	0.72	0.59	0.95	0.65	0.84
C. parapsilosis	0.96	0.86	0.84	0.98	0.85	0.92
C. tropicalis	0.94	0.74	0.65	0.96	0.69	0.85
Macro average	**0.95**	**0.75**	**0.76**	**0.97**	**0.83**	**0.86**
Overall accuracy 0.74						

species, i.e., *C. parapsilosis*, *C. orthopsilosis* and *C. metapsilosis*, which are phenotypically indistinguishable [55]. A detailed study of the per-species classification performance can be seen in Table 4, in which different metrics are calculated for each *Candida* species, for the best 1-D CNN model.

Finally, to further assess the degree of overfitting and therefore the generalization capability of the models, Y-randomization, also known as permutation testing, was applied. This technique has been suggested as a necessary test for any regression or classification model dealing with spectroscopic data, not only to detect overfitting but also to assure that the dataset has real discriminative information to differentiate between classes [56]. It is based on re-training the models with a random permutation of the class labels. If good performance is obtained, i.e., the re-trained model can find a correlation between the spectral information and the (now) randomized labels, it is an indication that it is overfitted to spurious information unrelated to actual inter-class differences. By contrast, if low classification performance (close to chance) is obtained by the re-trained model, even for the new, randomized train dataset, we

Table 5
Results of the Y-randomization test for some representative models and different preprocessing of the dataset. Those models showing a high train accuracy with the randomized labels (*) likely suffer a high degree of overfitting and lack of generalization. All of the models, however, show a test accuracy close to chance, thus indicating that the dataset includes relevant discriminative information.

Model	Data	Train / Unseen acc. (original data)	Train / Unseen acc. (randomized labels)
RF*	Preprocessed + FS	1.00 / 0.64	**1.00** / 0.12
LDA	Preprocessed + FS	0.73 / 0.67	0.17 / 0.10
SVM	Preprocessed + FS	0.82 / 0.72	0.18 / 0.11
RF*	Preprocessed + DA	1.00 / 0.66	**1.00** / 0.12
MLP	Preprocessed + DA	1.00 / 0.82	0.10 / 0.12
SVM	Preprocessed + DA	1.00 / 0.80	0.08 / 0.09
1-D CNN	Preprocessed + DA	0.91 / 0.77	0.10 / 0.06

can assume that the model has found and is exploiting a significant class structure based on real discriminant information [57]. We have applied this test to some of the previous models, considering two different transformations that result in high accuracy and should help to reduce overfitting: preprocessed with feature selection (PREP + FS) and preprocessed with data augmentation (PREP + DA). For each model, we have obtained its overall accuracy for the train and for the hold-out (unseen) dataset, with and without label randomization. The results are shown in Table 5. It can be seen that some of the models (in particular, tree-based such as Random Forest) show a high accuracy for the (randomized) training dataset, thus suggesting a high probability of overfitting. On the contrary, SVM, MLP and 1-D CNN obtain very low accuracy, thus suggesting a lower degree of overfitting. Nevertheless, all of them show a low accuracy (close to random, 9 %, for 11 classes) for the randomized unseen dataset, thus suggesting that the spectroscopic data has relevant information that ties the spectral data to the actual classes.

4. Conclusions

Raman spectra from 11 species of the genus *Candida* (*C. albicans, C. dubliniensis, C. glabrata, C. guilliermondii, C. inconspicua, C. krusei, C. lusitaniae, C. metapsilosis, C. orthopsilosis, C. parapsilosis, C. tropicalis*) were obtained from pure cultures under controlled and repetitive experimental conditions.

The dataset includes between 220 and 350 spectra from each one of the 11 species (a total of 3126 Raman spectra), identified at the isolate level (67 different isolates). The quality of the spectra, the intra-class and inter-class variability, and the presence of outliers were assessed. Twenty-four Raman features were clearly identified in the spectra, that were assigned to specific molecular bands, based on previous literature. We have shown that some species have a low variability of measured spectra while others exhibit a high variance in several Raman features. We have found that the presence of secretions in some of the samples could have an impact on the classification performance of any Raman-based approach and need further research. A study of the importance of the features on different models has also been carried out and we have determined that there is a Raman band assigned to genetic material with much higher importance than other features.

A total of 9 representative machine learning and deep learning models were trained and evaluated in this work: Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Random Forest (RF), Extra Trees Classifier (ET), XGBoost (XGB), Light Gradient Boosting Machine (LGBM), Multilayer Perceptron (MLP) and a one-dimensional Convolutional Neural Network (1-D CNN). Each one was trained with different variations of the dataset: raw spectra, preprocessed (energy normalization, smoothing, baseline correction and standardization), with and without feature selection (selecting the 43 most significant features), with and without data augmentation (linear combination of spectra and noise addition), with and without outliers removal (5 % of spectra with higher outlier factor). A hyperparameter optimization to find the best parameter combination or architecture was performed on all models. Trained models achieve overall accuracy values between 0.7 and 0.8, a very good result considering the number of classes, but with different degrees of overfitting. To further assess the overfitting of each trained model, and thus their generalization capabilities, a Y-randomization test was performed. We consider that the best model to classify the *Candida* dataset at the species level automatically is the 1-D CNN, achieving slightly above 0.8 overall accuracy value but with a low degree of overfitting, and a robust performance irrespective of the preprocessing applied to the spectra. A per-species analysis of the classification performance shows values of specific accuracies ranging from 0.90 to 0.98, in accordance with their phenotypic similarities.

We, therefore, believe that it is feasible to automatically classify 11 pathogenic species of *Candida* with high accuracy based on their Raman spectra. Future works are focused on classification at the isolate level, for which genotype variations could not be directly related to phenotype differences, the improvement of the classification performance, and the inclusion of other species, in particular, *C. auris*, due to its worldwide expansion as a multi-drug resistant yeast. For the task of automatic pathogen identification, Raman spectroscopy has shown to be a powerful tool that could be transferable to daily clinical routine due to its simplicity of use, easy automation, reproducibility, rapid acquisition, low cost and minimum sample preparation, especially when combined with machine learning and deep learning algorithms.

CRediT authorship contribution statement

María Gabriela Fernández-Manteca: Writing – original draft, Investigation, Methodology, Software. **Alain A. Ocampo-Sosa:** Writing – original draft, Investigation, Methodology, Conceptualization, Validation, Resources, Supervision. **Carlos Ruiz de Alegría-Puig:** Methodology, Conceptualization. **María Pía Roiz:** Methodology, Conceptualization. **Jorge Rodríguez-Grande:** Formal analysis, Software. **Fidel Madrazo:** Resources, Investigation. **Jorge Calvo:** Conceptualization, Supervision. **Luis Rodríguez-Cobo:** Writing – review & editing. **José Miguel López-Higuera:** Project administration, Funding acquisition. **María Carmen Fariñas:** Project administration, Supervision. **Adolfo Cobo:** Writing – original draft, Software, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the R + D projects INNVAL19/17 (funded by Instituto de Investigación Valdecilla-IDIVAL), PID2019-107270RB-C21 (funded by MCIN/ AEI /10.13039/501100011033) and by Plan Nacional de I + D + and Instituto de Salud Carlos III (ISCIII), Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Ciencia, Innovación y Universidades, Spanish Network for Research in Infectious Diseases (REIPI RD16/0016/0007), CIBERINFEC (CB21/13/00068), CIBER-BBN (BBNGC1601), cofinanced by European Development Regional Fund “A way to achieve Europe”. A. A. O.-S was financially supported by the Miguel Servet II program (ISCIII-CPII17-00011).

References

[1] S. Bhattacharya, S. Sae-Tia, B.C. Fries, Candidiasis and Mechanisms of Antifungal Resistance, Antibiotics 9 (2020) 1–19, https://doi.org/10.3390/ANTIBIOTICS9060312.
[2] N. Yapar, Epidemiology and risk factors for invasive candidiasis, Ther. Clin. Risk Manag. 10 (2014) 95, https://doi.org/10.2147/TCRM.S40160.
[3] T.P. McCarty, C.M. White, P.G. Pappas, Candidemia and Invasive Candidiasis, Infect. Dis. Clin. North Am. 35 (2021) 389–413, https://doi.org/10.1016/J.IDC.2021.03.007.
[4] C. Cervera, Candidemia and invasive candidiasis in the adult: clinical forms and treatment, Enferm. Infecc. Microbiol. Clin. 30 (2012) 483–491, https://doi.org/10.1016/J.EIMC.2012.02.003.
[5] M.A. Pfaller, M. Castanheira, Nosocomial Candidiasis: Antifungal Stewardship and the Importance of Rapid Diagnosis, Med. Mycol. 54 (2016) 1–22, https://doi.org/10.1093/MMY/MYV076.
[6] R. Ben-Ami, D.P. Kontoyiannis, Resistance to Antifungal Drugs, Infect. Dis. Clin. North Am. 35 (2021) 279–311, https://doi.org/10.1016/J.IDC.2021.03.003.
[7] B. Aslam, W. Wang, M.I. Arshad, M. Khurshid, S. Muzammil, M.H. Rasool, M. A. Nisar, R.F. Alvi, M.A. Aslam, M.U. Qamar, M.K.F. Salamat, Z. Baloch, Antibiotic resistance: a rundown of a global crisis, Infect. Drug Resist. 11 (2018) 1645–1658, https://doi.org/10.2147/IDR.S173867.

[8] N. Martins, I.C.F.R. Ferreira, L. Barros, S. Silva, M. Henriques, Candidiasis: Predisposing Factors, Prevention, Diagnosis and Alternative Treatment, Mycopathologia 177 (2014) 223–240, https://doi.org/10.1007/S11046-014-9749-1.

[9] O. Epelbaum, R. Chasan, Candidemia in the Intensive Care Unit, Clin. Chest Med. 38 (2017) 493–509, https://doi.org/10.1016/J.CCM.2017.04.010.

[10] P.L. White, A.E. Archer, R.A. Barnes, Comparison of non-culture-based methods for detection of systemic fungal infections, with an emphasis on invasive Candida infections, J. Clin. Microbiol. 43 (2005) 2181–2187, https://doi.org/10.1128/JCM.43.5.2181-2187.2005.

[11] R.I. Amann, W. Ludwig, K.H. Schleifer, Phylogenetic identification and in situ detection of individual microbial cells without cultivation, Microbiol. Rev. 59 (1995) 143–169, https://doi.org/10.1128/MMBR.59.1.143-169.1995.

[12] K. Kivirand, T. Rinken, Introductory Chapter: Why Do We Need Rapid Detection of Pathogens?, Biosensing Technol, Detect. Pathog. - A Prospect. W. Rapid Anal. (2018), https://doi.org/10.5772/INTECHOPEN.74670.

[13] N. Singhal, M. Kumar, P. K. Kanaujia, J. S. Virdi, Maldi-tof mass spectrometry: an emerging technology for microbial identification and diagnosis, Front. Microbiol. 6. doi:10.3389/fmicb.2015.00791.

[14] V. Ruelle, B. El Moualij, W. Zorzi, P. Ledent, E. De Pauw, Rapid identification of environmental bacterial strains by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, Rapid Commun. Mass Spectrom. 18 (2004) 2013–2019, https://doi.org/10.1002/RCM.1584.

[15] H. Chen, A. Das, L. Bi, N. Choi, J.-I. Moon, Y. Wu, S. Park, J. Choo, Recent advances in surface-enhanced raman scattering-based microdevices for point-of-care diagnosis of viruses and bacteria, Nanoscale 12 (2020) 21560–21570, https://doi.org/10.1039/D0NR06340A.

[16] J. Guicheteau, S. Christesen, D. Emge, A. Tripathi, Bacterial mixture identification using Raman and surface enhanced raman chemical imaging, J. Raman Spectrosc. 41 (2010) 1632–1637, https://doi.org/10.1002/jrs.2601.

[17] [Wang2020] D. Wang, P. He, Z. Wang, G. Li, N. Majed, A. Z. Gu, Advances in single cell Raman spectroscopy technologies for biological and environmental applications, Curr. Opin. Biotechnol. 64 (2020) 218–229, analytical Biotechnology. doi:10.1016/j.copbio.2020.06.011.

[18] H. Sato, Y. Maeda, M. Ishigaki, B. Andriana, Biomedical Applications of Raman Spectroscopy (2000) 1–12, https://doi.org/10.1002/9780470027318.a9281.

[19] G. Auner, S. Koya, C. Huang, B. Broadbent, M. Trexler, Z. Auner, A. Elias, K. Mehne, M. Brusatori, Applications of raman spectroscopy in cancer diagnosis, Cancer Metastasis Rev. 37 (2018), https://doi.org/10.1007/s10555-018-9770-9.

[20] A.Y.F. You, M.S. Bergholt, J.-P. St-Pierre, W. Kit-Anan, I.J. Pence, A.H. Chester, M. H. Yacoub, S. Bertazzo, M.M. Stevens, Raman spectroscopy imaging reveals interplay between atherosclerosis and medial calcification in the human aorta, Science, Advances 3 (12) (2017), https://doi.org/10.1126/sciadv.1701156.

[21] K. Eberhardt, C. Stiebing, C. Matthäus, M. Schmitt, J. Popp, Advantages and limitations of raman spectroscopy for molecular diagnostics: an update, Expert Rev. Mol. Diagn. 1 (6) (2015) 773–787, https://doi.org/10.1586/14737159.2015.1036744.

[22] L. Pan, P. Zhang, C. Daengngam, S. Peng, M. Chongcheawchamnan, A review of artificial intelligence methods combined with raman spectroscopy to identify the composition of substances, J. Raman Spectrosc. 53 (1) (2022) 6–19, https://doi.org/10.1002/jrs.6225.

[23] H. Byrne, P. Knief, M. Keating, F. Bonnier, Spectral pre and post processing for infrared and raman spectroscopy of biological tissues and cells, Chem. Soc. Rev. 45 (2016), https://doi.org/10.1039/c5cs00440c.

[24] [Chouthai2015] N. Chouthai, A. Shah, H. Salimnia, O. Palyvoda, S. Devpura, M. Klein, and B. Asmar, Use of raman spectroscopy to decrease time for identifying the species of Candida growth in cultures, Avicenna journal of medical biotechnology, vol. 7, pp. 45–8, 04 2015.

[25] O. Samek, K. Rebrošová, S. Bernatova, J. Jezek, V. Krzyzanek, M. Siler, P. Zemanek, F. růžička, V. Holá, and M. Mahelová,, Candida parapsilosis biofilm identification by raman spectroscopy, Int. J. Mol. Sci. 15 (12 2014.) 23924–23935, https://doi.org/10.3390/ijms151223924.

[26] [Pezzotti2021] G. Pezzotti, M. Kobara, T. Asai, T. Nakaya, N. Miyamoto, T. Adachi, T. Yamamoto, N. Kana- mura, E. Marin, W. Zhu, I. Nishimura, O. Mazda, T. Nakata, and K. Makimura, Raman imaging of pathogenic Candida auris: Visualization of structural characteristics and machine-learning identification, Frontiers in Microbiology, vol. 12, 11 2021.

[27] E. Witkowska, et al., Detection and identification of human fungal pathogens using surface-enhanced Raman spectroscopy and principal component analysis, Anal. Methods 8 (48) (2016) 8427–8434, https://doi.org/10.1039/C6AY02957D.

[28] P.HC. Eilers, H.FM. Boelens, Baseline correction with asymmetric least squares smoothing, Leiden University Medical Centre Report 1 (1) (2005) 5.

[29] T.J. White, et al., Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, PCR protocols: a guide to methods and applications 18 (1) (1990) 315–322.

[30] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation-based anomaly detection, ACM Transactions on Knowledge Discovery from Data (TKDD) 6 (1) (2012) 1–39.

[31] J. Liu, et al., Deep convolutional neural networks for Raman spectrum recognition: a unified solution, Analyst 142 (21) (2017) 4067–4074, https://doi.org/10.1039/C7AN01371J.

[32] Z. Li, et al., A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Trans. Neural Networks Learn. Syst. (2021).

[33] [Gristanj2022] Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on tabular data?." arXiv preprint arXiv:2207.08815 (2022).

[34] F. Pedregosa, et al., Scikit-learn: Machine learning in Python, the Journal of machine Learning research 12 (2011) 2825–2830.

[35] [Tensorflow2022] https://github.com/tensorflow (accessed: 16 September 2022), doi: 10.5281/zenodo.4724125.

[36] M. Ali, PyCaret: An open source, low-code machine learning library in Python, PyCaret version 2 (2020).

[37] [Optuna2019] Akiba, Takuya, et al. "Optuna: A next-generation hyperparameter optimization framework." Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019.

[38] T. Lemma, et al., Identifying yeasts using surface enhanced Raman spectroscopy, Spectrochim. Acta A Mol. Biomol. Spectrosc. 218 (2019) 299–307, https://doi.org/10.1016/j.saa.2019.04.010.

[39] A. Mushtaq, et al., "Surface-Enhanced Raman Spectroscopy (SERS) for Monitoring Colistin-Resistant and Susceptible E. Coli Strains". Spectrochimica Acta. Part A, Molecular and Biomolecular, Spectroscopy 278 (2022), 121315, https://doi.org/10.1016/j.saa.2022.121315.

[40] A. Zdaniauskienė, et al., Shell-isolated nanoparticle-enhanced Raman spectroscopy for characterization of living yeast cells, Spectrochim. Acta A Mol. Biomol. Spectrosc. 240 (2020), 118560, https://doi.org/10.1016/j.saa.2020.118560.

[41] S. Pebotuwa, et al., "Influence of the Sample Preparation Method in Discriminating Candida spp. Using ATR-FTIR Spectroscopy". Molecules vol. 25 (2020), https://doi.org/10.3390/molecules25071551.

[42] A. Silge, et al., The application of UV resonance Raman spectroscopy for the differentiation of clinically relevant Candida species, Anal. Bioanal. Chem. 410 (23) (2018) 5839–5847, https://doi.org/10.1007/s00216-018-1196-2.

[43] M. Kashif, et al., Surface-enhanced Raman spectroscopy for identification of food processing bacteria, Spectrochim. Acta A Mol. Biomol. Spectrosc. 261 (2021), 119989.

[44] de Siqueira e Oliveira, Fernanda SantAna,, et al., Discrimination of selected species of pathogenic bacteria using near-infrared Raman spectroscopy and principal components analysis, J. Biomed. Opt. 17 (10) (2012), 107004, https://doi.org/10.1117/1.JBO.17.10.107004.

[45] G. Pezzotti, et al., Raman Spectroscopy of Oral Candida Species: Molecular-Scale Analyses, Chemometrics, and Barcode Identification, Int. J. Mol. Sci. 23 (10) (2022) 5359, https://doi.org/10.3390/ijms23105359.

[46] H. Noothalapati, et al., Label-free Chemical Imaging of Fungal Spore Walls by Raman Microscopy and Multivariate Curve Resolution Analysis, Sci. Rep. 6 (2016) 27789, https://doi.org/10.1038/srep27789.

[47] S. Bashir, et al., Surface-enhanced Raman spectroscopy for the identification of tigecycline-resistant E. coli strains, Spectrochim. Acta A Mol. Biomol. Spectrosc. 258 (2021), 119831.

[48] P. Pradhan, et al., Deep learning a boon for biophotonics? J. Biophotonics 13 (6) (2020) e201960186.

[49] M. Cavalheiro, M.C. Teixeira, "Candida biofilms: threats, challenges, and promising strategies, Front Med (Lausanne) 5 (2018) 28.".

[50] J. Gareth, et al., An introduction to statistical learning: with applications in R, Springer, 2013.

[51] F.T. Liu, K.M. Ting, Z.-H. Zhou, On detecting clustered anomalies using SCiForest. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, 2010.

[52] M. Rasheed, N. Kumar, R. Kaur, Global secretome characterization of the pathogenic yeast Candida glabrata, J. Proteome Res. 19 (1) (2019) 49–63.

[53] A.D. Gavrilov, et al., Preventing model overfitting and underfitting in convolutional neural networks, International Journal of Software Science and Computational Intelligence (IJSSCI) 10 (4) (2018) 19–28.

[54] D. Ma, et al., Classifying breast cancer tissue by Raman spectroscopy with one-dimensional convolutional neural network, Spectrochim. Acta A Mol. Biomol. Spectrosc. 256 (2021), 119732.

[55] [Tavanti2005] A. Tavanti, A. Davidson, N. Gow, M. Maiden, and F. Odds, "Candida Orthopsilosis and Candida Metapsilosis spp. Nov. to replace Candida Parapsilosis groups II and III," Journal of clinical microbiology, vol. 43, pp. 284–92, 01 2005.

[56] E.l. Haddad, L.C. Josette, B. Bousquet, Good practices in LIBS analysis: Review and advices, Spectrochim. Acta B At. Spectrosc. 101 (2014) 171–182.

[57] M. Ojala, G.C. Garriga, Permutation tests for studying classifier performance, J. Mach. Learn. Res. 11 (2010) 6.