TYPE Original Research
PUBLISHED 10 July 2024
DOI 10.3389/fchem.2024.1412288

Check for updates

OPEN ACCESS

EDITED BY
Hassan Rasouli,
Kermanshah University of Medical Sciences, Iran

REVIEWED BY
Chunfen Jin,
Honeywell UOP, United States
Mariana Gallo,
University of Parma, Italy

*CORRESPONDENCE
Kássio M. G. Lima,
✉ kassio.lima@ufrn.br
Rafael W. Bastos,
✉ rafawbastos@gmail.com

‡These authors jointly supervised the work

†These authors have contributed equally to this work

RECEIVED 04 April 2024
ACCEPTED 10 June 2024
PUBLISHED 10 July 2024

CITATION
Nascimento ALF, de Medeiros AGJ, Neves ACO, de Macedo ABN, Rossato L, Assis Santos D, dos Santos ALS, Lima KMG and Bastos RW (2024), Near-infrared spectroscopy and multivariate analysis as effective, fast, and cost-effective methods to discriminate *Candida auris* from *Candida haemulonii*.
Front. Chem. 12:1412288.
doi: 10.3389/fchem.2024.1412288

Near-infrared spectroscopy and multivariate analysis as effective, fast, and cost-effective methods to discriminate *Candida auris* from *Candida haemulonii*

Ayrton L. F. Nascimento[1†], Anthony G. J. de Medeiros[2†], Ana C. O. Neves[1], Ana B. N. de Macedo[2], Luana Rossato[3], Daniel Assis Santos[4,5], André L. S. dos Santos[6], Kássio M. G. Lima[1*‡] and Rafael W. Bastos[2,5*‡]

[1]Laboratório de Química Biológica e Quimiometria, Instituto de Química, Universidade Federal do Rio Grande do Norte, Natal, Brazil, [2]Laboratório de Uso Comum, Centro de Biociências, Universidade Federal do Rio Grande do Norte, Natal, Brazil, [3]Laboratório de Pesquisa em Ciências da Saúde, Universidade Federal da Grande Dourados, Dourados, Brazil, [4]Laboratório de Micologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [5]National Institute of Science and Technology in Human Pathogenic Fungi, Ribeirão Preto, Brazil, [6]Instituto de Microbiologia Paulo de Góes, Universidade Federal do Rio de Janeiro, Rio de Janeiro-RJ, Brazil

Candida auris and *Candida haemulonii* are two emerging opportunistic pathogens that have caused an increase in clinical cases in the recent years worldwide. The differentiation of some *Candida* species is highly laborious, difficult, costly, and time-consuming depending on the similarity between the species. Thus, this study aimed to develop a new, faster, and less expensive methodology for differentiating between *C. auris* and *C. haemulonii* based on near-infrared (NIR) spectroscopy and multivariate analysis. *C. auris* CBS10913 and *C. haemulonii* CH02 were separated in 15 plates per species, and three isolated colonies of each plate were selected for Fourier transform near-infrared (FT-NIR) analysis, totaling 90 spectra. Subsequently, principal component analysis (PCA) and variable selection algorithms, including the successive projections algorithm (SPA) and genetic algorithm (GA) coupled with linear discriminant analysis (LDA), were employed to discern distinctive patterns among the samples. The use of PCA, SPA, and GA algorithms associated with LDA achieved 100% sensitivity and specificity for the discriminations. The SPA-LDA and GA-LDA algorithms were essential in selecting the variables (infrared wavelengths) of most importance for the models, which could be attributed to binding of cell wall structures such as polysaccharides, peptides, proteins, or molecules resulting from yeasts' metabolism. These results show the high potential of combined FT-NIR and

Abbreviations: NIR, near-infrared ; FT-NIR, Fourier transform near-infrared ; PCA, principal component analysis; SPA, successive projection algorithm; GA, genetic algorithm; LDA, linear discriminant analysis; S/N, signal-to-noise ratio; HCA, hierarchical cluster analysis; SDA, Sabouraud dextrose agar; BHI, brain heart infusion; EMSC, extended multiplicative scatter correction; SENS, sensitivity; SPEC, specificity; G-Score, geometric score; AUC, area under the curve; TP, true positive; TN, true negative; FP, false positive; FN, false negative; TPR, true-positive rate; FPR, false-positive rate; PCs, principal components.

multivariate analysis techniques for the classification of *Candida*-like fungi, which can contribute to faster and more effective diagnosis and treatment of patients affected by these microorganisms.

KEYWORDS

***Candida auris*, *Candida haemulonii*, near-infrared spectroscopy, principal component analysis-linear discriminant analysis, successive projections algorithm-linear discriminant analysis, genetic algorithm-linear discriminant analysis, multivariate analysis**

Introduction

Candida is a heterogeneous genus of yeasts, which may or may not undergo dimorphic transition to pseudohyphae or hyphae (invasive forms). The main pathogenic species of *Candida* in humans are *C. albicans, C. parapsilosis, C. glabrata, C. krusei, and C. tropicalis*, and additionally *C. haemulonii* infections have been reported more frequently since the year 2000, especially in tropical areas (Bastos et al., 2021; Françoise et al., 2023). These fungi can be part of the human body's microbiota, and are either commensal or mutualistic. However, due to changes in the host (such as pregnancy, aging, prematurity, chronic diseases, and stress) or extrinsic factors (use of antibiotics, corticosteroids, contraceptives, antiblastic drugs, surgical interventions, trauma, and burns), it can transition to a parasitic stage, causing infectious diseases collectively called candidiasis (Colombo et al., 2013). In addition to these classically known species of *Candida*, *Candida auris*, an emerging pathogenic fungus, has become the target of more recent studies due to its ability to infect hospitalized patients, causes outbreak, to manifest severe forms of the disease, to be persistent in the hospital environment and patient's skin, and to also be resistant to antifungals (Ahmad and Alfouzan, 2021).

Candida auris was first described in 2009, when it was isolated from the ear canal of a patient in Japan, hence the name "*auris*." Since then, *C. auris* infections have been reported in more than a dozen countries, including the United States, Canada, Colombia, Germany, India, Israel, Japan, Kenya, Norway, Pakistan, Spain, South Africa, South Korea, United Kingdom, Venezuela, Kuwait, Oman, and, recently, in Brazil (Carvajal et al., 2021; de Almeida et al., 2021). *C. auris* can be recovered from various clinical specimens, including sterile body fluids, ear, wounds, and mucocutaneous swabs. However, the main clinical manifestations are invasive and bloodstream infections (Rudramurthy et al., 2017).

Correctly identifying *C. auris* presents a challenge due to its close phylogenetic proximity to other species, such as those in the *Candida haemulonii* complex. Consequently, *C. auris* and the *C. haemulonii* complex share several phenotypic characteristics, complicating their differentiation (Osei Sekyere, 2018). Both species typically display yeast-like growth on standard laboratory media, forming smooth, creamy colonies with similar morphological features, as observed under microscopy. Furthermore, they exhibit resistance to multiple classes of antifungal drugs, including azoles, echinocandins, and polyenes, which complicates treatment strategies and emphasizes the importance of precise identification (Osei Sekyere, 2018).

Distinguishing between *C. auris* and *C. haemulonii* poses significant challenges in clinical laboratories. Conventional phenotypic assays, such as biochemical profiling and morphological characterization, often fail to provide conclusive results due to overlapping traits and variations within the *Candida* genus. Moreover, the absence of species-specific diagnostic markers complicates accurate differentiation, leading to misidentification and potential treatment failures. In fact, misidentifications by biochemical methods are frequent, even with updated databases (Gómez-Gaviria et al., 2023). Therefore, a more accessible, simple, and effective identification becomes essential for studying these multidrug-resistant microorganisms to recognize their specific characteristics. Precisely, there is a need for a more accurate diagnosis for treating infections more quickly and efficiently, without prescribing incorrect medications that ultimately may generate drug-resistant microorganisms. Thus, efforts are needed for discoveries and development of new methods for identification and diagnosis of microorganisms in order to fill gaps and offer medical professionals more possibilities, agility, or precision, depending on their needs.

Infrared spectroscopy is a vibrational technique that has the ability to analyze biological systems, as complex molecules such as proteins, lipids, carbohydrates, and nucleic acids exhibit distinct vibrational behaviors according to their structural and molecular conformation (Neves et al., 2016). Through the emission of electromagnetic radiation in the near-infrared (NIR)—a smaller portion of the infrared spectrum between 900 and 2600 nm—a rapid and accurate diagnosis can be obtained for pathogens isolated from hospital environments and patients (de Sousa Marques et al., 2013).

The use of NIR in recent years has proven to be highly effective for the analysis of various organic, inorganic, and biological substances. NIR offers several advantages, such as rapid identification of various parameters. This efficiency enhances the accuracy of diagnoses, preventing erroneous sample identifications (Cebrián et al., 2021). As NIR is a non-destructive technique requiring little or no sample preparation, it also reduces environmental damage by avoiding or minimizing the use of reagents, which often cause harm to nature. However, for biological samples, this technique itself may not provide sufficient specificity in the search for biomarkers, as many biomolecules are contributing to the entire signal, leading to a large amount of complex data. On the other hand, multivariate analysis has proven to be effective in overcoming this disadvantage (Neves et al., 2016).

There are two classes of multivariate analysis techniques for pattern recognition: unsupervised and supervised methods. The former aims to detect similarities and differences within a dataset composed, for example, of spectra from different classes without prior information about the class to which they belong. Principal component analysis (PCA) is the most popular unsupervised method. On the other hand, in supervised methods, there is prior information of different classes. They are based on two successive steps: first, samples whose class is known are used to build a model with suitable parameters that optimize the discrimination between data from different classes, and then, unknown samples are assigned to an appropriate class using the parameters optimized during the first stage. Linear discriminant analysis (LDA) is an effective supervised approach (Lasalvia et al., 2022).

All measured spectra can be represented as a dataset or matrix X, with n rows corresponding to measured samples and m columns, each corresponding to the spectral signal for a specific wavenumber value. The first objective of PCA is to reduce the dimensionality of large datasets by finding new variables, which are linear functions of those in the original dataset, which successively maximize variance and are uncorrelated with each other (Lasalvia et al., 2022).

LDA is based on a linear transformation of m variables describing n samples belonging to different classes so that samples of the same class are close, but samples from different classes are distant from each other. This goal is achieved through a mathematical classification algorithm (based on calculating the Mahalanobis distance between samples for each class) that maximizes the distance between the means of the classes, while minimizing the variance within each class. Thus, a predicted class is assigned to each sample. After building the classification model, it is used to allocate new and unknown samples to the most likely class. However, the LDA method is generally restricted to problems with few dimensions and cannot be applied when the number of spectral variables is greater than the number of samples (m < n) due to the risk of overfitting since the large number of variables have a high collinearity/redundancy (Morais et al., 2019). This problem can be solved by combining LDA with algorithms that reduce these dimensions, such as PCA, successive projection algorithm (SPA), and genetic algorithm (GA) (José et al., 2005; Lasalvia et al., 2022).

The objectives of this work are to overcome the difficulties of differentiation between two much related species of *Candida*, *C. auris* and C. *haemulonii,* with a new reliable, fast, and relatively less costly method for optimizing diagnoses and methods of research for identifying these yeasts.

Methodology

Microorganisms and growth

We used two strains of *Candida auris* and *Candida haemulonii*: *C. auris* CBS10913 from the Westerdijk Institute collection and *C. haemulonii* CH02, provided by Dr. André Dos Santos. *C. haemulonii* CH02 was isolated from a patient and identified phenotypically using CHROMagar *Candida* (CHROMagar Company) and VITEK 2 (bioMérieux) with the YST card. Additionally, it was identified by sequencing the ITS1-5.8S-ITS2 gene (Ramos et al., 2015).

The yeast strains were cryopreserved in brain heart infusion (BHI) growth medium supplemented with 10% glycerol at −80°C until required for experimentation. For experimental procedures, fungi from frozen stocks underwent two successive subcultures on Sabouraud dextrose agar (SDA) for 48 h at 37°C, followed by another cultivation cycle on SDA under the same conditions. Subsequently, isolated colonies were subjected to NIR spectroscopy. The experiment was conducted across 15 plates, with three colonies selected from each plate to generate spectra. In total, 45 colonies per strain were analyzed.

Preparation of samples for NIR spectra acquisition

One of the advantages and suggestions of this study was the acquisition of spectra without any sample preparation. Using a transflectance probe, each spectrum was obtained by placing the probe directly above the plates containing the colonies. The samples could be used for further analysis following the above method by other researchers, avoiding time loss and reagent consumption.

Obtaining near-infrared spectra

The 90 colonies, 45 per species, were subjected to NIR spectroscopy by a Fourier transform spectrometer ARCspectro ANIR (ARCoptix, Switzerland) with a 99% reflectance reference underneath, in the region between 900 and 2,600 nm. The detector gain was adjusted to extreme, at one scan, and a Boxcar filter was applied every 10 nm in triplicate (isolated colonies) to obtain as much variability within the same sample and among different samples.

Multivariate analysis of infrared data

MATLAB software (MathWorks Inc, Natick, MA, USA) was used to import the dataset, perform pretreatment, and construct multivariate classification models (PCA-LDA, SPA-LDA, and GA-LDA). A total of 30 samples were separated for model training and 15 for testing, applying the Kennard–Stone algorithm for infrared spectra (Kennard and Stone, 1969), i.e., a proportion of 70%–30% for training and testing, respectively. Training samples were used to build and optimize the models (selection of variables using the SPA and GA algorithms), while the test samples were used to evaluate their classification using LDA.

A dataset with many variables can be problematic for LDA classification since the probability functions between classes can spread and overlap very easily. Therefore, the number of variables can be simplified by performing data reduction. PCA is a well-known method for reducing the number of variables, creating new ones called principal components, which are linear combinations of the original variables, in which the spectral matrix X is decomposed as follows:

$$X = TP^{t} + E,$$

where X is the I × J data matrix, T is the I × A matrix of score vectors (representing the sample projection in the new space), the score vectors are orthogonal ($T^{t}T = diag(\lambda_a)$ and λ_a are the eigenvalues of the matrix $X^{t}X$), P is the J × A matrix of loading vectors (weights of the variables), and E is the $I \times$ J residual matrix. *I* is the number of objects, J is the number of variables, and A is the number of calculated components (de Sousa Marques et al., 2013).

One strategy to avoid overfitting in the SPA-LDA and GA-LDA models is to use a validation set to guide variable selection. The optimal number of variables for SPA-LDA and GA-LDA was determined from the minimum of the cost function G calculated for a given validation dataset as (de Sousa Marques et al., 2013)

$$G = \frac{1}{N_v}\sum_{n=1}^{N_v} g_n,$$

where N_v is the number of validation samples and g_n is defined as

$$g_n = \frac{r^2\left(x_n, m_{I(n)}\right)}{min_{I(m)\neq I(n)}\, r^2\left(x_n, m_{I(n)}\right)},$$

where I(n) is the true class index for the nth validation object x_n; the numerator $r^2\left(x_n, m_{I(n)}\right)$ is the squared Mahalanobis distance between the object x_n (of class index I(n)) and the sample mean $m_{I(n)}$ of this true class; and the denominator is the squared Mahalanobis distance between object x_n and the error center of the nearest class, i.e., the wrong class (Neves et al., 2016).

To obtain a discriminant profile, the LDA classification score (L_{ij}) is calculated for a given class k by the following equation:

$$L_{ik} = (x_i - \bar{x}_k)^T \Sigma^{-1}_{pooled}(x_i - \bar{x}_k) - 2\log_e \pi_{k,}$$

where x_i is an unknown measurement vector for the sample i; $\bar{x}_k$ is the measurement vector for the mean of the classes k; Σ_{pooled} is the pooled matrix of covariance; and π_k is the prior probability of class k (Neves et al., 2016).

The GA algorithm was set to a minimum of 40 generations and a maximum of 80 generations. Crossover and mutation probabilities were set to 60% and 10%, respectively, and repeated three times, starting from different random initial populations.

Accuracy (number of samples correctly classified considering true and false negatives) (Morais et al., 2020), sensitivity (SENS, the confidence that a positive result for a sample of the label class is obtained) (Morais et al., 2020), specificity (SPEC, the confidence that a negative result for a sample of the non-label class is obtained) (Morais et al., 2020), G-score (model performance not accounting for class size) (Morais et al., 2020), and AUC (area under the curve that measures the relation between true positives and false positives, giving the probability of a model to classify a random positive example higher than a random negative example) (López et al., 2013) were calculated as important quality parameters in test evaluation.

$$Accuracy\ (\%) = \frac{TP + TN}{TP + FP + TN + FN} \times 100,$$

$$SENS\,(\%) = \left(\frac{TP}{TP + FN}\right) \times 100,$$

$$SPEC\,(\%) = \left(\frac{TN}{TN + FP}\right) \times 100,$$

$$Gscore = \sqrt{SENS \times SPEC,}$$

$$AUC = \frac{1 + TPR - FPR,}{2}$$

where

$$TPR = \left(\frac{TP}{TP + FN}\right),$$

and

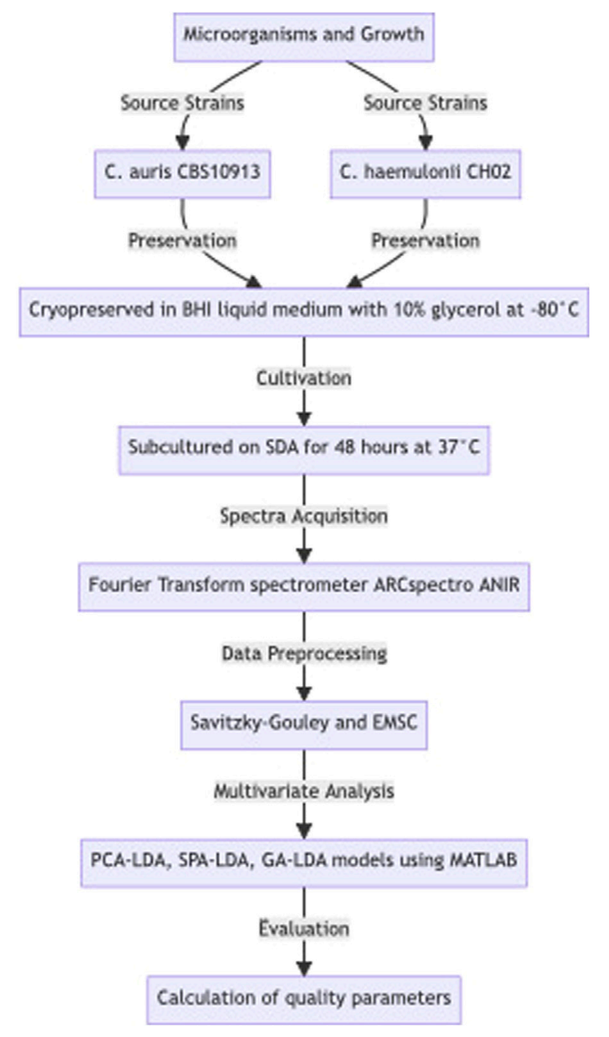
Microorganisms and Growth
Source Strains
Source Strains
C. auris CBS10913
C. haemulonii CH02
Preservation
Preservation
Cryopreserved in BHI liquid medium with 10% glycerol at -80°C
Cultivation
Subcultured on SDA for 48 hours at 37°C
Spectra Acquisition
Fourier Transform spectrometer ARCspectro ANIR
Data Preprocessing
Savitzky-Gouley and EMSC
Multivariate Analysis
PCA-LDA, SPA-LDA, GA-LDA models using MATLAB
Evaluation
Calculation of quality parameters

FIGURE 1
Flowchart describing the methodology.

$$FPR = \left(\frac{FP}{FP + TN}\right).$$

TPR and FPR are true-positive rate (percentage of positive instances correctly classified) and false-positive rate (percentage of negative instances misclassified), respectively, FN is defined as false negative, and FP is false positive. TP and TN are defined as true positive and true negative, respectively (López et al., 2013).

Herein, for all calculations, *C. auris* CBS1093 samples were considered the positive class ("disease group") and *C. hemulonii* CH02 samples as the negative class ("control group"). Figure 1 shows a flowchart describing the methodology of this work.

Results and discussion

Differentiating between *C. auris* and *C. haemulonii* complex species is challenging in clinical practice because these species have phylogenetic proximity and share similar morphological and

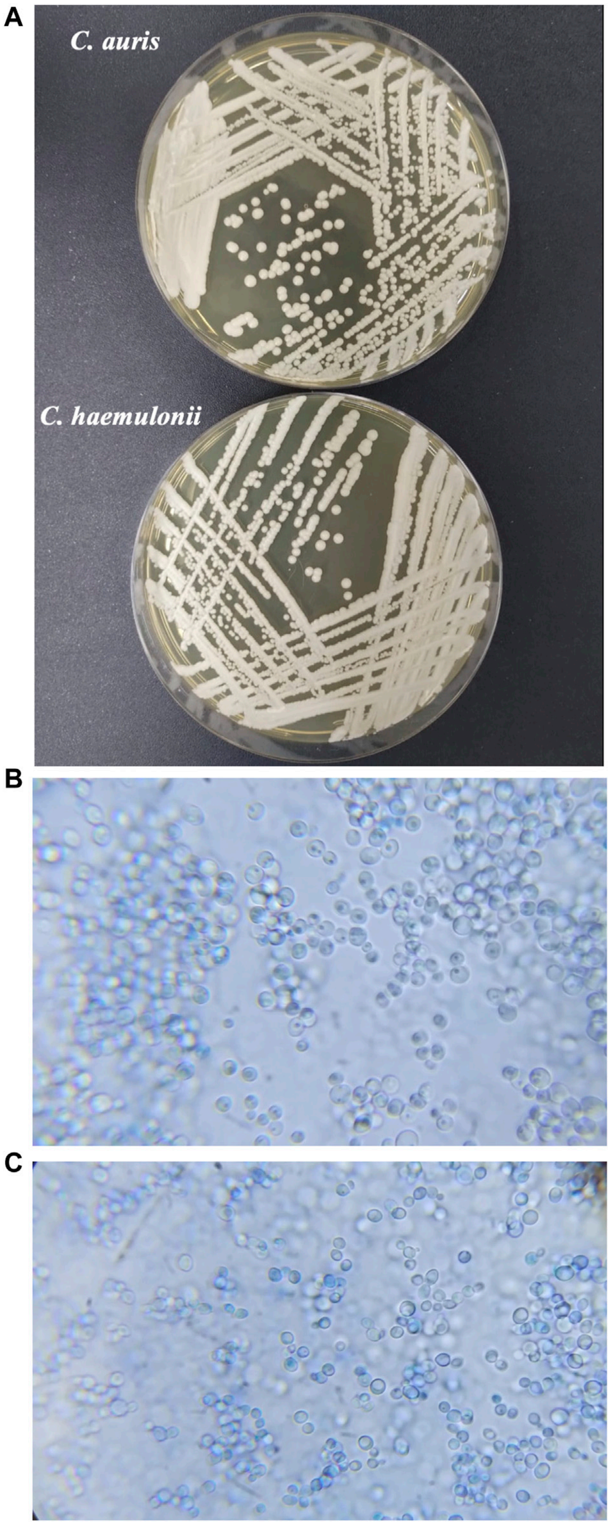

FIGURE 2
(A) *Candida auris* CBS10913 and *C. haemulonii* CH02 samples for visual comparation. Microscopy of *C. auris* CBS10913 **(B)** and **(C)** *C. haemulonii* CH02 in methylene blue at ×100 magnification by an optical microscope.

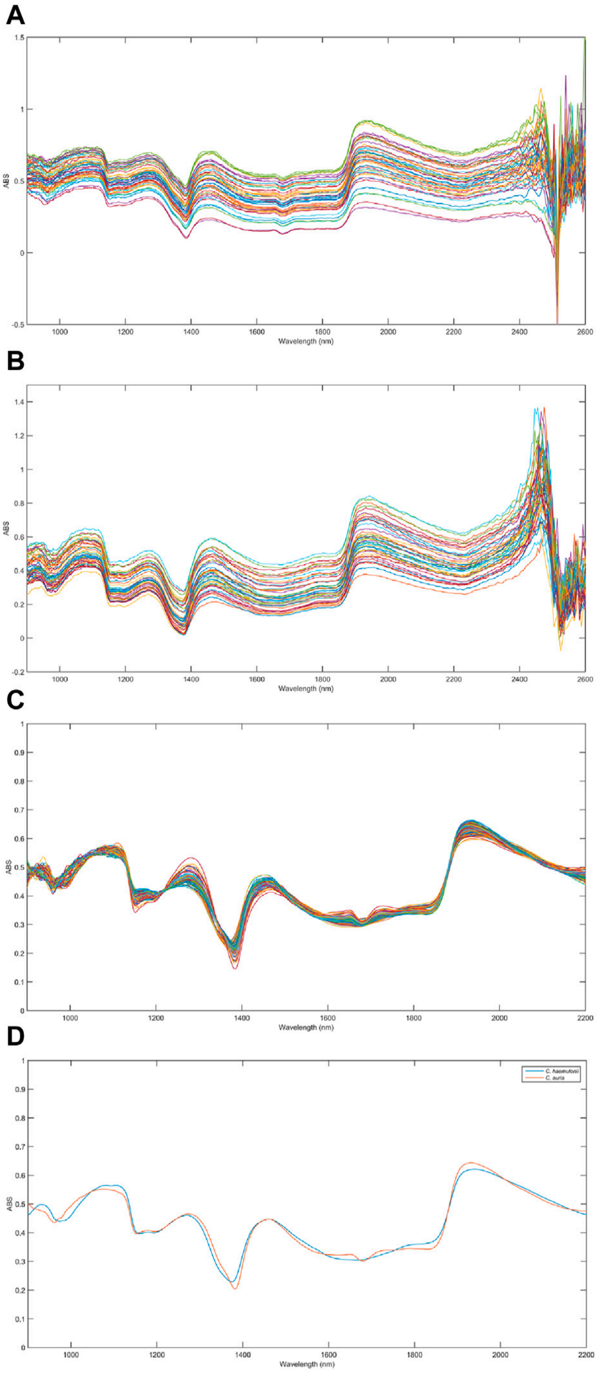

FIGURE 3
Raw infrared spectra for **(A)** *Candida auris* CBS10913 and **(B)** *Candida haemulonii* CH02 samples. **(C)** All pretreated spectra for both species and **(D)** mean pretreated spectra for both species were *C. auris* CBS10913 and *C. haemulonii* CH02 are represented by red and blue lines, respectively.

physiological characteristics. Figure 2 shows the macro-morphology of the colonies (Figure 2A) and the micro-morphology of the cells (Figures 2B, C) of the two species, highlighting how they are morphologically similar. Both species typically display yeast-like growth on standard laboratory media, forming smooth, creamy colonies with similar morphological features observed under microscopy (Figure 2).

Figures 3A, B show the raw NIR spectra obtained for individual colonies (Figure 2A) of *C. auris* CBS10913 and *C. haemulonii* CH02,

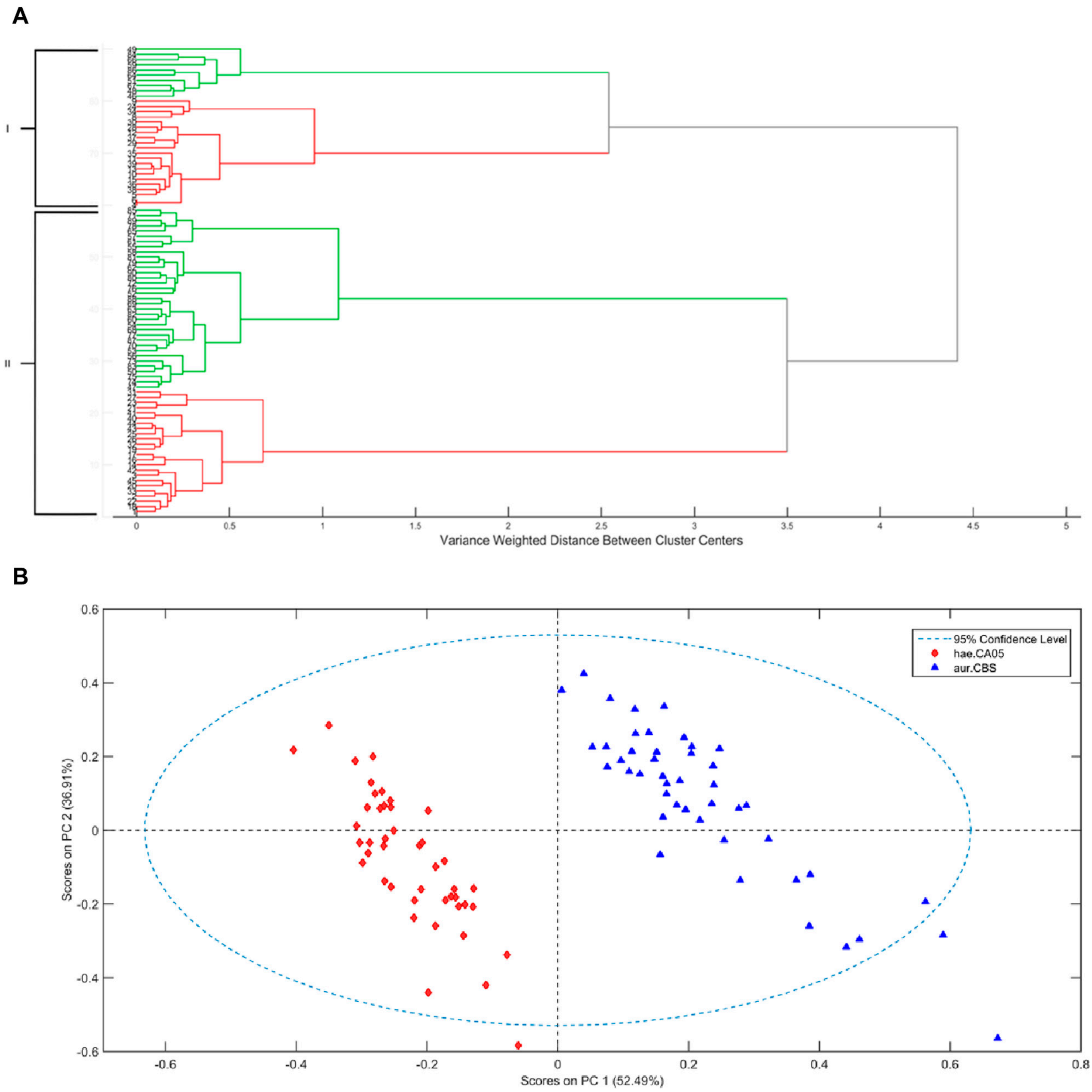

FIGURE 4
HCA **(A)** of NIR for *Candida auris* CBS10913 (red) and *C. haemulonii* CH02 (green). PCA scores for *Candida auris* CBS10913 and *C. haemulonii* CH02 **(B)**.

respectively. Figure 3C shows all spectra, for both species, after pretreatment. Figure 3D presents the mean pretreated spectra for both species. The region between 2,200 and 2,600 nm from the raw NIR spectra (Figures 3A, B) showed a poor signal-to-noise ratio (S/N) and was removed before building the discrimination models as it may not provide any useful information. As pretreatments, Savitzky–Golay smoothing filter (5 points window) and extended multiplicative scatter correction (EMSC) were applied, both from the PLS ToolBox (Eigenvector Research, Inc., Manson, WA, USA) in the MATLAB environment, to improve the signal and correct it for light scatterings, respectively. Figure 3D shows the spectral similarity between classes. The spectra are slightly shifted downward, relative to each other, but these are mean spectra, and distinguishing an isolated spectrum from another to separate the sample class is very difficult, necessitating computational analysis to identify markers responsible for differences between species.

First, an exploratory analysis was carried out by applying hierarchical cluster analysis (HCA) and PCA to observe the behavior of samples regarding their division in clusters related to the species of *Candida* without the need of any prior class information. To perform the HCA, it is necessary to have a metric function for sample distances (in this case the Mahalanobis distance was applied), a linkage criterion among groups (Ward’s linkage was used), and the agglomerative hierarchical

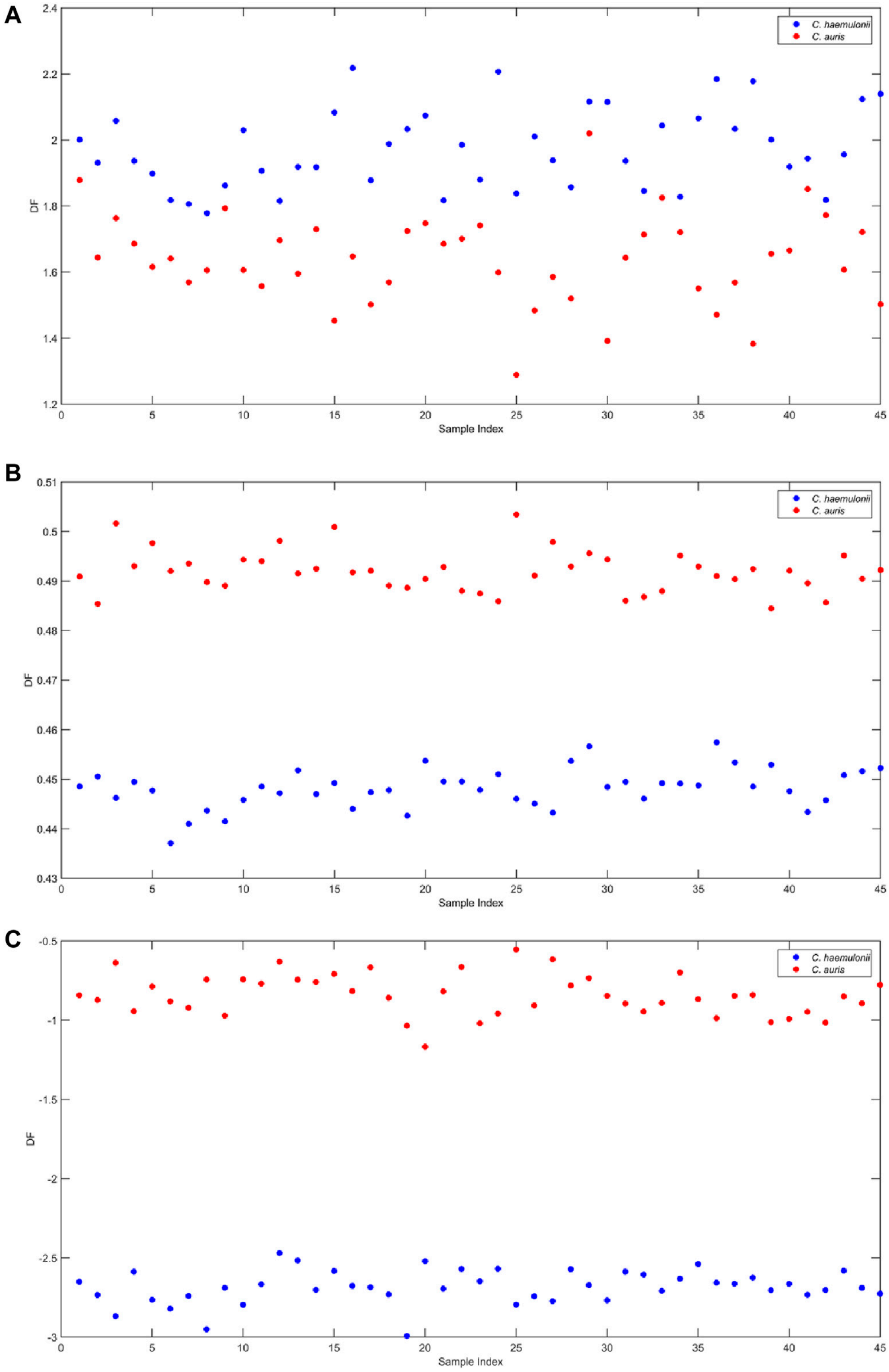

FIGURE 5
(A) PCA-LDA discriminant function over the NIR spectra for *C. auris* CBS10913 and *C. haemulonii* CH02. **(B)** SPA-LDA discriminant function over the NIR spectra for *C. auris* CBS10913 and *C. haemulonii* CH02. **(C)** GA-LDA discriminant function over the NIR spectra for *C. auris* CBS10913 and *C. haemulonii* CH02.

clustering technique was used. Initially, each sample is considered an individual cluster, and subsequently, pairs of clusters are merged based on their similarities. This process results in a dendrogram, a two-dimensional tree-like diagram that shows the groups of merged samples (Neves et al., 2021). The dendrogram in Figure 4A shows the presence of two main

TABLE 1 Table of confusion from PCA-LDA, SPA-LDA, and GA-LDA models.

Actual class		*C. auris*	*C. haemulonii*
PCA-LDA			
	C. auris	15	0
	C. haemulonii	0	15
SPA-LDA			
	C. auris	15	0
	C. haemulonii	0	15
GA-LDA			
	C. auris	15	0
	C. haemulonii	0	15

clusters, where the two classes of *Candida* are mixed, highlighting their similarities. When the NIR data are analyzed by PCA, two distinct groups for each class are formed, as shown in the PCA score plots (Figure 4B). Although the first two principal components of PCA accounted for 89.4% of the explained variance and were able to separate the two distinct groups for each class, a mathematical function is still needed to predict the classes so that the model can be used for unknown samples in future diagnoses.

Candida auris CBS10913 vs. *Candida* haemulonii CH02

Figure 5A represents the Fisher discriminant function for PCA-LDA. Only two PCs were needed to explain 92.22% of data variation and showed good separation. Although the PCA-LDA model satisfactorily discriminated the classes, considering attempts to correlate classification results with biomarker searches by attributing functional groups and/or chemical bonds reflected in NIR wavelengths, PCA-LDA may not be the best option available due to its nature of creating a new space and losing the original information. An alternative to that issue is using SPA and GA algorithms to select the most important variables (wavelengths) for the model. Figures 5B, C show the discriminant function over the spectral observation points for SPA-LDA and GA-LDA, respectively. In addition to selecting the ideal number of variables, these algorithms were able to increase the visual separation.

Table 1 displays the confusion matrix showing real and predicted classes and the number of samples in which each algorithm classified them. For this case, the positive class is represented by *C. auris* CBS10913 and the negative class is represented by *C. haemulonii* CH02. As shown in Table 2, the three methods were capable of classifying the species with maximum sensitivity (meaning their capability to correctly identify patients positive for *C. auris* infection) and specificity (meaning the capability of models to identify patients positive for *C. hemulonii* infection). Although PCA-LDA results are equally satisfactory, SPA-LDA and GA-LDA have identified the most important variables for the models, as depicted in Figures 6A, B.

SPA-LDA selected the following wavelengths in nanometers: 931, 953, 1389, 1830, and 1894. These wavelengths are associated with the third overtone of C-H bonds in hydrocarbons, alcohol secondary overtones, C-H combination bands associated with long aliphatic molecules, O-H/C-H combination bands of polysaccharides, and C=O stretch second overtone of carboxylic acids, respectively. GA-LDA selected 940, 1135, 1384, 1557, 1830, and 1903. These wavelengths are associated with the second overtone of alcohol combination bands, hydrogen bond secondary amide second overtone, C-H combination bands associated with long aliphatic molecules, hydrogen bond secondary amide first overtone, O-H/C-H combination bands of polysaccharides, and hydrogen bond in P-OH group first overtone, respectively (Siesler et al., 2001; Workman and Weyer, 2008). The wavelengths selected by SPA-LDA and GA-LDA correspond highly to the subtle spectral differences observed between the species, which could be attributed to important intrinsic variations such as different polysaccharides, peptides, or protein structures in the cell wall and metabolic products (Garcia-Rubio et al., 2019; Oliver et al., 2020).

The discriminatory ability of the models presented herein is not only equivalent but also outstanding when considering the results achieved for all the quality performance parameters evaluated in this study, as seen in Table 2. However, it is important to note that, among the other reduction algorithms in this study, SPA has the advantage of being deterministic, i.e., it always returns the same selected variables each run. In contrast, GA may vary slightly in the selection of variables across different runs due to its random nature, while the original variables are not present in PCA anymore, as it generates two new latent variables to explain the variance within the data. This aspect of SPA might play an important role when considering the assignment of the main variables responsible for the discrimination and their correlation with groups of molecules that may act as biomarkers. The area under the curve (AUC) of

TABLE 2 Quality performance values from PCA-LDA, SPA-LDA, and GA-LDA models.

Quality performance feature	PCA-LDA	SPA-LDA	GA-LDA
Accuracy (%)	100	100	100
Sensitivity (%)	100	100	100
Specificity (%)	100	100	100
G-score	100	100	100
AUC	1	1	1

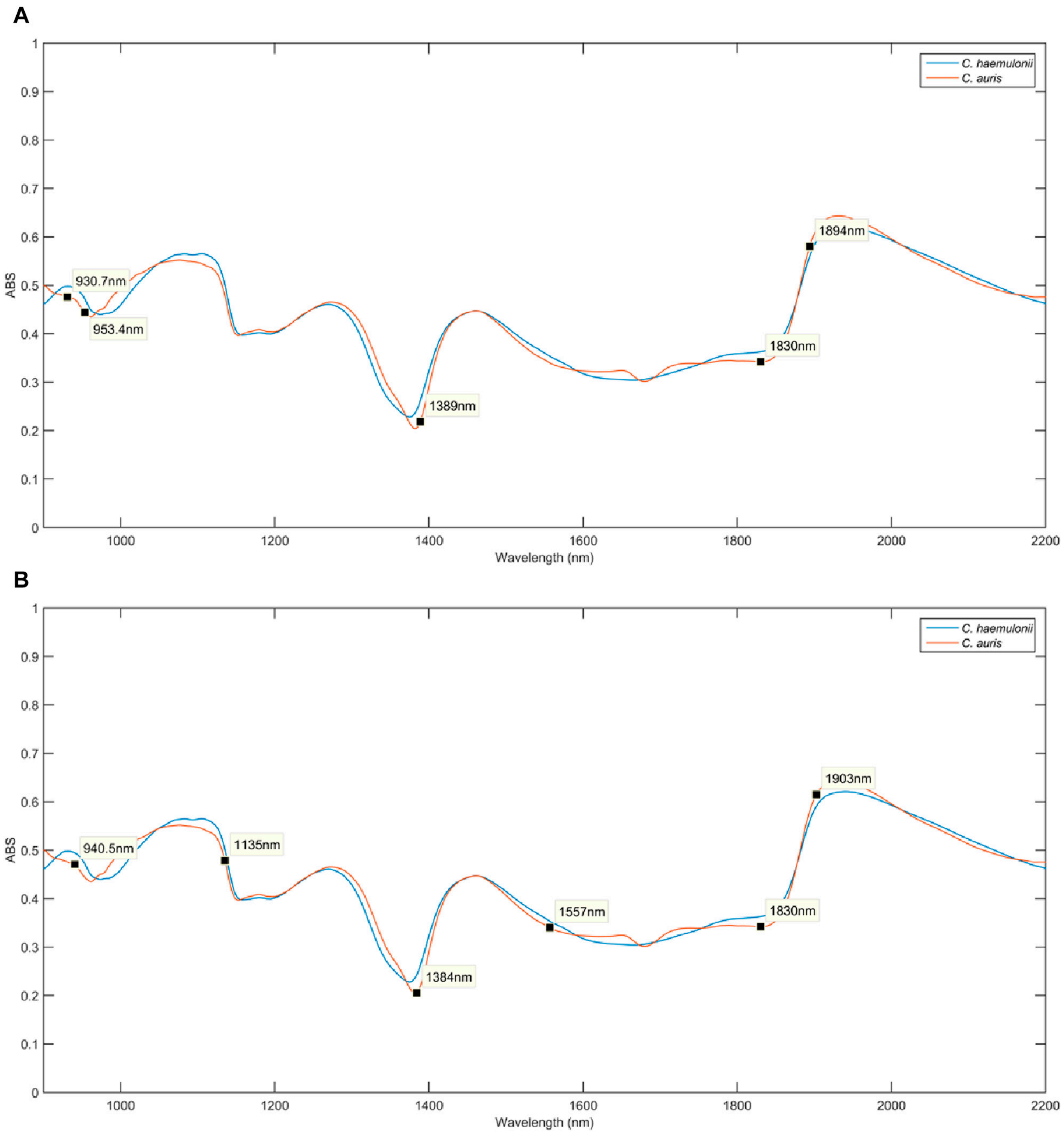

FIGURE 6
(A) Variable selection of SPA-LDA over the mean pretreated spectra for *Candida auris* CBS10913 (orange line) and *C. haemulonii* CH02 (blue line). **(B)** Variable selection of GA-LDA over the mean pretreated spectra for *Candida auris* CBS10913 (orange line) and *C. haemulonii* CH02 (blue line).

prediction samples for SPA-LDA, which evidences the model's capacity to truly classify and not just give random results, accounted for the maximum value of 1.

In addition to differentiating two emergent yeast species, our results are very important in the context of clinical practice. Individuals who are at risk of acquiring *C. auris* infections are primarily hospitalized and nursing home patients. These patients generally have comorbidities that, together with the multidrug-resistant nature of the yeast, contribute to the lethality of this infection, which can reach up to 60% of cases (Wang et al., 2018). Additionally, *Candida haemulonii* has emerged as an opportunistic pathogenic fungus associated with nail infections, onychomycosis, paronychia, vaginal candidiasis, blood infections, and several fungemia related to catheters, osteitis, and outbreaks in ICUs (Leite-Jr et al., 2023). The rapid identification of these species, facilitated by the method employed in this study, enables the prompt application of targeted treatments.

Standard methods to diagnose candidiasis, in general, can be laborious or highly costly. For example, through direct mycological examination (slides of biological material, whether oral, vaginal, or bloodstream mucosa) treated with KOH solution and stained with an appropriate dye and/or by cultivation in specific mycological media, with subsequent identification of the pathogen by microscopic examination of its structures, through automation, MALDI-TOF, or molecular methods (Colombo et al., 2013; Montes et al., 2019). On the other hand, our study provides a more efficient, simple, and cost-effective method to discriminate between *Candida auris* and *Candida haemulonii* since there is no need for highly specialized personnel, sample preparation, or very expensive materials and equipment.

Previous studies have recognized the potential of NIR spectroscopy in mycology. For instance, Cebrián et al. (2021) presented data demonstrating the utility of NIR in identifying substances produced by molds, while Santos et al. (2010) demonstrated the capability of identification and characterization of filamentous fungi and yeast forms through Fourier transform infrared spectroscopy (Santos et al., 2010; Cebrián et al., 2021). Essendoubi et al. (2005) used Fourier transform infrared spectroscopy with hierarchical clustering analysis to distinguish *Candia* species (*Candida albicans*, *Candida glabrata*, *Candida parapsilosis*, *Candida tropicalis*, *Candida krusei*, and *Candida kefyr*) (Essendoubi et al., 2005). However, our study marks the first instance of employing NIR spectroscopy in conjunction with supervised algorithms for this specific purpose. This approach has been shown to be more precise than unsupervised methods, as discrimination is based on a mathematical function that calculates the probability of a sample belonging to a given class, i.e., creates a model. Notably, our methodology successfully differentiated between two *Candida* species, highlighting its efficacy.

Conclusion

This study successfully differentiated between the closely related species *C. auris* CBS10913 and *C. haemulonii* CH02 using NIR spectroscopy combined with multivariate analysis. The SPA-LDA and GA-LDA models achieved 100% accuracy, sensitivity, and specificity in distinguishing these species. These models identified crucial spectral features, demonstrating robust discriminatory capabilities. SPA-LDA showed a significant advantage in biomarker identification due to its deterministic nature. The selected wavelengths correlated with subtle spectral variations, potentially linked to polysaccharides, peptide or protein structures, and metabolic products. These findings suggest that NIR spectroscopy, coupled with advanced multivariate analysis, can offer a rapid, accurate, and cost-effective method for yeast identification, improving clinical diagnostics and treatment strategies. Further validation with larger datasets is recommended to extend this approach to other *Candida* species.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

ANa: data curation, formal analysis, methodology, software, and writing–review and editing. AMe: formal analysis, methodology, validation, and writing–review and editing. ANe: conceptualization, investigation, methodology, validation, and writing–review and editing. AMa: formal analysis, investigation, methodology, validation, and writing–review and editing. LR: formal analysis, methodology, validation, and writing–review and editing. DA: formal analysis, methodology, validation, and writing–review and editing. AS: formal analysis, methodology, validation, and writing–review and editing. KL: funding acquisition, project administration, supervision, writing–original draft, and writing–review and editing. RB: conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, and writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The authors of this work would like to thank the Chemistry Postgraduation Program and Biosciences Center from Federal University of Rio Grande do Norte and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (grant number 405934/2022-0), the National Institute of Science and Technology INCT Funvir from Brazil, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and FAPESP for their financial support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmad, S., and Alfouzan, W. (2021). *Candida auris*: epidemiology, diagnosis, pathogenesis, antifungal susceptibility, and infection control measures to combat the spread of infections in healthcare facilities. *Microorganisms* 9, 807. doi:10.3390/MICROORGANISMS9040807

Bastos, R. W., Rossato, L., Goldman, G. H., and Santos, D. A. (2021). Fungicide effects on human fungal pathogens: cross-resistance to medical drugs and beyond. *PLoS Pathog.* 17, e1010073. doi:10.1371/JOURNAL.PPAT.1010073

Carvajal, S. K., Alvarado, M., Rodríguez, Y. M., Parra-Giraldo, C. M., Varón, C., Morales-López, S. E., et al. (2021). Pathogenicity assessment of colombian strains of *candida auris* in the galleria mellonella invertebrate model. *J. Fungi* 7, 401. doi:10.3390/jof7060401

Cebrián, E., Núñez, F., Rodríguez, M., Grassi, S., and González-Mohino, A. (2021). Potential of near infrared spectroscopy as a rapid method to discriminate OTA and non-OTA-producing mould species in a dry-cured ham model system. *Model Syst.* 13, 620. doi:10.3390/toxins13090620

Colombo, A. L., Garnica, M., Aranha Camargo, L. F., Da Cunha, C. A., Bandeira, A. C., Borghi, D., et al. (2013). *Candida glabrata*: an emerging pathogen in Brazilian tertiary care hospitals. *Med. Mycol.* 51, 38–44. doi:10.3109/13693786.2012.698024

de Almeida, J. N., Francisco, E. C., Hagen, F., Brandão, I. B., Pereira, F. M., Presta Dias, P. H., et al. (2021). Emergence of *Candida auris* in Brazil in a COVID-19 intensive care unit. *J. Fungi* 7, 220. doi:10.3390/JOF7030220

de Sousa Marques, A., Celeste Nunes de Melo, M., André Cidral, T., and Michell Gomes de Lima, K. (2013). Feature selection strategies for identification of *Staphylococcus aureus* recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection. *A case study*. doi:10.1016/j.mimet.2013.12.015

Essendoubi, M., Toubas, D., Bouzaggou, M., Pinon, J. M., Manfait, M., and Sockalingum, G. D. (2005). Rapid identification of *Candida* species by FT-IR microspectroscopy. *Biochim. Biophys. Acta Gen. Subj.* 1724, 239–247. doi:10.1016/j.bbagen.2005.04.019

Françoise, U., Desnos-Ollivier, M., Govic, Y.Le, Sitbon, K., Valentino, R., Peugny, S., et al. (2023). *Candida haemulonii* complex, an emerging threat from tropical regions? *PLoS Negl. Trop. Dis.* 17, e0011453. doi:10.1371/journal.pntd.0011453

Garcia-Rubio, R., de Oliveira, H. C., Rivera, J., and Trevijano-Contador, N. (2019). The fungal cell wall: *Candida, cryptococcus*, and *Aspergillus* species. *Front. Microbiol.* 10, 2993. doi:10.3389/FMICB.2019.02993

Gómez-Gaviria, M., Martínez-álvarez, J. A., Chávez-Santiago, J. O., and Mora-Montes, H. M. (2023). *Candida haemulonii* complex and *Candida auris:* Biology, virulence factors, immune response, and multidrug resistance. *Infect. Drug Resist* 16, 1455–1470. doi:10.2147/IDR.S402754

José, M., Pontes, C., Kawakami, R., Galvão, H., César, M., Araú Jo, U., et al. (2005). *The successive projections algorithm for spectral variable selection in classification problems*. doi:10.1016/j.chemolab.2004.12.001

Kennard, R. W., and Stone, L. A. (1969). Computer aided design of experiments. *Technometrics* 11, 137–148. doi:10.2307/1266770

Lasalvia, M., Capozzi, V., and Perna, G. (2022). A comparison of PCA-LDA and PLS-DA techniques for classification of vibrational spectra. *Appl. Sci. Switz.* 12, 5345. doi:10.3390/app12115345

Leite-Jr, D. P., Vivi-Oliveira, V. K., Maia, M. L. S., Macioni, M. B., Oliboni, G. M., De Oliveira, I. D., et al. (2023). The *Candida* genus complex: Biology, evolution, pathogenicity virulence and one health aspects, beyond the *Candida albicans* paradigm. A comprehensive review. *Virology Immunol. J.* 7, 1–38. doi:10.23880/vij-16000331

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inf. Sci. (N Y)* 250, 113–141. doi:10.1016/j.ins.2013.07.007

Montes, K., Ortiz, B., Galindo, C., Figueroa, I., Braham, S., and Fontecha, G. (2019). Identification of *candida* species from clinical samples in a honduran tertiary hospital. *Pathogens* 8, 237. doi:10.3390/pathogens8040237

Morais, C. L. M., Lima, K. M. G., Singh, M., and Martin, F. L. (2020). Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat. Protoc.* 15, 2143–2162. doi:10.1038/s41596-020-0322-8

Morais, C. L. M., Paraskevaidi, M., Cui, L., Fullwood, N. J., Isabelle, M., Lima, K. M. G., et al. (2019). Standardization of complex biologically derived spectrochemical datasets. *Nat. Protoc.* 14, 1546–1577. doi:10.1038/s41596-019-0150-x

Neves, A. C. O., Silva, P. P., Morais, C. L. M., Miranda, C. G., Crispim, J. C. O., and Lima, K. M. G. (2016). ATR-FTIR and multivariate analysis as a screening tool for cervical cancer in women from northeast Brazil: a biospectroscopic approach. *RSC Adv.* 6, 99648–99655. doi:10.1039/c6ra21331f

Neves, A. C. O., Viana, A. D., Menezes, F. G., Wanderlei Neto, A. O., Melo, M. C. N., and Gasparotto, L. H. S. (2021). Biospectroscopy and chemometrics as an analytical tool for comparing the antibacterial mechanism of silver nanoparticles with popular antibiotics against *Escherichia coli*. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 253, 119558. doi:10.1016/j.saa.2021.119558

Oliver, J. C., Laghi, L., Parolin, C., Foschi, C., Marangoni, A., Liberatore, A., et al. (2020). Metabolic profiling of *Candida* clinical isolates of different species and infection sources. *Sci. Rep.* 10, 16716. doi:10.1038/S41598-020-73889-1

Osei Sekyere, J. (2018). *Candida auris*: a systematic review and meta-analysis of current updates on an emerging multidrug-resistant pathogen. *Microbiologyopen* 7, e00578. doi:10.1002/MBO3.578

Ramos, L. S., Figueiredo-Carvalho, M. H. G., Barbedo, L. S., Ziccardi, M., Chaves, A. L. S., Zancopé-Oliveira, R. M., et al. (2015). *Candida haemulonii* complex: species identification and antifungal susceptibility profiles of clinical isolates from Brazil. *J. Antimicrob. Chemother.* 70, 111–115. doi:10.1093/JAC/DKU321

Rudramurthy, S. M., Chakrabarti, A., Paul, R. A., Sood, P., Kaur, H., Capoor, M. R., et al. (2017). *Candida auris* candidaemia in Indian ICUs: analysis of risk factors. *J. Antimicrob. Chemother.* 72, 1794–1801. doi:10.1093/JAC/DKX034

Santos, C., Fraga, M. E., Kozakiewicz, Z., and Lima, N. (2010). Fourier transform infrared as a powerful technique for the identification and characterization of filamentous fungi and yeasts. *Res. Microbiol.* 161, 168–175. doi:10.1016/J.RESMIC.2009.12.007

Siesler, H. W., Ozaki, Y., Kawata, S., and Heise, H. M. (2001). *Near-infrared spectroscopy*. 1st Edn. United States: Wiley VCH.

Wang, X., Bing, J., Zheng, Q., Zhang, F., Liu, J., Yue, H., et al. (2018). The first isolate of *Candida auris* in China: clinical and biological aspects. *Emerg. Microbes Infect.* 7, 1–9. doi:10.1038/S41426-018-0095-0

Workman, J., and Weyer, L. (2008). *Practical guide to interpretive near-infrared spectroscopy*. 1st Edn. United States: CRC Press.