Version of Record: https://www.sciencedirect.com/science/article/pii/S0924203123000425
Manuscript_61770fba1e7964067a517d172c5fd9a8

Automatic Recognition of Food Bacteria Using Raman Spectroscopy and Chemometrics: A Comparative Study of Multivariate Models

O.H. Dib [1,3], A. Assaf [1*], E. Grangé[1], J.F. Morin[4], C.B.Y. Cordella [2,3], G. Thouand [1]

[1] Nantes Université, ONIRIS, CNRS, GEPEA, UMR 6144, F-85000, La Roche-sur-Yon, France

[2] Département Sciences des Aliments, Laboratoire de Recherche et de Traitement de l'Information Chimiosensorielle – LARTIC, Université Laval, Pavillon Paul-Comtois, 2425, rue de l'Agriculture, Canada

[3] UMR 914 Physiologie de la Nutrition et du Comportement Alimentaire, Groupe *Analyse-Chimiométrie-Modélisation,* INRA/AgroParisTech, Université Paris-Saclay, 16 Rue Claude Bernard F-75005 Paris, France

[4] Eurofins Alimentaire, 9 Rue Pierre Adolphe Bobierre, BP 42301, 44323 Nantes, Cedex 3, France.

* Corresponding author: ali.assaf1@univ-nantes.fr

Author contributions

All authors have contributed equally to this work

Abstract

Food safety is the foundation of trust for food stakeholders. Contamination, especially from biological sources, at food processing plants can threaten this foundation, resulting in negative impacts on consumer health and substantial economic losses. Therefore, a rapid, effective, and noninvasive method for detecting bacteria in the food industry is essential. In this study, Raman micro-spectroscopy with advanced statistical tools is proposed as a mean of detecting and differentiating between various types of bacteria. This approach circumvents the complexities of traditional culture-based detection methods. Specifically, fifty-two bacterial strains of 39 different genera were analyzed using Raman spectroscopy. As a result, about 2,563 Raman spectra were generated and integrated into the database. This huge amount of spectral data was analyzed using several chemometric tools, including principal component analysis (PCA), factorial discriminant analysis (FDA) k-nearest neighbors' algorithm (KNN), and convolutional neural network (CNN). Our multivariate data analysis showed that the developed method is rapid and capable of distinguishing several strains. While, FDA models showed mediocre performances, KNN models provided good bacterial classification for most of the analyzed strains (average correct classification 90–95%). In comparison, CNN achieved a higher classification accuracy, of 97%, compared with other models. Combining Raman spectroscopy with chemometric tools yields a robust bacterial assessment method that is simple, rapid, and efficient.

Keywords: Bacteria, Chemometric tools, Classification, Raman spectroscopy

1. Introduction

Foodborne diseases (FBD) triggered by pathogenic bacteria are still considered a significant cause of morbidity and mortality worldwide. According to the World Health Organization (WHO), at least 420 thousand deaths occur every year due to the consumption of contaminated food, based on their "Estimates of the global burden of foodborne diseases" report [1]. Furthermore, these diseases can also be an economic burden, with recent estimates showing an annual cost of up to 90 billion dollars in the United States [2]. Although, current pathogen detection methods, including enzyme-linked immunosorbent assay (ELISA) [3], loop-mediated isothermal amplification (LAMP) [4], and conventional biochemical detection [5], are reliable, they are lengthy procedures that cannot produce test results rapidly. Therefore, for food quality testing, a fast and accurate detection tool for foodborne pathogens is needed.

Over recent years, advancements in Raman spectroscopy have opened new research avenues by allowing rapid and non-destructive analysis. This optical method uses the inelastic scattering of light to produce a structural fingerprint of high specificity. The molecular vibrations caused by the interaction of light with the target sample reveal almost all chemical components, including nucleic acids, carbohydrates, lipids, and proteins [6, 7]. The high sensitivity of this technique allows it to identify organisms at the level of single cells [8, 9]. However, like many other modern analytical instruments, Raman spectroscopy produces a significant amount of information (variables) for a large number of samples in a relatively short time. This results in multivariate data matrices that require the use of chemometric tools to extract the maximum useful information.

The most common chemometric tools used to explore the potential features and classify Raman signals are principal component analysis (PCA), linear discriminant analysis, support vector machines (SVMs), and deep learning methods such as convolutional neural networks

(CNN) [10, 11]. However, due to the complexity of microbial composition, large datasets acquired during Raman analyses mean that unsupervised methods are no longer adequate for data analysis [12]. In fact, PCA is widely used in processing Raman spectra, but can only perform dimensionality reduction, which may not be helpful for some discrimination situations. In addition, processing Raman spectra by PCA in groups results in low-efficiency data usage and many essential features cannot be easily identified using scatterplots or other PCA output [13]. In comparison with unsupervised methods (PCA), supervised learning algorithms, such as factorial discriminant analysis (FDA), K-nearest neighbors (KNN), and CNN, rely on model training. This is done by presenting known samples to identify features and employing them to perform classification. Machine learning methods have been successfully applied in clustering, regression, and classification tasks on large data matrices, especially for the differentiation and identification of bacterial organisms [5, 8, 9]. For example, Tang et al.[14] used ten supervised machine learning methods, including KNN and CNN, to analyze 117 *Staphylococcus* strains and found CNN to have the highest accuracy, at 98.22%, followed by KNN, at 96.22%. In addition, Ho et al. [9] utilized CNN to identify 30 common bacterial pathogens, with an accuracy above 82%.

This study aims to enrich the Raman database with a diverse range of food-contaminating bacterial species (n = 52) from various families and genera., In terms of the number of species included, this study has the highest diversity compared to other studies done in the same field [5,8,9,14]. Furthermore, this paper reports two strategies for the rapid identification of bacteria, utilizing several chemometric tools such as FDA, KNN, and CNN. These two strategies meet all the requirements for the rapid differentiation of the presented bacterial species, including reduced time (with minimal sample preparation and atomization) and analysis costs (with minimal use of reagents).

2. Materials and methods

2.1 Culture medium

The M92 medium was made by adding 30 g of trypticase (Difco, ref. 211825, France) and 3 g of yeast extract (Thermo Fisher, ref. 212750, US) to 1 L of distilled water. The pH was verified and adjusted to 7.0–7.2 with HCl (1M) or NaOH (1M). For agar plates, 15 g of agar (Biokar diagnostics, ref. A1012HA, France) was added to M92 medium (1 L). The solutions were autoclaved at 121°C for 20 minutes and aliquoted into sterile Petri dishes or stored in an amber glass bottle at 4°C until use.

2.2 Bacterial strains

Fifty-two bacterial species belonging to 22 different families were used in this study. These families are divided into 39 genera, including 25 Gram-positive and 27 Gram-negative species. Fig. 1 and Table S1 summarize all the strains analyzed.

All strains were inoculated from cryogenic tubes stored at -80°C in trypticase soy yeast extract medium (M92) (Difco, ref. 211825, France) with a cryoprotectant (15 v/v of sterile glycerol). The precultures were carried out in flasks (100 ml) containing 10 ml of M92 and incubated overnight at 30°C with stirring at 250 rpm (Innova® 42R, Eppendorf, France). The precultures were used to inoculate each bacterial strain in triplicate in 250 ml flasks containing 50 ml of culture medium. The starting optical density (OD) was equal to 0.1 for the cultures, which were grown in triplicates (UV-Vis spectrophotometer, Helios Ɛ, UVE 082917, France). Bacterial growth was followed by measuring the optical density at 620 nm over time (h). At the exponential growth phase, 15 ml of each culture was centrifuged at 6,000 *g* for 5 minutes (Awel, MF 20-R, France), and 10 μl of the biomass obtained was used for Raman analysis, as described in an earlier study [15].

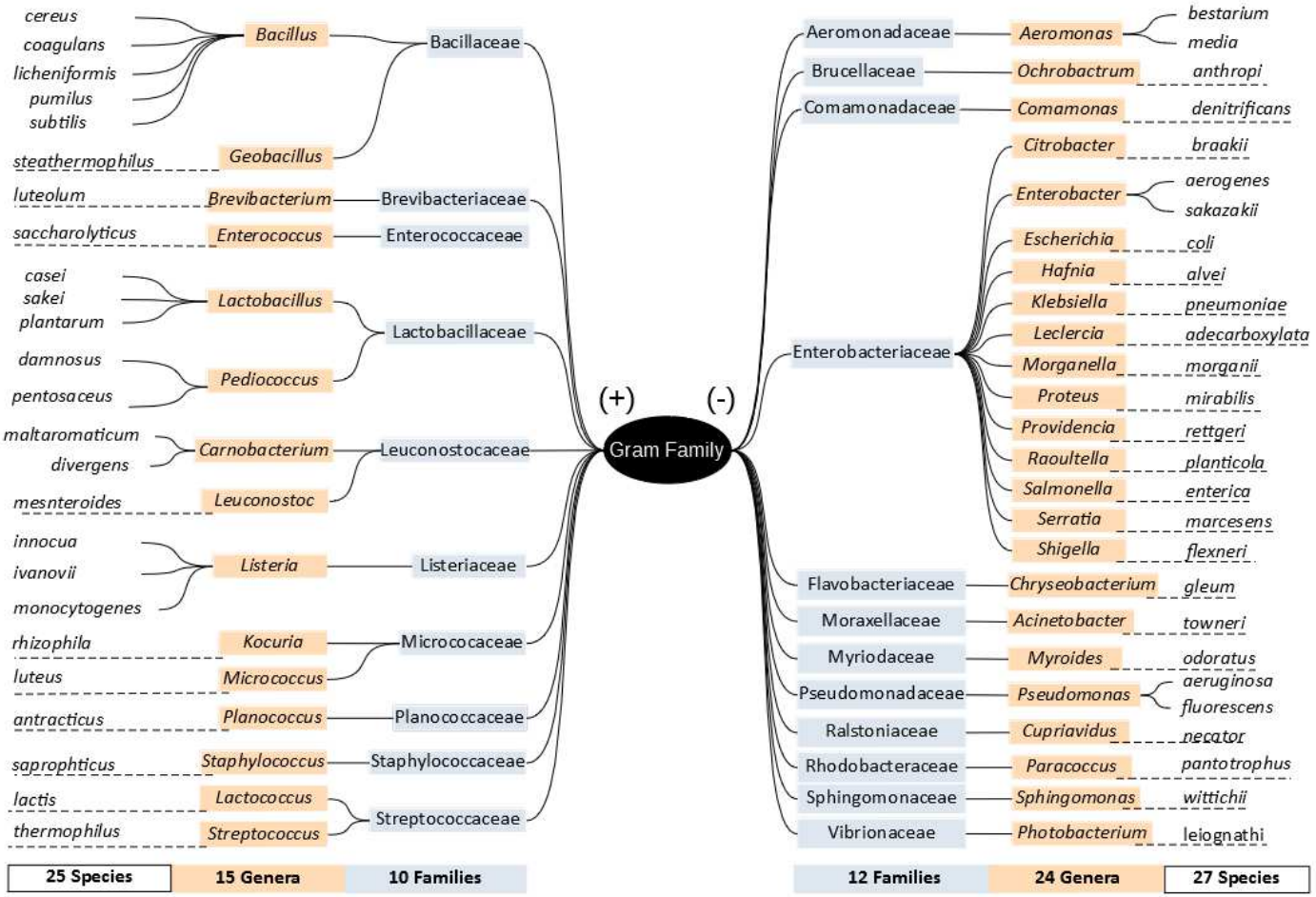


Fig. 1 Family tree of the 52 bacterial species involved in this study. These species belong to 22 different families divided into 39 genera, including 25 Gram positive and 27 Gram negative species

2.3 Raman micro-spectroscopy measurements

A Raman spectrometer (Senterra, Bruker Optics, France) driven by Opus software (Bruker Optics GmBH, V 7.2, Germany) was used to acquire the Raman spectra. This spectrometer is equipped with two gratings (400 and 1200 lines/mm), a CCD camera cooled to -60°C, and a BX51 Olympus microscope with multiple objectives (the objective LCLanN50x/0.5 was used in these analyses, the laser spot = 1.91 µm). The analyzes were performed at 785 nm with a laser power of 25 mW on the samples. The spectral resolution was approximately equal to 8 cm^{-1}. Five acquisitions of 10 seconds each were necessary for each spectrum. In total, 45 spectra were obtained on each bacterial deposit. The Raman spectrometer was automatically calibrated by the patented SureCal^{TM} technology as reported in a previous study [16].

Gold surfaces were prepared according to Assaf et al., 2014 and Kanso et al., 2008 [15,17]. Briefly, the glass microscope slides (ISO 8037, ref. RS, France) were cut into rectangular areas (26x9x1 mm) and cleaned with Piranha solution (70% v/v H2SO4 and 30% v/v H2O2) at 70 °C for 30 minutes. Then the slide surfaces were washed several times with deionized water and ethanol (Labogros, ref. 9006902). After drying, a 30 nm chromium layer (Sigma Aldrich, ref. 266264) and a 100 nm gold layer (Goodfellow, ref. AU005160/72) were deposited onto surfaces by Physical Vapor Deposition using Alcatel-built machine [15].

2.4 Spectral preprocessing

Raman spectra were processed using Opus software (Bruker optics GmBH, V 7.2, Germany). The spectral range 250–3100 cm^{-1} was used in this study for the classification of bacteria. All spectra were baseline corrected using an elastic concave method (64° and ten iterations) and then normalized using min-max normalization. The selection of good-quality spectra was directed by choosing the spectra of bacteria in the exponential phase, especially those with a high DNA/RNA ratio at 780–820 cm^{-1}, as reported in a previous study [15].

2.5 Data analysis

Two strategies were implemented before starting the data analysis. The first strategy involved creating an FDA (based on PCA scores) [18] and a KNN model for each level of bacterial classification, based on the Gram family of the species. The objective was to reduce computational costs, improve the testing parameters, and avoid overfitting.

The second strategy, using a deep learning network, involved the direct classification of bacteria at the species level, irrespective of the Gram family.

2.5.1 Principal Component Analysis (PCA)

PCA is considered one of the most widely used exploration methods in the data sciences [19]. PCA uses an orthogonal transformation to represent the initial data matrix X (original matrix) by a product of two new matrices, T and P, respectively the scores matrix and the loading matrix while maintaining the maximum variance [20].

$$\mathbf{TP}^t + \mathbf{E} = \mathbf{X} \quad (1)$$

where P^t corresponds to the transposed matrix of P, and E is the residuals matrix.

A PCA model with an initial X data matrix of 2925 × 5641 was made. All data were normalized using the standard normal variate (SNV) before running the PCA. Six principal components (PCs) were extracted, covering 93.7% of the total variance, whereas the first two components accounted for 72.8% of the variability in the dataset. The number of selected PC components was determined using a scree plot.

2.5.2 Factorial discriminant analysis (FDA)

Stepwise factorial discriminant analysis (FDA) was done according to Bertrand et al. [18]. The basic idea of this method is to assess a factorial discriminant analysis on the scores of a previous PCA computed on the initial X data matrix. The most discriminating PC scores were selected based on the maximization of the trace of $T^{-1}B$, where **T** is the total variance–

covariance matrix, and **B** is part of the **T** matrix describing the variability among the groups. The ***fda2*** function of the SAISIR® package was used to execute the FDA [21]. The first 10 PC scores were used in the calculation, and the displayed results of the FDA were the average of 100-fold iterations. The data was divided into a calibration and validation set where the latter compromised ¼ of the total spectrum (640 spectra). It is worth mentioning that the replicates from each sample were maintained in the same dataset to ensure that the cross-dispatching of replicates does not impact the validation model.

2.5.3 K-nearest neighbors (KNN)

KNN is a supervised machine learning distance-based algorithm first developed by Fix and Hodges [22] and improved by many authors including Coomans and Massart (23). It is used to predict a test input according to the k training samples (which are the nearest neighbors to the test input), and to assign this test input to the class that has the largest class probability [24]. This is called the majority vote procedure. The advantage of the k-nearest-neighbor classifier is its simplicity as there are only two parameter choices to be set: the number of neighbors, k, and the distance metric to be used.

Parameter k has a great impact on the classification rate of the KNN model. We determined the optimal value of k = 5 during the KNN calibration process. The value k = 5 means that the five closest samples were used to assign missing data. The number of neighbors was selected by testing the quality of the classifier on a test dataset. For the distance metric, the Euclidean distance between Raman spectra was used. Each spectrum is represented by a point in a mathematical space with P dimensions defined by the chosen wavelengths (P = number of spectral elements). In this space, we can calculate a Euclidean distance between two spectra, a and b, from their ordinates, a(i) and b(i), by:

$$d_{a,b} = \sqrt{\sum_{i=1}^{N} (a_i - b_i)^2} \quad (2)$$

Similar strains with very similar spectra will therefore be represented by close points and will thus delimit areas in space. Data partitioning was executed in the same way as for the FDA.

2.5.4 Convolutional neural networks (CNN)

Deep learning networks such as recurrent neural networks, deep neural networks, and convolutional neural networks are a branch of machine learning. They have been applied in numerous fields, including text analysis [25], speech recognition [26], drug design [27], image analysis [28], etc. The function of such networks is to automatically learn associations and extract abstract features from big data [29]. The term "deep" in deep learning refers to the use of multiple layers in the network. Here, a multilayer CNN model (13 layers) was designed to classify 52 bacterial species, as shown in Fig. 1. The overall structure of the CNN model is made up of two convolutional layers, two max-pooling layers, a dropout layer, two full connection layers, and a softmax layer [30].

A one-dimensional vector (1×5641) containing the entire Raman spectrum range was used as input to the CNN. This transformation is crucial to adjust to the next layer since it is a 2D convolutional layer [30]. Each convolutional layer applies convolution to its input (as in Equation 2) followed by batch normalization and relu layer,

$$Y_i = X * W_i + b_i \tag{3}$$

where Y_i represents the output calculated by the i-th convolutional kernel W_i in the convolutional layer and the corresponding bias b_i. X is the input spectrum (1×5641).

The filter size of both convolutional layers was set at [1,3]. The number of filters was set at 32 for the first convolutional layer and 64 for the second convolutional layer.

CNN is known to be a hungry model that requires a large volume of data. For this reason, we used the additive white Gaussian noise function in MATLAB for data augmentation. White Gaussian noise with signal-to-noise ratios of 15, 25, and 30 were added to the original matrix. After augmentation, the total number of Raman spectra was 5126. Sixty percent of the

augmented spectral dataset (3076 spectra) was used for training (making sure that the replicates of the same sample were indeed in the same dataset), 20% of the data (1025 spectra) was used for validation, and 20% of the data (1025 spectra) was used for prediction. The prediction set was set aside, and the remaining sets were used for model training and validation. To train the model, five-fold cross-validation was performed and, based on the results of several preliminary tests, the following hyperparameters were selected: stochastic gradient descent with momentum (SGDM) optimizer; learning rate = 0.001; L2 regularizer = 0.005; batch size = 64.

The PCA and FDA models were computed using the SAISIR Package [21] while the KNN and CNN models were done in MATLAB using the Matlab Statistics and Machine Learning Toolbox (version R2019b, MathWorks Inc, Natick, MA, USA). All classification models were run on an NVIDIA Quadro P400 GPU (NVIDIA Corporation, Santa Clara, CA, USA).

3. Results and Discussion

Raman spectroscopy is an effective tool for identifying bacterial species as it provides valuable information on the chemical bonds present in bacteria. For instance, Fig. 2 shows the Raman spectra for some of the bacterial species present in this study, while Table S2 provides an overview of the various molecular vibrations present in a bacterium's spectrum. The Raman bands at around 657, 720, 850, and 1000 cm-1 are associated with the stretching vibrations of the C-C and C-N bonds in the amino acid residues

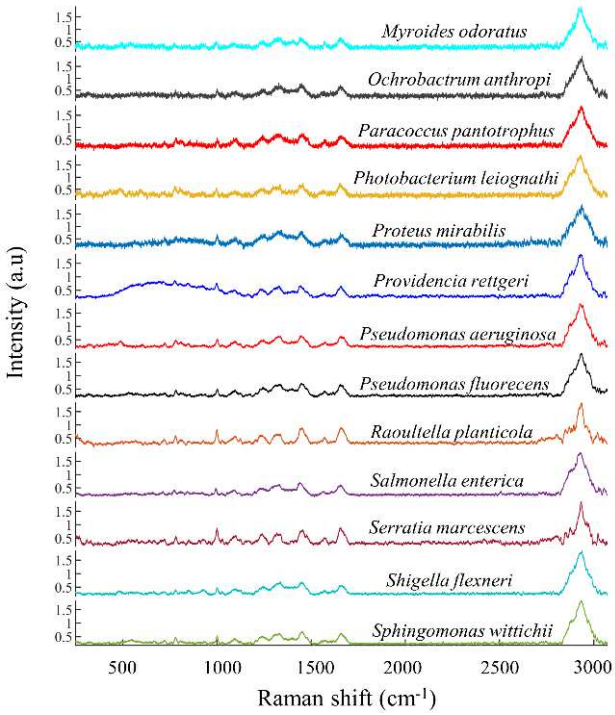


Fig 2. Raman spectra of some of the bacterial species present in this study. (See Fig.S1 for all 52 bacterial species)

of bacterial proteins. Another band at around 1650 cm-1 is attributed to the amide I vibration. Other prominent Raman bands in bacteria include those corresponding to nucleic acids (779 and 810 cm-1), lipids (2880, 2933 cm-1) and carbohydrates (550 cm-1) [31,32]. Although visually differentiating between bacterial species based on their molecular vibration is difficult (Fig.2 and Fig.S1), the bacterial spectra were utilized in conjunction with chemometric tools using two strategies to facilitate the identification of bacterial species.

3.1 Strategy 1 – Level-by-level identification

This strategy enables the identification of bacterial species level-by-level, tracing them from the Gram level to the species level. Three chemometric tools were used in this scenario: PCA, FDA and KNN. PCA was applied exclusively at the Gram family (+ or -) level. FDA was conducted on the PCA scores and applied to each level of bacterial classification. KNN was carried out on the initial data matrix and also applied as a level-by-level bacterial classifier.

3.1.1. PCA

PCA was performed to provide an overview of the Raman spectra under study (Fig.2 & Fig

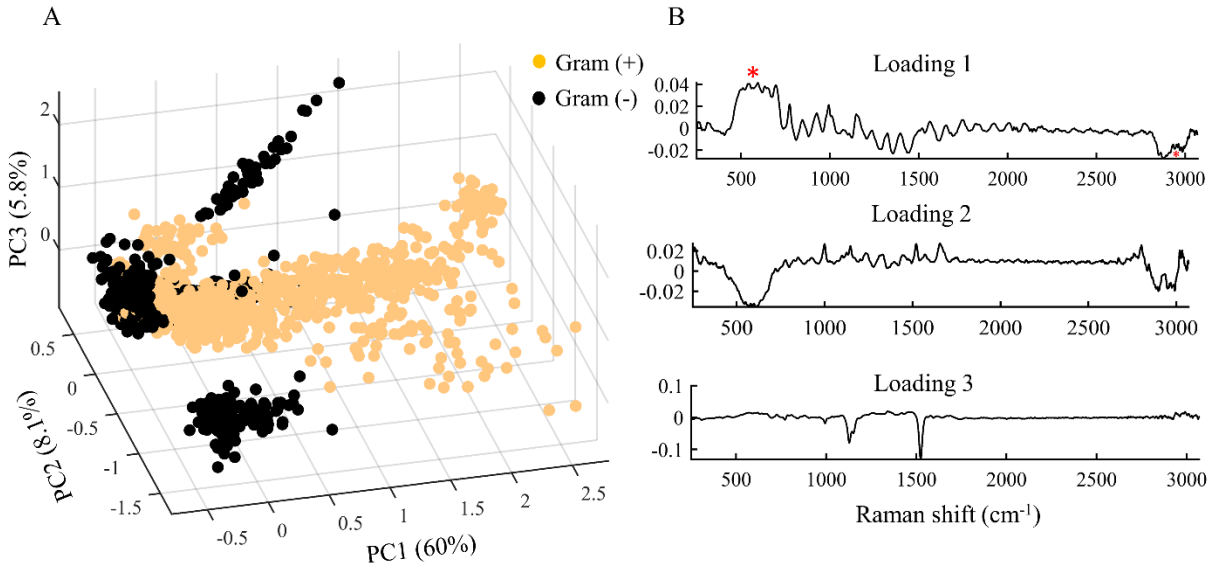


Fig. 3 A view of PCA scores (A) and loadings (B) for discriminating spectra of Gram (+) from Gram (-) bacteria. (B) Loading 1provided interesting information on the structural differences of Gram bacteria, specifically the bands at 550 and 2933 cm^{-1}(highlighted by an Asterix). However, loadings 2 and 3 presented insufficient information to validate the discrimination of bacteria at the Gram level.

S1) and visualize whether the bacteria could be discriminated at the Gram family level. The PCA analysis shown in Fig. 3 indicates that the discrimination of bacteria strains is possible, to some extent, at the Gram level with respect to our Raman analytical conditions; however, the Gram (+)/Gram (-) separation is not complete. Principal component 1 (PC1) accounts for 60% of the variability, while PC2 and PC3 account for 8.1% and 5.8%, respectively. Taking a closer look at PC1 reveals two sets of bacteria. The first set, with positive PC1 scores, belongs mainly to Gram (+), while the second set of bacteria, with negative PC1 scores, is composed mostly of Gram (-) bacteria (Figure 3A). The main contributions for PC1 were from the carbohydrate band (550 cm^{-1}) correlating with a positive sign of loadings (Figure 3B) and from the lipid band (2933 cm^{-1}), correlating with a negative sign of loadings (Figure 3B). These results are consistent with the structural differences between Gram (+) and Gram (-) bacteria. Gram (+) bacteria have a thick peptidoglycan layer (carbohydrates), while Gram (-) bacteria have a thinner peptidoglycan layer sandwiched between two lipid layers [33]. Furthermore, examination of PC1 revealed a significant variation among the Gram (+) bacterial species in contrast to the more clustered distribution of Gram (-) species based on their positions along PC2 and PC3. This vast variation made it challenging to use other Raman bands (Table S2) to explain in depth the difference between Gram (+) and Gram (-) (Fig. 3). These results suggest that the structure can be even more intricate depending on the phenotypic variation in the data. PCA, in this study, remains an informative tool that gives an initial impression of the existing clusters in the data. Therefore, to improve the discriminating capabilities of PCA, it can be combined with a more powerful supervised classification method, such as FDA, to additionally consider the variation within each group rather just the total variation.

3.1.2. FDA

FDA was performed to maximize the ratio of variability between classes and minimize the ratio of variability within classes. However, before conducting this FDA, it was necessary to perform a step of variable reduction using PCA. This step ensures that the number of variables does not exceed the number of samples, thus avoiding the impossibility of computing the required Mahalanobis distance and overfitting problems [18, 34].

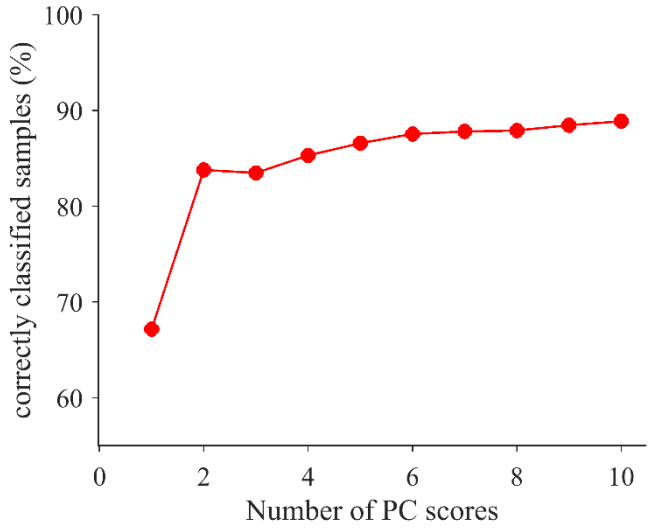


Fig. 4 Selecting the optimal number of PCs to determine the Gram group of the bacterial species

After the generation of orthogonal eigenvectors and the corresponding sample scores for each level of classification, the PCA scores (PCs) were incorporated into the FDA model in such a way as to maximize its discriminant ability. For each model, the optimum number of scores was selected [35]. For instance, Fig. 4 shows the weight of each successive PC on the percentage of samples correctly classified by FDA at the Gram classification level for the dataset of bacterial species. The percentage of correct classification steadily increased with the number of PCs before reaching a plateau. This means that additional PCs would not offer any improvement and the optimal number of PCs had been detected (Fig. 4 & Table 1). All models showed a medium classification performance with a calibration percentage ranging from 75.65 to 93.21 and a validation percentage ranging from 72.06 to 81.01. The best classification of bacterial species using PCA / FDA was obtained at the Gram level. The correct classification rates were 91.46% for calibration and 91.01% for validation. This result mimics those of the PCA and indicates that the best discrimination was obtained with the first two discriminant scores (PC1 and PC2). This moderate performance may be due to the

complexity of the data under study as in Raman spectra or to the similarities between bacterial species belonging to different/same families and genera. Also, the features extracted by PCA were not sufficiently suitable or informative for the FDA, leading to only medium classification accuracy.

Table 1 Capabilities of FDA and KNN to classify microorganisms using the Raman database

			FDA		KNN	
	Level of Classification	**Number of groups**	**Calibration (%)**	**Validation (%)**	**Calibration (%)**	**Validation (%)**
	Gram	2	91.41 (6*)	91.01	99.69	99.28
Gram (+)	Family	10	81.96 (9)	77.99	98.28	94.85
	Genus	15	88.75 (6)	76.98	99.22	98.37
	Species	25	88.77 (6)	77.51	97.05	95.57
Gram (-)	Family	12	83.22(8)	81.5	99.38	98.23
	Genus	24	82.49 (10)	79.07	98.03	95.82
	Species	27	87.26 (10)	81.01	95.92	90.05

*Optimum number of PCA scores allowed to enter in the FDA model

3.1.3. KNN

Machine learning methods such as KNN were also applied to improve the bacterial classification. In the same way as FDA, KNN was used at each level of bacterial classification. Table 1 shows the great improvement offered by KNN compared with FDA. As mentioned above, the calibration and validation percentages increased for all levels of classification.

The KNN model with the highest accuracy corresponded to the Gram level of identification. This model, therefore, allows us to detect the Gram family of the species. This means that if the bacterial species detected by the model is Gram (-), the other 25 Gram (+) species are eliminated. Based on the latter result, a list of families is then proposed, and another KNN model is applied (at the family level). This model had an accuracy of 95–98%. If the family

determined by the KNN identifies the strain in question, the strain is confirmed. Otherwise, we would apply the genus and species KNN models to decide. The KNN genus model also showed an excellent classification performance with a validation percentage of 95.82% for Gram (-) and 98.37% for Gram (+) strains.

However, when KNN was applied at the species level, the classification accuracy only reached about 90%. This was attributed to the exceptionally high level of resemblance among strains, especially in the case of Gram (-) ones. This level-by-level recognition process would allow us to identify the Gram classification, family, genus, and species of the strain with an excellent classification performance, always greater than 90% and mostly higher than 95%.

3.2 Strategy 2 – Direct bacterial classification

The second strategy, aided by CNN, aims to classify bacteria at the species level. CNN achieved the highest performance among the tested methods, as shown in Table 2, with a calibration accuracy of 99.71%, a validation accuracy of 98.93%, and a prediction accuracy of 97.95%.

Table 2 Recognition accuracy of microbes using Raman spectra with CNN

	CNN			
Level of Classification	**Number of species**	**Calibration (%)**	**Validation (%)**	**Prediction (%)**
Species	52	99.71	98.93	97.95

To verify the performance of the model, the receiving operating characteristic (ROC), accuracy-epoch, and loss-epoch curves were introduced (Fig. 5). The CNN model achieved an accuracy above 95% after 33 epochs and the training loss fell below 0.5 after 26 epochs.

As for the ROC, the area under the curve (AUC) was 0.99. Thus, the CNN model achieved a high performance at classifying the 52 bacterial species through Raman spectra.

Meanwhile, the confusion matrix provided the classification details of CNN, as shown in Fig. 6. This matrix can help evaluate the performance of CNN for every class of bacteria and find the types of bacteria where the model has the weakest recognition capabilities. The diagonal of

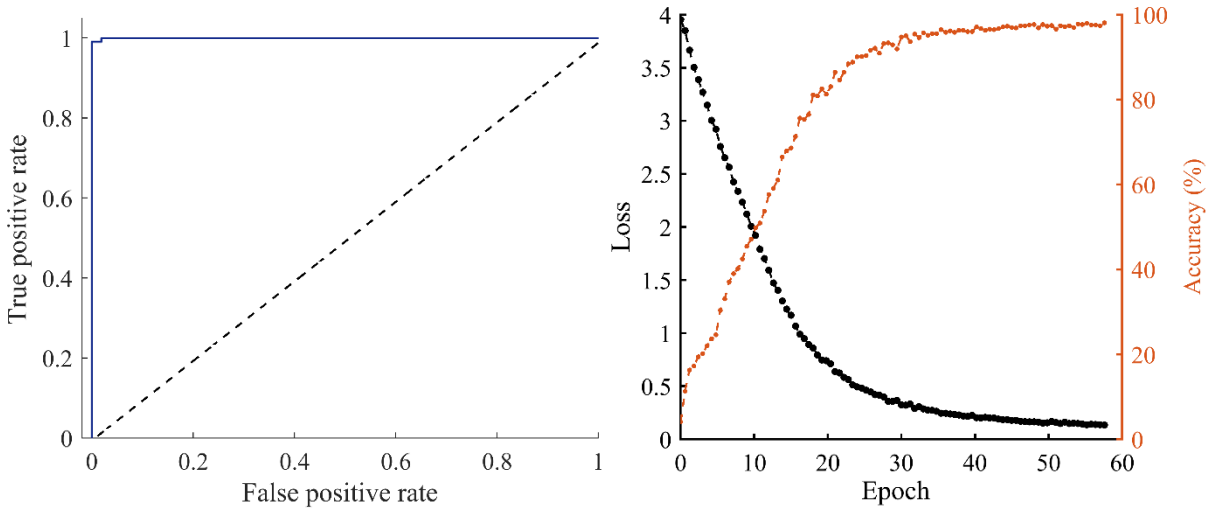


Fig. 5 Receiving operating characteristic (ROC), accuracy and loss curves of CNN

the confusion matrix represents the percentage of correctly predicted species, while the off diagonals show the percentage of misprediction for each species. For example, the model in question has the lowest accuracy for *Pseudomonas fluorecens*. This bacterium was misclassified as *Aeromonas bestiarum* (30% of cases), *Leclercia adecarboxylata*, and *Leuconostoc mesnteroides*. Another misclassification (30%) was that of the *Photobacterium leiognathi* with *Salmonella enterica*. These two bacteria share the same Gram family (Gram -) which might lead to misprediction by the model. Fig. 6 and Fig. S2 also show the misprediction of *Salmonella enterica* with *Escherichia coli*. This is due to the DNA similarity between the two species as this can reach up to 90% [36]. All remaining species were highly predictable by the model with accuracy ranging from 90% to 100%.

3.3 The debate (Strategy I vs Strategy II)

Both strategies seem to handle bacterial classification well. The leads of this debate, KNN and CNN, achieved a classification accuracy higher than 90 %. However, selecting the best tool requires consideration of other factors as each chemometric tool has its own advantages and disadvantages. KNN is a simple technique with an easy and effective algorithm that does not require training time. However, choosing the value of k involves making an expensive estimate. In contrast, CNN has high classification accuracy but requires a great deal of training data and has a high computational cost [37]. Numerous studies have investigated the use of different machine learning algorithms for bacterial classification, and have found both KNN and CNN to be effective methods. For instance, H. Tang et al. [14] found that both KNN and CNN achieved high accuracy in classifying bacterial species using Raman spectroscopy data, while Uysal et al. [38] found that KNN had a classification accuracy of 97.8% for bacterial classification using Raman spectroscopy. Ho et al. [9] also reported that CNN performed well, with an identification accuracy of 97 %, in classifying bacterial species.

As for FDA/PCA, this model did not have an adequate classification performance, except at the Gram level. The database comprising 52 different outputs showed a high level of difficulty with in-class learning for this model, even though the bacterial species were grouped to ease the classification process. However, the discriminating ability of FDA/PCA can be improved by variable selection Nonetheless, this feature was avoided in our study as it included a neural network. To maximize the neural network's potential, it is necessary to maintain a large number of variables in the initial data, and we wanted to keep the possibility of comparing the two strategies by having the same variables at the start.

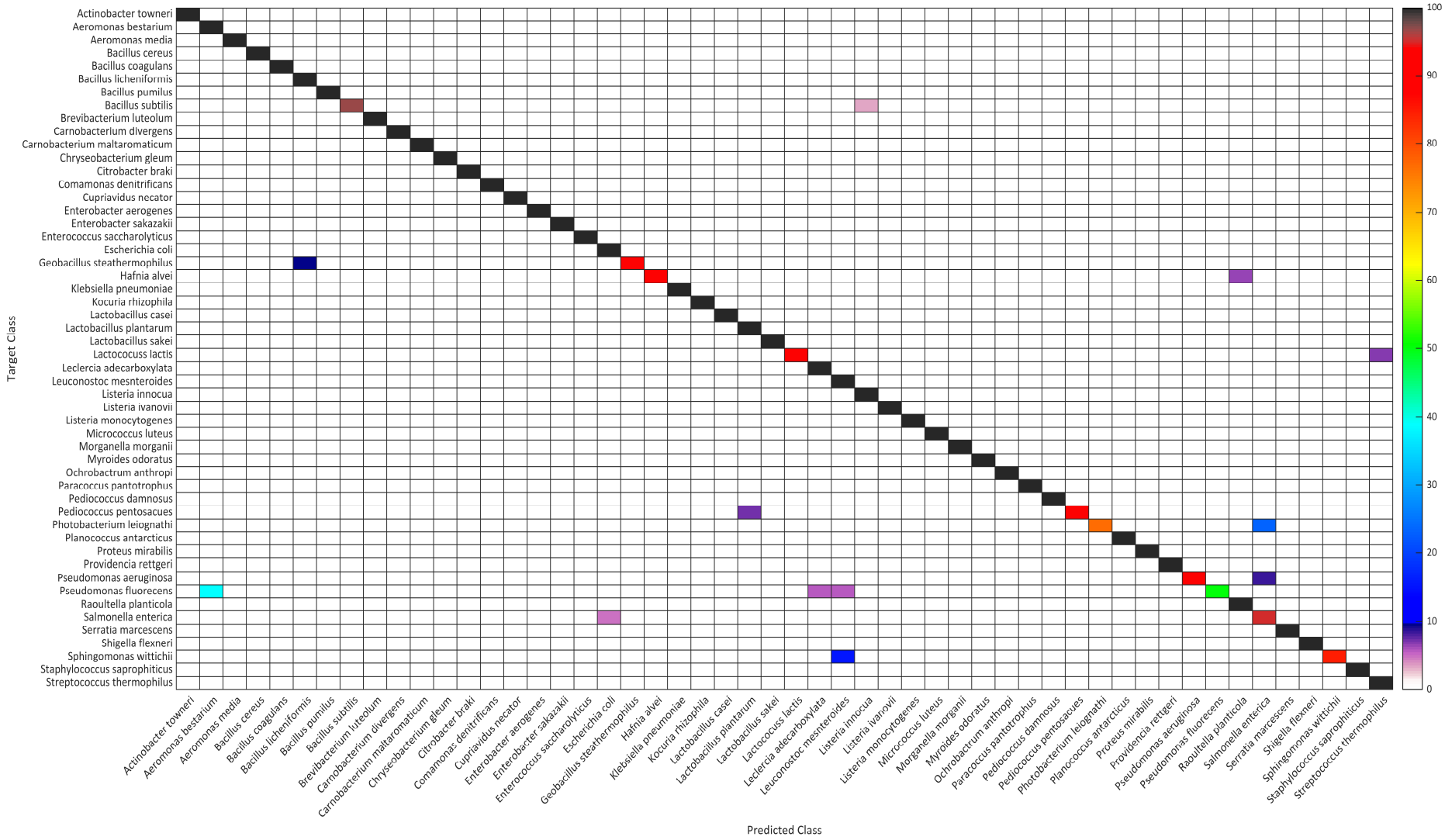


Fig. 6 Confusion matrix of the developed neural network in the prediction set (see Fig. S2 for a closer look at the mispredictions)

4. Conclusion

The chemometric tools implemented in this study showed a very good successful classification rate when applying both strategies. Despite FDA/PCA exhibiting a medium classification performance, it proved to be exceptional in bacterial classification at the Gram level. Additionally, our results show that CNN and KNN provided the best classification models. However, the choice of tool and its usage mainly depends on the end goal of the user. Although the neural approach once again showed superiority over the factorial approach, particularly when the number of classes is higher than three or four, KNN and other nonparametric techniques that rely on distance calculations, can compete rather well with neural networks. Many chemometric tools are worth exploring for Raman analysis and combining these tools with Raman spectroscopy, could enable more complex applications, such as analyzing bacterial pools in food samples. Overall, our study provides valuable insights into the use of chemometric tools for bacterial classification and highlights the potential of Raman spectroscopy as a powerful analytical technique in the field of microbiology.

5. Acknowledgments

The authors would like to thank the European Community, the European funds for regional development (FEDER), Eurofins Alimentaire France, the Region of Pays de la Loire, the University of Nantes, and the Paris Institute for Life, Food and Environmental Sciences (AgroParisTech) for their contributions to completing this work.

6. Funding

This study was supported by OSEO-ISI research grant program and the AgriFoodGPS project.

7. Conflict of interest

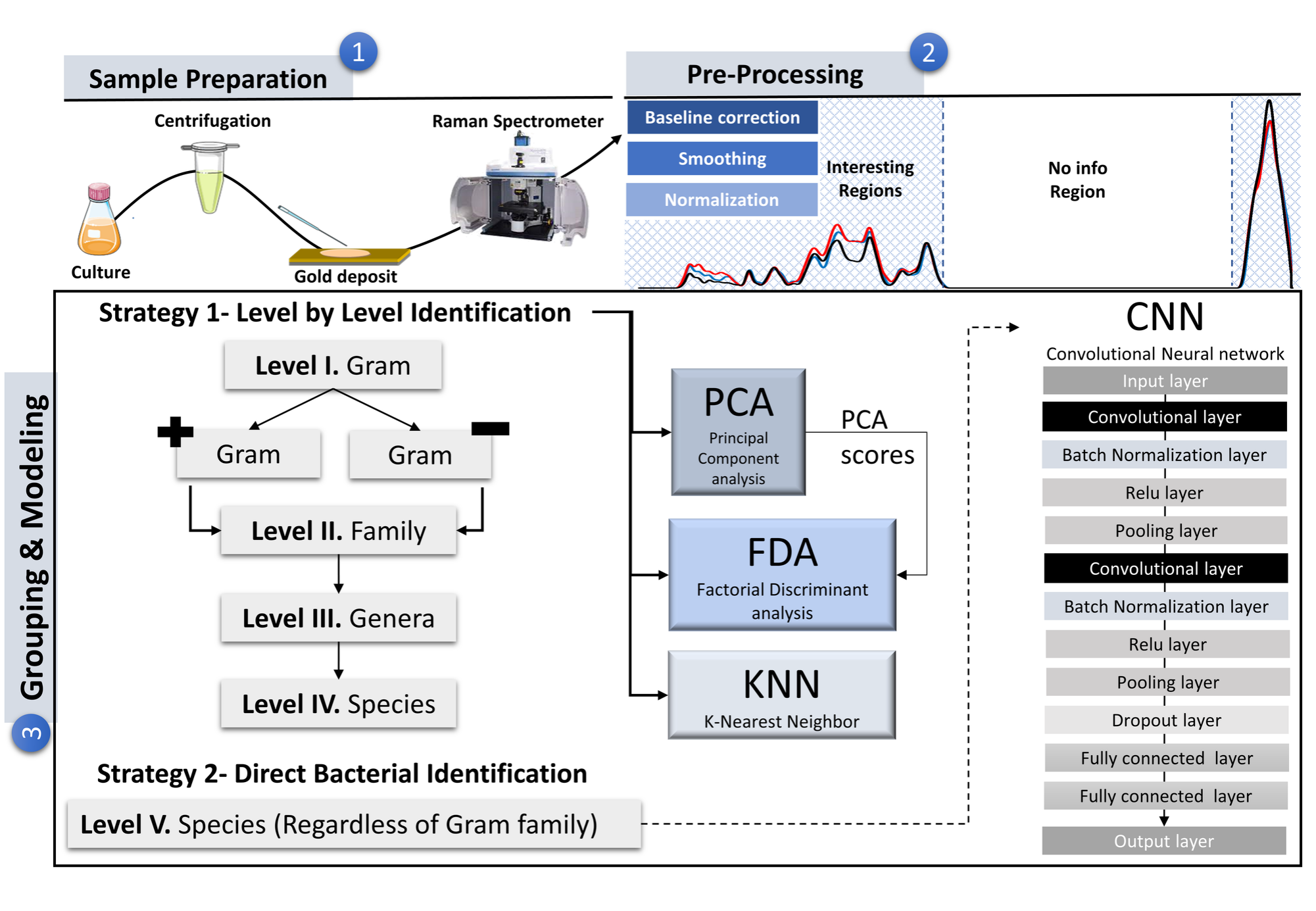
The authors declare that there are no conflicts of interest.

8. References

1. WHO. WHO estimates of the global burden of foodborne diseases 2022. https://www.who.int/publications/i/item/9789241565165. Accessed 11 Jan 2022.
2. Scharff RL. Food Attribution and Economic Cost Estimates for Meat- and Poultry-Related Illnesses. J Food Prot. 2020;83(6):959-67.
3. Zhu L, He J, Cao X, Huang K, Luo Y, Xu W. Development of a double-antibody sandwich ELISA for rapid detection of Bacillus Cereus in food. Sci. Rep. 2016;6(1):16092. https://doi.org/10.1038/srep16092
4. Srimongkol G, Ditmangklo B, Choopara I, Thaniyavarn J, Dean D, Kokpol S, et al. Rapid colorimetric loop-mediated isothermal amplification for hypersensitive point-of-care Staphylococcus aureus enterotoxin A gene detection in milk and pork products. Sci. Rep. 2020;10(1):7768. https://doi.org/10.1038/s41598-020-64710-0
5. Law JW, Ab Mutalib NS, Chan KG, Lee LH. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. Front Microbiol. 2014;5:770.
6. Baker MJ, Byrne HJ, Chalmers J, Gardner P, Goodacre R, Henderson A, et al. Clinical applications of infrared and Raman spectroscopy: state of play and future challenges. Analyst. 2018;143(8):1735-57. http://dx.doi.org/10.1039/C7AN01871A
7. Li Y-S, Church JS. Raman spectroscopy in the analysis of food and pharmaceutical nanomaterials. J Food Drug Anal. 2014;22(1):29-48. https://www.sciencedirect.com/science/article/pii/S1021949814000040
8. Yan S, Wang S, Qiu J, Li M, Li D, Xu D, et al. Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. Talanta. 2021;226:122195. https://www.sciencedirect.com/science/article/pii/S0039914021001168
9. Ho C-S, Jean N, Hogan CA, Blackmon L, Jeffrey SS, Holodniy M, et al. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. Nat. Commun. 2019;10(1):4927. https://doi.org/10.1038/s41467-019-12898-9
10. Sohn WB, Lee SY, Kim S. Single-layer multiple-kernel-based convolutional neural network for biological Raman spectral analysis. J. Raman Spectrosc. 2020;51(3):414-21. https://doi.org/10.1002/jrs.5804
11. Senger RS, Scherr D. Resolving complex phenotypes with Raman spectroscopy and chemometrics. Curr. Opin. Biotechnol. 2020;66:277-82. https://www.sciencedirect.com/science/article/pii/S0958166920301294
12. Lussier F, Thibault V, Charron B, Wallace GQ, Masson J-F. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. TrAC. 2020;124:115796. https://www.sciencedirect.com/science/article/pii/S0165993619305783
13. Li Z, Li Z, Chen Q, Ramos A, Zhang J, Boudreaux JP, et al. Detection of pancreatic cancer by convolutional-neural-network-assisted spontaneous Raman spectroscopy with critical feature visualization. Neural Networks. 2021;144:455-64. https://www.sciencedirect.com/science/article/pii/S0893608021003567
14. Tang J-W, Liu Q-H, Yin X-C, Pan Y-C, Wen P-B, Liu X, et al. Comparative Analysis of Machine Learning Algorithms on Surface Enhanced Raman Spectra of Clinical Staphylococcus Species. Front Microbiol. 2021;12. https://www.frontiersin.org/article/10.3389/fmicb.2021.696921

15. Assaf A, Cordella CBY, Thouand G. Raman spectroscopy applied to the horizontal methods ISO 6579:2002 to identify Salmonella spp. in the food industry. Anal. Bioanal. Chem. 2014;406(20):4899-910. https://doi.org/10.1007/s00216-014-7909-2
16. Lancelot E, Fontaine J, Grua-Priol J, Assaf A, Thouand G, Le-Bail A. Study of structural changes of gluten proteins during bread dough mixing by Raman spectroscopy. Food Chem. 2021;358:129916. https://www.sciencedirect.com/science/article/pii/S0308814621009225
17. Kanso, M., Cuenot, S. & Louarn, G. Sensitivity of Optical Fiber Sensor Based on Surface Plasmon Resonance: Modeling and Experiments. Plasmonics. 2008 (3), 49-57, doi:10.1007/s11468-008-9055-1.
18. Bertrand D, Courcoux P, Autran J-C, Meritan R, Robert P. Stepwise canonical discriminant analysis of continuous digitalized signals: Application to chromatograms of wheat proteins. J. Chemom. 1990;4(6):413-27. https://doi.org/10.1002/cem.1180040605
19. Abdi H, Williams LJ. Principal component analysis. WIREs Computational Statistics. 2010;2(4):433-59. https://doi.org/10.1002/wics.101
20. Cordella C. PCA : The basic building block of chemometrics. In: Ira SK, editor. Analytical Chemistry: IntechOpen; 2012. p. 146 p.
21. Cordella CBY, Bertrand D. SAISIR: a new general chemometric toolbox. TrAC. 2014;54.
22. Fix E, Hodges JL. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. International Statistical Review / Revue Internationale de Statistique. 1989;57(3):238-47. http://www.jstor.org/stable/1403797
23. Coomans D, Massart DL. Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules. Anal. Chim. Acta. 1982;136:15-27. https://www.sciencedirect.com/science/article/pii/S0003267001953590
24. Kim KS, Choi HH, Moon CS, Mun CW. Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. Current Applied Physics. 2011;11(3):740-5. https://www.sciencedirect.com/science/article/pii/S1567173910004153
25. Yu D, Deng L. Deep Learning and Its Applications to Signal and Information Processing [Exploratory DSP]. IEEE Signal Processing Magazine. 2011;28:145-54.
26. Dimmita N, Siddaiah P. Speech Recognition Using Convolutional Neural Networks. J. Eng. Technol (UAE). 2018;7:133-7.
27. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discovery Today. 2018;23(6):1241-50. https://www.sciencedirect.com/science/article/pii/S1359644617303598
28. Traore BB, Kamsu-Foguem B, Tangara F. Deep convolution neural network for image recognition. Ecological Informatics. 2018;48:257-68. https://www.sciencedirect.com/science/article/pii/S1574954118302140
29. Qi H, editor Derivation of Backpropagation in Convolutional Neural Network (CNN)2016.
30. Weng S, Yuan H, Zhang X, Li P, Zheng L, Zhao J, et al. Deep learning networks for the recognition and quantitation of surface-enhanced Raman spectroscopy. Analyst. 2020;145(14):4827-35. http://dx.doi.org/10.1039/D0AN00492H

31. Maquelin, K.; Kirschner, C.; Choo-Smith, L. P.; van den Braak, N.; Endtz, H. P.; Naumann, D.; Puppels, G. J., Identification of medically relevant microorganisms by vibrational spectroscopy. Journal of Microbiological Methods. 2002b; 51 (3), 255-271.
32. Schuster, K. C.; Urlaub, E.; Gapes, J. R., Single-cell analysis of bacteria by Raman microscopy: spectral information on the chemical composition of cells and on the heterogeneity in a culture. Journal of Microbiological Methods. 2000a; 42 (1), 29-38.
33. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. Cold Spring Harbor perspectives in biology 2. 2010. doi:10.1101/cshperspect.a000414.
34. Lu G-F, Zheng W. Complexity-reduced implementations of complete and null-space-based linear discriminant analysis. Neural Networks. 2013;46:165-71. https://www.sciencedirect.com/science/article/pii/S0893608013001494
35. Monakhova YB, Godelmann R, Kuballa T, Mushtakova SP, Rutledge DN. Independent components analysis to increase efficiency of discriminant analysis methods (FDA and LDA): Application to NMR fingerprinting of wine. Talanta. 2015;141:60-5. https://www.sciencedirect.com/science/article/pii/S0039914015001903
36. Winfield MD, Groisman EA. Phenotypic differences between Salmonella and Escherichia coli resulting from the disparate regulation of homologous genes. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(49):17162-7. https://www.pnas.org/content/pnas/101/49/17162.full.pdf
37. Makkar T, Kumar Y, Dubey AK, Á R, Goyal A, editors. Analogizing time complexity of KNN and CNN in recognizing handwritten digits. 2017 Fourth International Conference on Image Information Processing (ICIIP); 2017 21-23 Dec. 2017.
38. Uysal Ciloglu F, Saridag AM, Kilic IH, Tokmakci M, Kahraman M, Aydin O. Identification of methicillin-resistant Staphylococcus aureus bacteria using surface-enhanced Raman spectroscopy and machine learning techniques. Analyst. 2020;145(23):7559-70. http://dx.doi.org/10.1039/D0AN00476F


1
Sample Preparation
Centrifugation
Raman Spectrometer
Culture
Gold deposit
2
Pre-Processing
Baseline correction
Smoothing
Normalization
Interesting Regions
No info Region
3
Grouping & Modeling
Strategy 1- Level by Level Identification
Level I. Gram
Gram
Gram
Level II. Family
Level III. Genera
Level IV. Species
PCA
Principal Component analysis
PCA scores
FDA
Factorial Discriminant analysis
KNN
K-Nearest Neighbor
Strategy 2- Direct Bacterial Identification
Level V. Species (Regardless of Gram family)
CNN
Convolutional Neural network
Input layer
Convolutional layer
Batch Normalization layer
Relu layer
Pooling layer
Convolutional layer
Batch Normalization layer
Relu layer
Pooling layer
Dropout layer
Fully connected layer
Fully connected layer
Output layer