LWT - Food Science and Technology 174 (2023) 114449

Contents lists available at ScienceDirect

LWT

journal homepage: www.elsevier.com/locate/lwt

Prediction of growth/no growth status of previously unseen bacterial strain using Raman spectroscopy and machine learning

Takashi Yamamoto [a,b], J. Nicholas Taylor [c], Shige Koseki [a], Kento Koyama [a,*]

[a] Graduate School of Agricultural Science, Hokkaido University, Kita-9, Nishi-9, Kita-ku, Sapporo, 060-8589, Japan
[b] Food Science Labs, Functional Food Division, Nippon Shinyaku Co., Ltd., 14, Nishinosho-Monguchi-cho, Kisshoin, Minami-ku, Kyoto, 601-8550, Japan
[c] Research Institute for Electronic Science, Hokkaido University, Kita-20, Nishi-1-0, Kita-ku, Sapporo, 001-0020, Japan

ARTICLE INFO

Keywords:
Spoilage bacteria
Growth/no growth boundary model
Neural network
Linear discriminant analysis

ABSTRACT

We develop a method to predict the growth/no growth response of previously unseen bacteria, using Raman spectral features and a machine-learning model. Twenty-one strains of bacteria were isolated from seven commercially available fresh-cut vegetables. Twenty Raman spectra of single cells were acquired for each isolated strain. The growth/no growth responses of each strain in a liquid medium were evaluated with two levels of sodium acetate concentrations, two incubation temperatures, and eight sampling times using optical density to confirm growth limit conditions. The Raman spectra of 20 strains and their forty-eight growth/no growth responses were used to train an artificial neural network model to predict the growth/no growth of unknown bacteria. Our model predicted the growth/no growth of 21 unknown bacteria with an overall accuracy of 90%. Such rapid characterization of unknown bacterial growth using Raman spectroscopy will be valuable for the food manufacturing industry, where unknown bacteria are often encountered.

1. Introduction

Microbial control is used to ensure food preservation and continuity of production. A great variety of bacteria across many microbial genera are found on food and are involved in food spoilage or food poisoning (Jay et al., 2005; Rawat, 2015). Predicting the growth behavior of target microorganisms is useful for determining environmental conditions for food safety and quality assurance during processing, distribution, and storage (Stavropoulou & Bezirtzoglou, 2019). Bacterial responses over time have been investigated using various combinations of environmental factors in predictive microbiology (Ross & McMeekin, 1994). Statistical models have been developed to predict the target bacterial population behavior in various food environmental conditions across varying temperature, pH, and water activity (Walls & Scott, 1997). These prediction models enable the quantitative evaluation of target bacterial population behavior in various types of foods for decision making for setting processing and storage conditions. Thus, research to predict the target pathogenic and spoilage bacterial behavior is important for protecting food safety and quality.

Some representative strains have been studied to predict bacterial population behavior. However, different strains of the same bacterial species can have different sensitivities to factors related to population growth behavior (Aryani et al., 2015; Dengremont & Membre, 1995; Lianou & Koutsoumanis, 2011, 2013). Even if the genus and species of unknown bacteria in food products were identified, bacterial stress responses are usually predicted based on a similar strain or closely related species, as reported in previous studies. The prediction of growth based on similar information may cause problems in terms of accuracy and reliability because there is a lack of objective links between target strains or species. Thus, to predict the population behavior at the strain level, efforts are needed to link the stress tolerance of each strain.

Recently, Raman spectroscopy has attracted attention for the identification of bacteria in the medical and food safety fields (Yan et al., 2021). Raman spectroscopy is capable of rapid and nondestructive measurement of molecular vibration (Jaafreh et al., 2019) and provides biological information at the molecular level (Moreira et al., 2008). Indeed, it is possible to classify bacterial genera or species using Raman spectroscopy (Huayhongthong et al., 2019; Yan et al., 2021). In addition, Raman spectroscopy and chemometrics can be useful for identifying the stress tolerance of spoilage bacteria to food additives (Yamamoto et al., 2021) and the characteristics of antimicrobial resistance (Germond et al., 2018). Since Raman spectra and chemometric analysis objectively capture information about the molecules and can classify bacterial resistance to diverse stresses, such as antibiotics and

* Corresponding author.
E-mail address: kkoyama@agr.hokudai.ac.jp (K. Koyama).

https://doi.org/10.1016/j.lwt.2023.114449
Received 25 August 2022; Received in revised form 20 December 2022; Accepted 6 January 2023
Available online 7 January 2023

food additives, the relationship between bacterial spectral fingerprints and known bacterial population behavior may be applied to predict unknown bacterial population growth capability.

The purpose of this study was to predict the growth behavior of unknown microorganisms in a culture medium using Raman spectroscopy and machine learning. Bacteria in freshly-cut vegetables were investigated because cut vegetables can be easily contaminated at all stages of production (Lehto et al., 2011). First, 21 strains across nine genera were isolated from seven freshly-cut vegetable samples, which were then measured to obtain the Raman spectra. Second, the growth of the isolated bacteria in a liquid medium was evaluated at various sodium acetate concentrations, incubation temperatures, and sampling times by optical density measurement. Finally, we predicted the growth probability of unknown bacteria based on the sodium acetate concentration, temperature, incubation time, and Raman spectral features using machine learning. Prediction of unknown bacterial behavior using a simple procedure would help the food industry quickly establish food processing and storage conditions.

2. Materials and methods

2.1. Isolation of bacterial strains from fresh-cut vegetables

The target of the bacteria is isolated from the food since there is a large variety of strain. We collected seven kinds of freshly-cut vegetables, including various ingredients, at supermarkets and convenience stores in Kyoto, Japan. The details of the freshly-cut vegetables are described in Supplementary Table 1. To isolate bacteria from the freshly-cut vegetable samples, a 25 g portion of fresh-cut vegetables was homogenized in 225 g of phosphate-buffered saline (PBS) for 3 min using a homogenizer (Pro-media SH-IIM, ELMEX, Japan) to obtain a suspension. The suspension was appropriately diluted with PBS and 1 mL of each suspension was transferred to sterile Petri dishes mixed with standard plate count agar (SPC agar, NISSUI, Japan). Once the medium had solidified, it was incubated for 48 h at 35 °C under aerobic conditions. The major colonies obtained from incubated SPC agar were randomly isolated and placed on SPC agar using a loop. In this study, three strains were isolated from each sample of the cut vegetables. Twenty-one strains were identified, and the isolated colonies were stored at 5 °C.

2.2. Identification of the isolated bacteria strains

Isolated bacteria were identified according to the method described in the 17th edition of the Japanese Pharmacopoeia (http://www.mhlw.go.jp/topics/bukyoku/iyaku/yakkyoku/english.html). Briefly, the genomic DNA of the isolated bacteria was extracted from the culture with PrepManTM Ultra Sample Preparation Reagent (Life Technologies, USA) and used directly as a PCR template to amplify the divergent region of the 16S rRNA gene using 10 F primer (5′-GTTTGATCCTGGCTCA-3′) and 800 R primer (5′-TACCAGGGTATCTAATCC-3′). PCR products (approximately 740 bp) were purified using ExoSAP-IT (Affymetrix part of Thermo Fisher Scientific, MA, USA) for sequencing. Sequencing reactions were performed in a Bio-Rad DNA Engine Dyad PTC-220 Peltier Thermal Cycler, using the BigDye™ Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific, MA, USA), according to the manufacturer's instructions. Single-pass sequencing was performed on each template using a 10 F primer. The fluorescently labeled fragments were purified from the unincorporated terminators either by the ethanol precipitation method or using the BigDye Terminator™ Purification Kit (Thermo Fisher Scientific, MA, USA). The samples were analyzed using a 3730xl DNA analyzer (Thermo Fisher Scientific, MA, USA).

The 16S rRNA gene sequences obtained were compared with those of the type strains available in the EzBioCloud database (https://www.ezbiocloud.net/) (Yoon et al., 2017) to determine the approximate phylogenetic affiliation of each strain. The evolutionary history was inferred using the maximum likelihood method and the Kimura 2-parameter model (Kimura, 1980). Initial trees for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the maximum composite likelihood approach and then selecting the topology with a superior log likelihood value. A discrete gamma distribution was used to model evolutionary rate differences among the sites (five categories ($+G$, parameter = 0.6550)). The analysis involved 35 nucleotide sequences. All positions containing gaps and missing data were eliminated. The final dataset contained 655 positions. Evolutionary analyses were conducted using MEGA X (Kumar et al., 2018). The non-rooted phylogenetic trees were visualized using an online website the Interactive Tree of Life website (https://itol.embl.de) (Letunic & Bork, 2021).

2.3. Sample preparation and Raman spectra measurements

In subsequent experiments, the isolated bacteria in the stock were transferred to SPC agar by platinum loop and then used after culturing for 48 h at 35 °C under aerobic conditions. The Raman spectrum of a single cell was collected using a laser Raman microscope (RAMAN touch, Nanophoton, Osaka, Japan). The excitation source was a 532 nm laser operated at 10 mW. A 20 × /0.45 objective lens (Nikon TU Plan Fluor, Nikon) with a laser spot size of approximately 720 nm was used to focus the excitation light on the sample. Raman spectra were acquired using a 300 lines/mm grating for 30 s. Two spectra from a single spot were averaged to obtain the mean spectra. The Raman shift was calibrated using silicon (520 cm^{-1}) before the spectra were acquired. Cultured cells were suspended in 50 μL of pure water using a 1-μL loop. Aliquots (1 μL) of the suspension were dropped onto a stainless steel piece (SUS430, HIKARI, Osaka, Japan), which was air-dried before starting the Raman measurement. Twenty averaged Raman spectra were obtained for each strain using a range of Raman shifts from 125 to 4690 cm^{-1} with increments of 5.0 cm^{-1}.

2.4. Bacterial growth experiment

In this study, we aimed to predict the growth/no growth of unknown strains in a certain culture medium using Raman spectroscopy and machine learning. Therefore, we decided to predict the growth/no growth of unknown strains under different combinations of sodium acetate, a food additive, and low temperatures. To construct the prediction model, we first investigated the growth/no growth of bacteria under the following conditions using optical density measurements. The conditions of this experiment were two incubation temperatures (5 and 10 °C), three sodium acetate concentrations (0, 0.25, and 0.50% w/v), and eight incubation time periods (0, 1, 2, 3, 4, 5, 6, and 7 d). Thus, the total number of experimental conditions used was 48.

The pH of tryptic soy broth (TSB) was adjusted to pH 6.5 using 1 M hydrochloric acid. The cultured cells were suspended in TSB using a 1-μL loop and diluted with TSB to a bacterial concentration of 10^4 CFU mL^{-1} in TSB at the start of the culture. Aliquots (100 μL) of the suspension were dispensed into wells of 96-well flat-bottom culture plates (Corning 3595 96-well cell culture plate, Corning, NY, USA) with three replicates for each condition. The optical density (absorbance at 600 nm, OD_{600}) of each well was measured using a microplate reader SH-9000 Lab (Corona Electric, Ibaraki, Japan) at the beginning of the experiment and every other day thereafter. Three iterations of the experiment were performed according to previous literature (Kuroda et al., 2019; Presser et al., 1998; Rodríguez-Saavedra et al., 2021). In this study, the threshold OD_{600} value was set to 0.1 (Yamamoto et al., 2021). If the OD_{600} value was less than 0.1, the population was considered to have no growth. If the OD_{600} value was $\geq$0.1, the bacterial population was considered to be growing.

2.5. Feature extraction from Raman spectra of bacteria

The preprocessing procedure consisted of four parts: removal of

cosmic rays and baseline contamination, smoothing, and standardization. First, all spectra with cosmic spikes were identified by visual inspection, and the cosmic spikes were removed using Raman spectroscopy software (Raman viewer, Nanophoton). The maximum intensity within a selected range of ±120 cm^{-1} was replaced with the median intensity in that range. Second, concave baseline correction was performed on all spectra using a seventh-degree recursive polynomial fitting algorithm (Lieber & Mahadevan-Jansen, 2003; Taylor et al., 2019). Third, the spectra were smoothed using Savitzky-Golay polynomial filters (polyorder 2, 25 cm^{-1} window length) (Savitzky & Golay, 1964). Fourth, each Raman spectrum was standardized such that the mean intensity was 0 and the variance was 1. Finally, linear discriminant analysis (LDA) was used for data-dimensionality reduction (FISHER, 1938). LDA is a method of maximizing the distance between prespecified classes and thereby finding the axis of dimensionality reduction (Schumacher et al., 2014). In this study, classes were specified for each strain. The Raman spectral data were dimensionally reduced from 914 to 10 using LDA.

2.6. Aggregated hierarchical clustering

Aggregated hierarchical clustering was performed using 10-dimensional LDA values for comparison with a phylogenetic tree based on the 16S rRNA information. Using SciPy (Virtanen et al., 2020) in Python packages, each strain was sorted by hierarchical clustering using mean linkage and Euclidean distance metrics, based on the average of each value from LDA1 to LDA10.

2.7. Model development and evaluation

We constructed models to predict bacterial growth/no growth from the incubation temperature, sodium acetate concentration, incubation time, and Raman spectral features of the bacteria. Independent variables of the classification model used were temperature, concentrations of sodium acetate, incubation time, and ten LDA values (LDA 1 to LDA 10). The dependent variable was the probability of growth in the range of 0–1. For training and evaluation of the classification model, 2880 data points were prepared per strain. The 2880 data points were generated by three replicates of multiplying 48 growth/no growth conditions and 20 measurements of Raman features (LDA values).

The flow from the model development to the prediction of the growth probability is shown in Fig. 1. First, the incubation temperature, concentration of sodium acetate, incubation time, and 1–10 LDA values were standardized in advance. Out of the 21 strains, 20 were used for training data and one was used for test data. The input values for the machine learning model were temperature, sodium acetate concentration, incubation time, and 10 LDA values (LDA 1 to LDA 10). The true label was observed growth/no growth, and a range of 0–1 output were used as the proliferation probability.

The artificial neural network (ANN) classification model consisted of an input layer, one hidden layer, and an output layer. An ANN with a single hidden layer was built using the TensorFlow (Abadi et al., 2016) and Keras (https://keras.io) packages in Python. ANN can approximate complex relationships between explanatory and observed variables without the need to assume the type of relationship or degree of nonlinearity between the various independent and dependent variables (Hajmeer & Basheer, 2002). The output of the model was the growth probability of the bacterial population. We chose ANN with a single hidden layer since the deeper layer model may overfit to the training data (Kuroda et al., 2019).

The hyperparameters of the ANN were determined using the Bayesian optimization implementation provided by Keras Tuner (O'Malley et al., 2019). Hyperparameter determination and model training was performed using the hold-out method, and 25% of the data for each strain in the training data were split into validation data. The number of hidden layer units (10, 15, 20, 25, 30, and 35), dropout rate

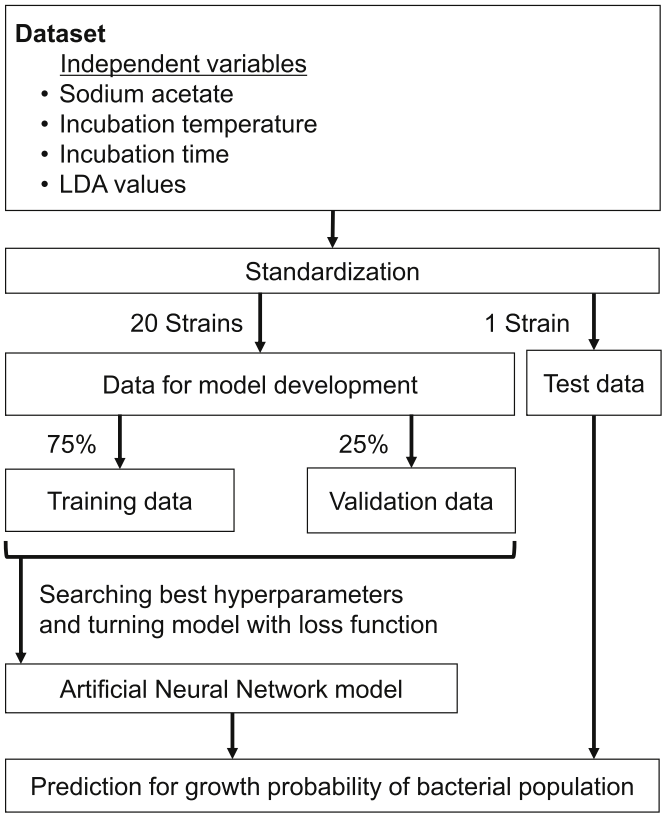

Fig. 1. Flow diagram of growth probability prediction model construction. 21 models were created for each test strain.

(0.1, 0.2, 0.3, 0.4, and 0.5) (Srivastava et al., 2014), and learning rate (0.1, 0.01, 0.001, and 0.0001) were hyperparameters for exploration. The optimizer used was adaptive moment estimation (Adam) (Kingma & Ba, 2015). The hidden layer of the activation function was a rectified linear unit (ReLU) (Nair & Hinton, 2010). A sigmoid activation function was used as the output layer. The batch size was set to one-tenth that of the training data, and the number of epochs was 50. We also added an early stopping function (Raskutti et al., 2014) to prevent overfitting. Learning was set to stop if no loss reduction of the validation data was observed during the five epochs. Binary cross-entropy was used as the loss function (De Boer et al., 2005; Lee & Song, 2019). The hyperparameters were determined each time the test strains were changed to create the model. The hyperparameters determined for each test strain are shown in Supplementary Table 2.

The performance of the constructed models was evaluated using the overall accuracy and area under the curve (AUC) as metrics of the predictive models. The data were classified into two classes (growth/no growth, threshold = 0.5, and growth probability). The overall accuracy was calculated as the percentage of correct responses for all the predictions. The AUC was calculated by plotting the true positive rate against the false positive rate at various thresholds to create a receiver operating characteristic (ROC) curve, taking a value between 0 and 1 in the region under the ROC curve. The AUC is useful for comparing multiple models when the class sampling is imbalanced (Saito & Rehmsmeier, 2015). and is 0.5 for classifiers whose prediction is random, and 1.0, for perfect classifiers (Hanley & McNeil, 1982). If the AUC exceeds 0.9, the model is considered highly accurate (Fischer et al., 2003).

3. Results

3.1. Strains isolated from fresh-cut vegetables and their genotypic identification

A total of 21 strains were isolated from seven types of cut vegetables (samples A to G). Based on the results of the 16S rRNA sequencing in the first half (approximately 720–740 bp) of each strain, a closely related species was estimated (Table 1). Nine genera were isolated from cut vegetables, each with seven strains of *Pseudomonas* sp. (Pseudomonas B1, Pseudomonas B3, Pseudomonas E1, Pseudomonas E2, Pseudomonas E3, Pseudomonas F1, and Pseudomonas F3), six strains (Rahnella A1, Rahnella A2, Rahnella A3, Rahnella G1, Rahnella G2, and Rahnella G3) of *Rahnella* sp., two strains (Erwinia B2 and Erwinia C2) of *Erwinia* sp., and one strain of *Arthrobacter* sp. (Arthrobacter D3), *Kluyvera* sp. (Kluyvera C3), *Lactococcus* sp. (Lactococcus F2), *Leuconostoc* sp. (Leuconostoc C1), *Serratia* sp. (Serratia D1), and *Pantoea* sp. (Pantoea D2). A non-root-based phylogenetic tree of the isolates from fresh vegetables is shown in Fig. 2. The strains were divided into three major groups: *Enterobacterales* (*Erwinia* sp., *Kluyvera* sp., *Pantoea* sp., *Rahnella* sp., and *Serratia* sp.), *Pseudomonas* sp., and Gram-positive bacteria (*Lactococcus* sp., *Leuconostoc* sp., and *Arthrobacter* sp.).

3.2. Extracted features obtained from Raman spectra

Twenty replicates of Raman spectra were obtained for the isolated 21 food-related strains, namely seven strains of *Pseudomonas* sp., six strains of *Rahnella* sp., two strains of *Erwinia* sp., one strain of *Arthrobacter* sp., one strain of *Kluyvera* sp., one strain of *Lactococcus* sp., one strain of *Leuconostoc* sp., one strain of *Serratia* sp., and one strain of *Pantoea* sp. Fig. 3 shows the preprocessed and averaged Raman spectra of each strain and the base substrate of SUS430. Raman spectra were similar to each other, except that Pantoea D2 showed specific spectra in the 1000 cm^{-1} 1200 cm^{-1} and 1500 cm^{-1} 1600 cm^{-1} regions (Fig. 3).

The dimensions of the Raman spectra were reduced using LDA for further analyses. Fig. 4 shows a pair plot created based on the LDA value obtained from the Raman spectrum. The pair plot created using LDA1–LDA5 confirmed a cluster of strains. The cluster of strains is shown separately for *Pantoea* sp. at LDA1 and LDA2-5, *Arthrobacter* sp. at LDA2 and LDA3, and *Pseudomonas* sp. mainly at LDA2 and LDA4–5 (Fig. 4). Fig. 5 shows the dendrograms created based on LDA1 to LDA10. The distances in the Raman information between *Rahnella* sp. (Rahnella A1, Rahnella A2, Rahnella A3, Rahnella G1, Rahnella G2, and Rahnella G3) were close to each other in the 16S rRNA classification (Figs. 2 and 5), but only *Pantoea* sp. (Pantoea D2) was far from the same *Enterobacteriaceae* strains in the Raman information (Figs. 2 and 5).

Table 1
Similarity of 16S rRNA of each isolation retrieved by EzBioCloud.

Cut-Vegetable sample No.	Named Strain	Top-hit taxon	Top-hit strain	Similarity (%)
A	Rahnella A1	*Rahnella aceris*	SAP-19	100.0
	Rahnella A2	*Rahnella aquatilis*	CIP 78.65	99.2
	Rahnella A3	*Rahnella aquatilis*	CIP 78.65	99.3
B	Pseudomonas B1	*Pseudomonas viridiflava*	DSM 6694	100.0
	Erwinia B2	*Erwinia persicina*	NBRC 102418	99.6
	Pseudomonas B3	*Pseudomonas rhodesiae*	CIP 104664	99.7
C	Leuconostoc C1	*Leuconostoc holzapfelii*	BFE 7000	99.2
	Erwinia C2	*Erwinia persicina*	NBRC 102418	99.9
	Kluyvera C3	*Kluyvera intermedia*	NBRC 102594	100.0
D	Serratia D1	*Serratia myotis*	12	99.6
	Pantoea D2	*Pantoea eucalypti*	LMG 24198	99.9
	Arthrobacter D3	*Arthrobacter oryzae*	KV-651	99.6
E	Pseudomonas E1	*Pseudomonas grimontii*	CFML 97-514	100.0
	Pseudomonas E2	*Pseudomonas viridiflava*	DSM 6694	100.0
	Pseudomonas E3	*Pseudomonas kitaguniensis*	MAFF 301498	99.9
F	Pseudomonas F1	*Pseudomonas extremorientalis*	KMM 3447	99.9
	Lactococcus F2	*Lactococcus lactis* subsp. *Cremoris*	NCDO 607	99.9
	Pseudomonas F3	*Pseudomonas extremorientalis*	KMM 3447	100.0
G	Rahnella G1	*Rahnella aquatilis*	CIP 78.65	100.0
	Rahnella G2	*Rahnella aquatilis*	CIP 78.65	99.7
	Rahnella G3	*Rahnella aquatilis*	CIP 78.65	99.3

Coverage of 16S rRNA: 48.9–49.8%.

3.3. Model performance

We evaluated whether the constructed classification model could predict the growth of previously unseen bacteria based on their Raman spectral features at arbitrary incubation times, sodium acetate concentrations, and temperatures. A representative result of growth/no growth response at 10 °C for two *Pseudomonas* spp. is shown in Fig. 6. Pseudomonas, E2, and F1 were different species (Table 1). Both strains exhibited different growth/no growth responses, especially tolerance to sodium acetate. Even though species variability was observed in the examined *Pseudomonas* spp., the developed classification model was able to express the growth/no growth boundary for each strain (Fig. 6). A representative result of growth/no growth response at 10 °C for Leuconostoc C1 and Pantoea D2 are shown in Fig. 7. Only Leuconostoc C1 and Pantoea D2 genera were contained in the test data. The results of predicting the growth probability of Leuconostoc C1 showed a growth/no-growth boundary (Fig. 7a). By contrast, the developed model could not predict the growth characteristics of Pantoea D2 (Fig. 7b). The developed model enabled the prediction of growth/no growth boundary of unknown species and unknown genera, except for Pantoea D2, which had the longest distance in the dendrogram constructed with Raman spectra (Fig. 5).

Table 2 summarizes the results of the overall accuracy and AUC of growth/no growth prediction for unknown bacteria using Raman spectral features. The mean overall accuracy and AUC for all the test data were 0.90 and 0.91. Each bacterium was treated as an unknown bacterium, and the growth/no growth of unknown bacteria was predicted. In some cases, the training set and predicted data contained bacteria of the same genus, whereas in other cases, the training data contained bacteria that were not in the predicted data. The mean Overall accuracy and AUC for the unknown levels of the test data were 0.86 and 0.83 for the unknown genera, 0.92 and 0.92 for the unknown species, and 0.92 and 0.94 for the unknown strains. The strain with the highest evaluation was Pseudomonas E2, with an overall accuracy of 1.00 and an AUC of 1.00. The lowest evaluation strain was Pantoea D2, with an overall accuracy of 0.77 and an AUC of 0.50. The mean value of the overall accuracy and AUC for unknown bacteria, except for Pantoea D2, was approximately 0.9. Strains other than Pantoea D2, whose distance in the dendrogram of Raman features was less than approximately 25 (Fig. 5), were predicted with a certain overall accuracy.

4. Discussion

In this study, we developed and validated a machine learning model to predict the growth/no growth of unknown bacteria isolated from

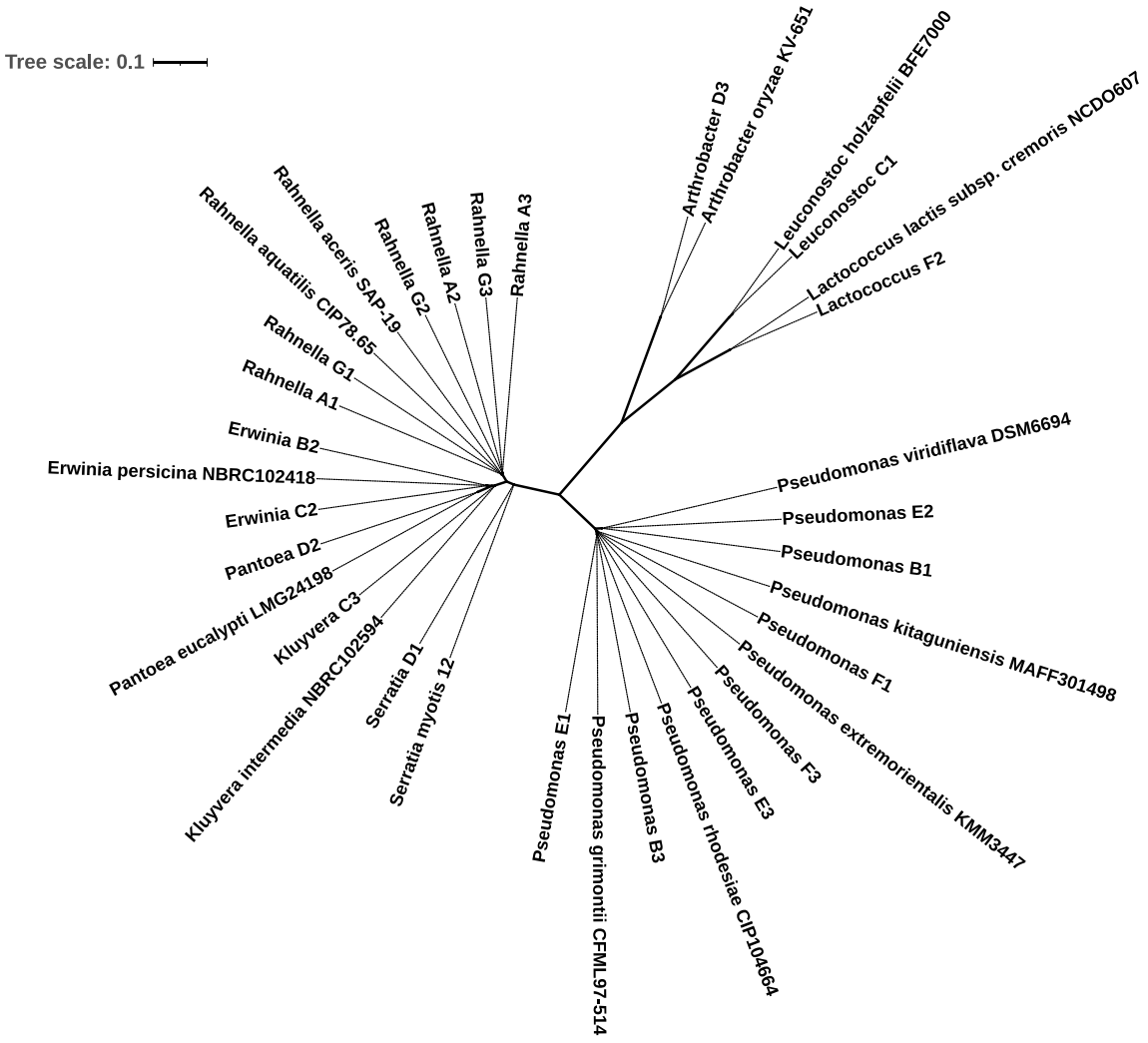

Fig. 2. Non-rooted phylogenetic trees of the isolates from the fresh-cut vegetable including the type strains are based on the partial 16S rRNA gene sequence. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site.

commercially available freshly cut vegetables using Raman spectroscopy (Figs. 6 and 7). The model predicted growth/no growth of unknown bacteria with an average overall accuracy of 90% (Table 2). Other currently used growth/no growth models with the function of environmental factors are limited to the prediction of only known bacteria in the training set because there is no link among strains. By contrast, in the present study, Raman spectra served to provide a link between unknown and known bacteria. Therefore, the combination of Raman spectroscopy and machine learning demonstrated the potential to rapidly provide the growth behavior of bacteria that were not used in the training of the model.

Raman spectra with chemometrics have been used to classify microorganisms by species and phylogeny (Klein et al., 2019; Meisel et al., 2014; Yan et al., 2021). Based on previous reports, we assumed that the processing of Raman spectra would enable the creation of clusters in terms of biological taxonomy. The dimensionality of the obtained spectra was reduced using LDA with the label for each strain (Fig. 4). Confirming the cluster distance of each strain for the Raman spectra-derived features created in our study, the closely related species in the results obtained by 16S rRNA analysis formed clusters at a close distance (Figs. 2 and 5). Clustering by 16S rRNA gene analysis infers the genus and species of an unknown bacterium based on its proximity to the genetic distance (Janda & Abbott, 2007). Thus, the bacteria were judged to be similar to the closest genus and species based on Raman spectra and LDA.

The bacterial and bacterial population behaviors were identified separately. Genetic approaches and spectroscopy methods have been used to identify bacteria (Ho et al., 2019; Yoon et al., 2017). Bacterial population behavior has been investigated using predictive microbiology. Our model successfully integrated Raman spectral data into the prediction of the unknown strain population behavior. Although sensory data have been suggested for use in assessing food safety and quality (Nychas et al., 2021), there is no research combining sensory data for the prediction of population behavior. Our approach enabled us to predict the behavior of the bacterial population. Our evaluation will be beneficial for dealing with foodborne diseases and food spoilage incidents.

Rapid and simple identification of bacterial species and their stress responses plays a key role in determining bacterial control methods. Raman spectroscopy has been used to identify food-related microorganisms in the previous studies (Huayhongthong et al., 2019; Meisel et al., 2014; Yilmaz et al., 2015). However, the stress tolerance of this strain remains unclear. Although the stress tolerance of a species is inferred from similar species, stress tolerance can differ among species. For example, some species of *Pseudomonas* sp. have different stress tolerances to sodium acetate (Fig. 6). Links among bacteria are required to predict bacterial responses. Our approach attempted to overcome this

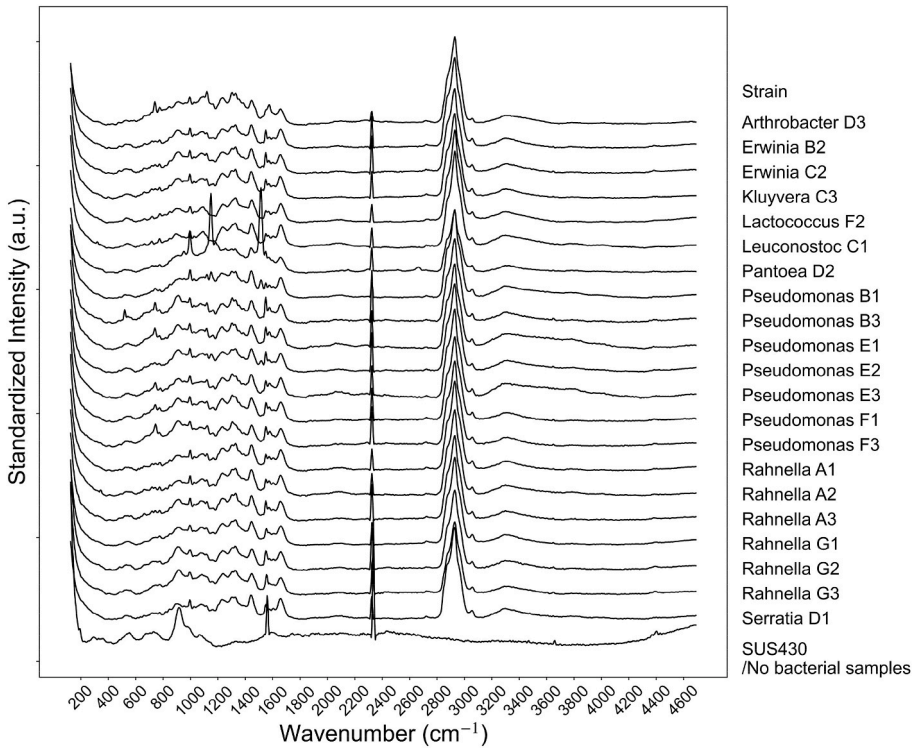

Fig. 3. Mean spectra of strain after preprocessing that cosmic ray removed, baseline corrected, smoothed and intensity standardized. SUS430 was the result of measurement without bacterial samples. The 20 single cell spectra were used 20 for each average spectrum.

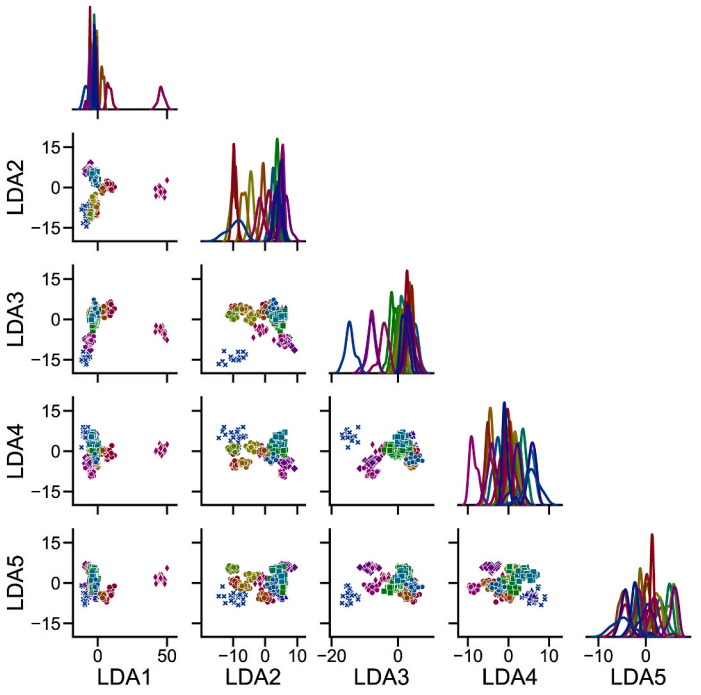

Fig. 4. A matrix of scatter plot of LDA 1 to LDA 5 derived from Raman spectra. Color-coded by strain and each symbol has a different type of genus by the similarity of 16S rRNA of each isolation (Table 1), as follows, × : Arthrobacter sp., ▲: Erwinia sp., ★: Kluyvera sp., ◆: Leuconostoc sp., ⬢: Lactococcus sp., ◆: Pantoea sp., ■: Rahnella sp., ●: Pseudomonas sp., ⬟: Serratia sp. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

problem by using the similarity between known and unknown bacteria characterized in terms of their Raman scattering properties. Considering that the Raman spectroscopy approach is rapid for characterizing bacteria compared to genetic methods, the method developed in this study could rapidly predict growth/no growth for the same genera with different stress tolerances.

The lowest AUC was 0.5 in the prediction of Pantoea D2 (Table 2). One likely explanation for the low prediction performance was that the Raman spectra and clusters of Pantoea D2 were significantly different from those of other strains (Figs. 4 and 5). Pantoea sp. have been reported to have different Raman features among bacteria (Polisetti et al., 2016). Because the prediction accuracy was low, a sample of Pantoea D2 should be assigned to an unpredictable condition. This constraint cannot usually be guaranteed *a priori* in our predictions. Therefore, the

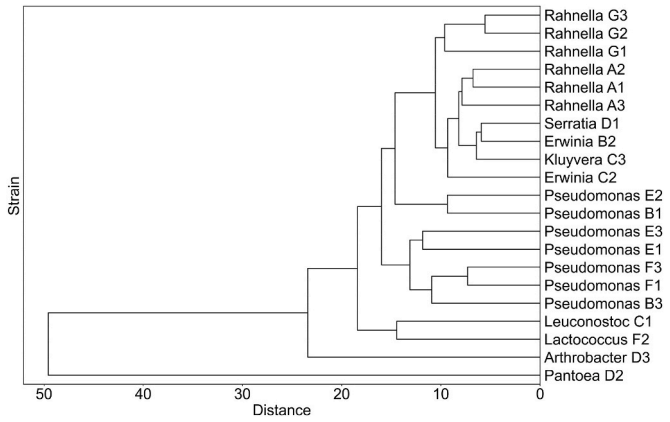

Fig. 5. Dendrogram of an agglomerative hierarchical cluster analysis performed on LDA values of the preprocessed Raman spectra of twenty-one strains. UPGMA algorithm was used for cluster analysis.

reliability of the prediction results was estimated. For example, unknown bacteria should be characterized by a Gaussian process (Kemmler et al., 2013) or binary classification (Meisel et al., 2012) before prediction. Analytical techniques should be developed to ensure the reliability of predictions for the growth/no-growth of unknown bacteria in the future.

This study has limited performance in probabilistic evaluation. Since the experimental repetition is three, the probabilistic part is roughly estimated. For a more detailed prediction, much more repetition is needed (Tsuruma et al., 2021). Nevertheless, the presented model can predict the boundary of growth/no growth.

Predicting bacterial growth in food products may require predictions under complex conditions because there are many food composition factors and environmental conditions that can affect microbial growth (Labuza & Fu, 1993). Environmental and food storage stresses (such as heating, cooling, and exposure to acidity or alkalinity) damage various cellular structures, such as microbial cell walls, cell membranes, proteins, RNA, and DNA (Wesche et al., 2009). Athamneh et al. (Athamneh et al., 2014) found that the Raman spectra of *E. coli* contain sufficient biochemical information to distinguish phenotypes induced by antibiotics with different mechanisms. Phenotypic profiling and discriminant analysis were used to demonstrate the potential for predicting the function (inhibition of protein, cell wall, DNA, and RNA synthesis) of antibiotics whose mechanism of action is unknown. In other words, since Raman spectroscopy might utilize information at the DNA and RNA levels as well as the protein and cell wall, we might be able to derive information on environmental and food storage stresses that act at the microbial cell wall, protein, and DNA levels. The combination of Raman spectroscopy and machine learning would allow us to predict the bacterial response to various environments with pH, osmotic pressure, and heat, which can be applied in food processing and preservation industries in the future.

5. Conclusion

Our machine learning model could predict the growth/no growth response of 21 unknown bacterial strains in culture media with 90% overall accuracy. Our modeling procedure enabled us to extract information from Raman spectra of known bacteria to predict the growth/no growth of unknown bacteria that were not used in the training of the

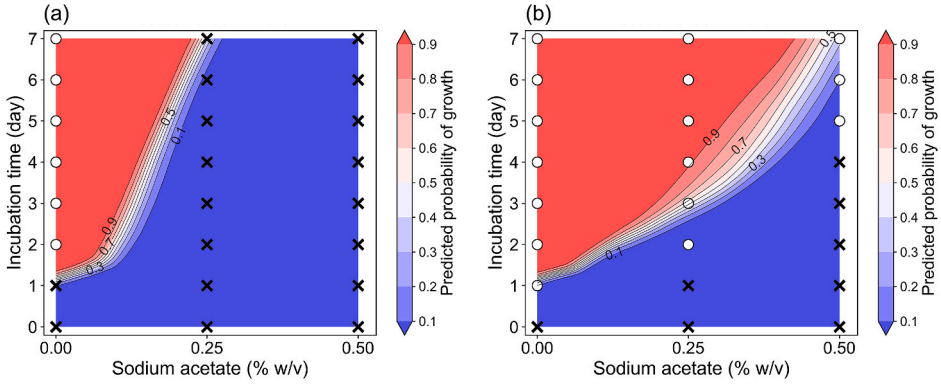

Fig. 6. The representative observation of growth (○)/no growth (×) of Pseudomonas E2 (a) and Pseudomonas F1 (b) as unknown bacteria at 10 °C and growth probability contour lines for unknown strain predicted by the artificial neural network model.

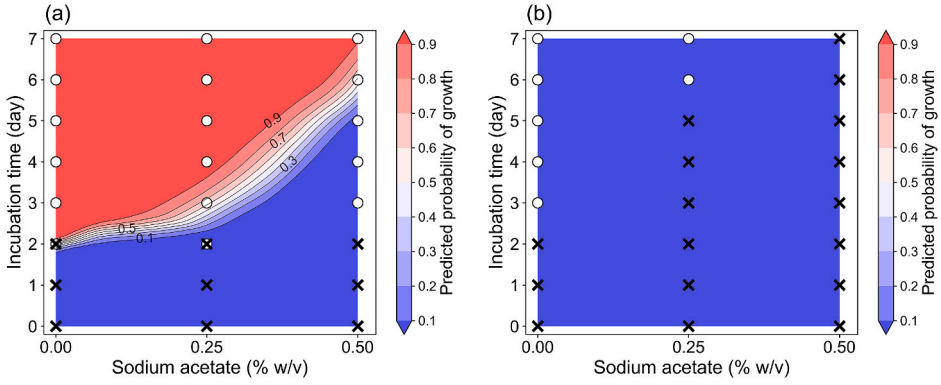

Fig. 7. The representative observation of growth (○)/no growth (×) of Leuconostoc C1 (a) and Pantoea D2 (b) as unknown bacteria at 10 °C and growth probability contour lines for unknown strain predicted by artificial neural network model.

Table 2
Overall accuracy and AUC when predicting growth/no growth with Raman spectral features. In all 21 strains, one strain was used as test data, and the remaining 20 strains were used as training data.

Test strains	Unknown level	Overall accuracy	AUC
Arthrobacter D3	Genera	0.82	0.88
Erwinia B2	Strain	0.89	0.94
Erwinia C2	Strain	0.95	0.96
Kluyvera C3	Genera	0.93	0.94
Lactococcus F2	Genera	0.82	0.86
Leuconostoc C1	Genera	0.88	0.89
Pantoea D2	Genera	0.77	0.50
Pseudomonas B1	Species	0.95	0.97
Pseudomonas B3	Species	0.86	0.84
Pseudomonas E1	Species	0.81	0.78
Pseudomonas E2	Species	1.00	1.00
Pseudomonas E3	Species	0.93	0.97
Pseudomonas F1	Strain	0.86	0.88
Pseudomonas F3	Strain	0.80	0.87
Rahnella A1	Species	0.94	0.96
Rahnella A2	Strain	0.93	0.95
Rahnella A3	Strain	0.94	0.97
Rahnella G1	Strain	0.94	0.97
Rahnella G2	Strain	0.96	0.97
Rahnella G3	Strain	0.98	0.99
Serratia D1	Genera	0.92	0.92
Average	Genera	0.86	0.83
	Species	0.92	0.92
	Strain	0.92	0.94
	Overall	0.90	0.91

model. Prediction of the growth behavior of unknown bacteria is valuable for the food manufacturing industry, where unknown bacteria are likely to be detected.

CRediT authorship contribution statement

Takashi Yamamoto: Conceptualization, Formal analysis, Data curation, Writing – original draft, were convinced of this concept, conducted an experiment, analyzed the data, and wrote the Python script, All authors have reviewed the manuscript. **J. Nicholas Taylor:** Conceptualization, were convinced of this concept, All authors have reviewed the manuscript. **Shige Koseki:** Conceptualization, were convinced of this concept, All authors have reviewed the manuscript. **Kento Koyama:** Conceptualization, were convinced of this concept, designed the computations, All authors have reviewed the manuscript.

Declaration of competing interest

There are no conflicts of interest to declare.

Data availability

The data that has been used is confidential.

Acknowledgments

This work was supported by the Kieikai Research Foundation and Tobe Maki Scholarship Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lwt.2023.114449.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). TensorFlow: Large-Scale machine learning on heterogeneous distributed systems. https://www.tensorflow.org/.

Aryani, D. C., den Besten, H. M. W., Hazeleger, W. C., & Zwietering, M. H. (2015). Quantifying strain variability in modeling growth of Listeria monocytogenes. *International Journal of Food Microbiology, 208*, 19–29. https://doi.org/10.1016/j.ijfoodmicro.2015.05.006

Athamneh, A. I. M., Alajlouni, R. A., Wallace, R. S., Seleem, M. N., & Sengera, R. S. (2014). Phenotypic profiling of antibiotic response signatures in Escherichia coli using Raman spectroscopy. *Antimicrobial Agents and Chemotherapy, 58*(3), 1302–1314. https://doi.org/10.1128/AAC.02098-13

De Boer, P. T., Kroese, D. P., Mannor, S., & Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research, 134*(1), 19–67. https://doi.org/10.1007/s10479-005-5724-z

Dengremont, E., & Membre, J. M. (1995). Statistical approach for comparison of the growth rates of five strains of Staphylococcus aureus. *Applied and Environmental Microbiology, 61*(12), 4389–4395. https://doi.org/10.1128/aem.61.12.4389-4395.1995

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. In *Intensive care medicine* (pp. 1043–1051). Springer. https://doi.org/10.1007/s00134-003-1761-8. Vol. 29, Issue 7.

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics, 8*(4), 376–386. https://doi.org/10.1111/j.1469-1809.1938.tb02189.x

Germond, A., Ichimura, T., Horinouchi, T., Fujita, H., Furusawa, C., & Watanabe, T. M. (2018). Raman spectral signature reflects transcriptomic features of antibiotic resistance in Escherichia coli. *Communications Biology, 1*(1). https://doi.org/10.1038/s42003-018-0093-8

Hajmeer, M., & Basheer, I. (2002). A probabilistic neural network approach for modeling and classification of bacterial growth/no-growth data. *Journal of Microbiological Methods, 51*(2), 217–226. https://doi.org/10.1016/S0167-7012(02)00080-5

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Ho, C. S., Jean, N., Hogan, C. A., Blackmon, L., Jeffrey, S. S., Holodniy, M., Banaei, N., Saleh, A. A. E., Ermon, S., & Dionne, J. (2019). Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nature Communications, 10*(1). https://doi.org/10.1038/s41467-019-12898-9

Huayhongthong, S., Khuntayaporn, P., Thirapanmethee, K., Wanapaisan, P., & Chomnawang, M. T. (2019). Raman spectroscopic analysis of food-borne microorganisms. *LWT, 114*(May), Article 108419. https://doi.org/10.1016/j.lwt.2019.108419

Jaafreh, S., Valler, O., Kreyenschmidt, J., Günther, K., & Kaul, P. (2019). In vitro discrimination and classification of Microbial Flora of Poultry using two dispersive Raman spectrometers (microscope and Portable Fiber-Optic systems)in tandem with chemometric analysis. *Talanta, 202*, 411–425. https://doi.org/10.1016/j.talanta.2019.04.082

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology, 45*(9), 2761–2764. https://doi.org/10.1128/JCM.01228-07

Jay, J. M., Loessner, M. J., & Golden, D. A. (2005). *Modern food microbiology*. Springer US. https://doi.org/10.1007/b100840

Kemmler, M., Rodner, E., Rösch, P., Popp, J., & Denzler, J. (2013). Automatic identification of novel bacteria using Raman spectroscopy and Gaussian processes. *Analytica Chimica Acta, 794*, 29–37. https://doi.org/10.1016/j.aca.2013.07.051

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution, 16*(2), 111–120. https://doi.org/10.1007/BF01731581

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. 3rd international conference on learning representations. *ICLR 2015 - Conference Track Proceedings, 1–15*.

Klein, D., Breuch, R., von der Mark, S., Wickleder, C., & Kaul, P. (2019). Detection of spoilage associated bacteria using Raman-microspectroscopy combined with multivariate statistical analysis. *Talanta, 196*(December 2018), 325–328. https://doi.org/10.1016/j.talanta.2018.12.094

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). Mega X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution, 35*(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

Labuza, T. P., & Fu, B. (1993). Growth kinetics for shelf-life prediction: Theory and practice. *Journal of Industrial Microbiology, 12*(3–5), 309–323. https://doi.org/10.1007/BF01584208

Lee, H., & Song, J. (2019). Introduction to convolutional neural network using Keras; an understanding from a statistician. *Communications for Statistical Applications and Methods, 26*(6), 591–610. https://doi.org/10.29220/CSAM.2019.26.6.591

Lehto, M., Kuisma, R., Määttä, J., Kymäläinen, H. R., & Mäki, M. (2011). Hygienic level and surface contamination in fresh-cut vegetable production plants. *Food Control, 22*(3–4), 469–475. https://doi.org/10.1016/j.foodcont.2010.09.029

Letunic, I., & Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research, 49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Lianou, A., & Koutsoumanis, K. P. (2011). Effect of the growth environment on the strain variability of Salmonella enterica kinetic behavior. *Food Microbiology, 28*(4), 828–837. https://doi.org/10.1016/j.fm.2010.04.006

Lianou, A., & Koutsoumanis, K. P. (2013). Strain variability of the behavior of foodborne bacterial pathogens: A review. *International Journal of Food Microbiology, 167*(3), 310–321. https://doi.org/10.1016/j.ijfoodmicro.2013.09.016

Lieber, C. A., & Mahadevan-Jansen, A. (2003). Automated method for subtraction of fluorescence from biological Raman spectra. *Applied Spectroscopy, 57*(11), 1363–1367. https://doi.org/10.1366/000370203322554518

Meisel, S., Stöckel, S., Elschner, M., Melzer, F., Rösch, P., & Popp, J. (2012). Raman spectroscopy as a potential tool for detection of Brucella spp. *Milk, 78*(16), 5575–5583. https://doi.org/10.1128/AEM.00637-12

Meisel, S., Stöckel, S., Rösch, P., & Popp, J. (2014). Identification of meat-associated pathogens via Raman microspectroscopy. *Food Microbiology, 38*, 36–43. https://doi.org/10.1016/j.fm.2013.08.007

Moreira, L. M., Silveira, L., Jr., Santos, F. V., Lyon, J. P., Rocha, R., Zângaro, R. A., Villaverde, A. B., & Pacheco, M. T. T. (2008). Raman spectroscopy: A powerful technique for biochemical analysis and diagnosis. *Spectroscopy, 22*(1), 1–19. https://doi.org/10.1155/2008/942758

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *International conference on machine learning*. https://openreview.net/pdf?id=rkb15iZdZB.

Nychas, G.-J., Sims, E., Tsakanikas, P., & Mohareb, F. (2021). Data science in the food industry. *Annual Review of Biomedical Data Science, 4*(1), 341–367. https://doi.org/10.1146/annurev-biodatasci-020221-123602

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., & Invernizzi, L. (2019). Keras tuner. Retrieved May, 21, 2020 https://github.com/keras-team/keras-tuner.

Polisetti, S., Bible, A. N., Morrell-Falvey, J. L., & Bohn, P. W. (2016). Raman chemical imaging of the rhizosphere bacterium Pantoea sp. YR343 and its co-culture with Arabidopsis thaliana. *Analyst, 141*(7), 2175–2182. https://doi.org/10.1039/c6an00080k

Presser, K. A., Ross, T., & Ratkowsky, D. A. (1998). Modelling the growth limits (growth/No growth interface) of Escherichia coli as a function of temperature, pH, lactic acid concentration, and water activity. *Applied and Environmental Microbiology, 64*, 1773–1779. https://doi.org/10.1128/AEM.64.5.1773-1779.1998

Raskutti, G., Wainwright, M. J., & Yu, B. (2014). Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research, 15*(1), 335–366. https://www.jmlr.org/papers/volume15/raskutti14a/raskutti14a.pdf.

Rawat, S. (2015). Food spoilage: Microorganisms and their prevention. *Pelagia Research Library Asian Journal of Plant Science and Research, 5*(4), 47–56. www.pelagiaresearchlibrary.com.

Rodríguez-Saavedra, M., Pérez-Revelo, K., Valero, A., Moreno-Arribas, M. V., & González de Llano, D. (2021). A binary logistic regression model as a tool to predict craft beer susceptibility to microbial spoilage. *Foods, 10*, 1–15. https://doi.org/10.3390/foods10081926

Ross, T., & McMeekin, T. A. (1994). Predictive microbiology. *International Journal of Food Microbiology, 23*(3–4), 241–264. https://doi.org/10.1016/0168-1605(94)90155-4

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One, 10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Schumacher, W., Stöckel, S., Rösch, P., & Popp, J. (2014). Improving chemometric results by optimizing the dimension reduction for Raman spectral data sets. *Journal of Raman Spectroscopy, 45*(10), 930–940. https://doi.org/10.1002/jrs.4568

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958. https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_campaign=buffer&utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com.

Stavropoulou, E., & Bezirtzoglou, E. (2019). Predictive modeling of microbial behavior in food. *Foods, 8*(Issue 12). https://doi.org/10.3390/foods8120654. mdpi.com.

Taylor, J. N., Mochizuki, K., Hashimoto, K., Kumamoto, Y., Harada, Y., Fujita, K., & Komatsuzaki, T. (2019). High-resolution Raman microscopic detection of follicular thyroid cancer cells with unsupervised machine learning. *Journal of Physical Chemistry B, 123*(20), 4358–4372. https://doi.org/10.1021/acs.jpcb.9b01159

Tsuruma, N., Doto, S., Ishida, W., Koyama, K., & Koseki, S. (2021). How many repetitions per condition are required for developing a stable growth/no growth boundary model for Bacillus simplex spores? *Food Control, 122*, Article 107756. https://doi.org/10.1016/j.foodcont.2020.107756

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Walls, I., & Scott, V. N. (1997). Use of predictive microbiology in microbial food safety risk assessment. *International Journal of Food Microbiology, 36*(2–3), 97–102. https://doi.org/10.1016/S0168-1605(97)01260-9

Wesche, A. M., Gurtler, J. B., Marks, B. P., & Ryser, E. T. (2009). Stress, sublethal injury, resuscitation, and virulence of bacterial foodborne pathogens. *Journal of Food Protection, 72*(Issue 5), 1121–1138. https://doi.org/10.4315/0362-028X-72.5.1121. meridian.allenpress.com.

Yamamoto, T., Taylor, J. N., Koseki, S., & Koyama, K. (2021). Classification of food spoilage bacterial species and their sodium chloride, sodium acetate and glycine tolerance using chemometrics analysis and Raman spectroscopy. *Journal of Microbiological Methods, 190*(September), Article 106326. https://doi.org/10.1016/j.mimet.2021.106326

Yan, S., Wang, S., Qiu, J., Li, M., Li, D., Xu, D., Li, D., & Liu, Q. (2021). Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. *Talanta, 226*(February), Article 122195. https://doi.org/10.1016/j.talanta.2021.122195

Yilmaz, A. G., Temiz, H. T., Acar Soykut, E., Halkman, K., & Boyaci, I. H. (2015). Rapid identification of Pseudomonas aeruginosa and Pseudomonas fluorescens using Raman spectroscopy. *Journal of Food Safety, 35*(4), 501–508. https://doi.org/10.1111/jfs.12200

Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H., & Chun, J. (2017). Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *International Journal of Systematic and Evolutionary Microbiology, 67*(5), 1613–1617. https://doi.org/10.1099/ijsem.0.001755