HHS Public Access

Author manuscript

Int J Biol Macromol. Author manuscript; available in PMC 2024 July 27.

Published in final edited form as:

Int J Biol Macromol. 2024 February ; 257(Pt 2): 128773. doi:10.1016/j.ijbiomac.2023.128773.

Machine learning enabled multiplex detection of periodontal pathogens by surface-enhanced Raman spectroscopy

Rathnayake A.C. Rathnayake[a,1], **Zhenghao Zhao**[b,1], **Nathan McLaughlin**[c], **Wei Li**[d], **Yan Yan**[b,*], **Liaohai L. Chen**[c], **Qian Xie**[e], **Christine D. Wu**[d], **Mathew T. Mathew**[f,g], **Rong R. Wang**[a,*]

[a]Department of Chemistry, Illinois Institute of Technology, Chicago, IL 60616, United States of America

[b]Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, United States of America

[c]Department of Surgery, University of Illinois Chicago, Chicago, IL 60612, United States of America

[d]Department of Pediatric Dentistry, University of Illinois Chicago, Chicago, IL 60612, United States of America

[e]Department of Endodontics, University of Illinois Chicago, Chicago, IL, United States of America

[f]Department of Restorative Dentistry, University of Illinois Chicago, Chicago, IL 60612, United States of America

[g]Department of Biomedical Sciences, University of Illinois Rockford, Rockford, IL 61107, United States of America

Abstract

Periodontitis is a chronic inflammation of the periodontium caused by a persistent bacterial infection, resulting in destruction of the supporting structures of teeth. Analysis of microbial composition in saliva can inform periodontal status. *Actinobacillus actinomycetemcomitans* (*Aa*), *Porphyromonas gingivalis* (*Pg*), and *Streptococcus mutans* (*Sm*) are among reported periodontal pathogens, and were used as model systems in this study. Our atomic force microscopic (AFM) study revealed that these pathogens are biological nanorods with dimensions of 0.6–1.1 μm in length and 500–700 nm in width. Current bacterial detection methods often involve complex preparation steps and require labeled reporting motifs. Employing surface-enhanced Raman spectroscopy (SERS), we revealed cell-type specific Raman signatures of these pathogens for

[*]Corresponding authors. yyan34@iit.edu (Y. Yan), wangr@iit.edu (R.R. Wang).

[1]Equal contributors.

Declaration of competing interest

The authors declare no competing financial interest.

CRediT authorship contribution statement

R.R.W., Y.Y., L.L.C. contributed to conceptualization and experimental design; R.A.C.R. & N.M. collected and analyzed Raman spectra and AFM images under the supervision of R.R.W., L.L.C.; Z.Z. & Y.Y. developed the ML methods and predictive models; Q.X., W.L., M.M. and C.D.W. resourced and cultured the oral bacteria. R.A.C.R., Z.Z., R.R.W., Y.Y. drafted the manuscript with input from all authors. All authors have reviewed and agreed to the published version of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2023.128773.

label-free detection. It overcame the complexity associated with spectral overlaps among different bacterial species, relying on high signal-to-noise ratio (SNR) spectra carefully collected from pure species samples. To enable simple, rapid, and multiplexed detection, we harnessed advanced machine learning techniques to establish predictive models based on a large set of raw spectra of each bacterial species and their mixtures. Using these models, given a raw spectrum collected from a bacterial suspension, simultaneous identification of all three species in the test sample was achieved at 95.6 % accuracy. This sensing modality can be applied to multiplex detection of a broader range and a larger set of periodontal pathogens, paving the way for hassle-free detection of oral bacteria in saliva with little to no sample preparation.

Keywords

SERS; Machine learning; Periodontal pathogens; Label-free; Multiplex detection

1. Introduction

Periodontal disease is a significant public health concern. Approximately 47 % of adults over age 30 have some form of the disease. Cases tend to become more common and severe with age. These characteristics place periodontitis as the 11th most prevalent disease, impacting hundreds of millions globally [1]. In dentistry, standard diagnostic criteria include assessing bleeding upon probing, clinical attachment levels, plaque index, gingival index, and bone loss. These parameters are effective in retrospectively identifying the presence of the disease, but have limitations in detecting ongoing conditions or predicting individuals at high risk of future periodontitis. Moreover, considerable tissue damage must occur before these diagnostic parameters become noticeable, often leading to irreversible periodontal damage.

The clinical shifts are driven by changes in the microbial composition that colonizes at or below the gingival margin [2,3]. The oral microbiota is the second-most complex microbial community in the human body. A milliliter of saliva can contain up to 10^8 microorganisms from over 700 prokaryotic species [4,5]. A small fraction of these microorganisms contributes to periodontal disease. Among them, *Actinobacillus actinomycetemcomitans* (*Aa*) is strongly implicated as an etiological agent in periodontitis. This gram-negative bacterium produces numerous virulent factors that contribute to invasion of periodontal tissues, initiation of connective tissue destruction, and interference with tissue repair [6–8]. *Porphyromonas gingivalis* (*Pg*), also gram-negative, is recognized as a major periodontal causative pathogen involved in alveolar bone loss and collagen degradation [9]. *Streptococcus mutans* (*Sm*), a gram-positive coccus thriving in acidic environments, plays a significant role in the initiation of dental caries and the progression of periodontitis [2,10]. These and other oral bacteria accumulate in periodontal pockets and work cooperatively to induce periodontal destruction, leading to gingival inflammation, soft tissue destruction, and bone loss. They serve as biomarkers to indicate the status of periodontal health [11]. In this study, *Aa*, *Pg* and *Sm* were employed as model systems to establish a machine learning enabled, label-free detection method.

There are various methods to detect periodontal pathogens, such as anaerobic cultures, immunoassays, polymerase chain reaction (PCR), as well as aptamers, DNA, RNA, oligonucleotide, ligand, or antibody-based sensing assays against periodontal pathogens [10,12–14]. While these assays are effective, they often involve labor-intensive and time-consuming procedures, and require critical preparation steps, such as cell lysis or fixation, method and reagents to promote DNA hybridization, and target amplification to enhance detection [10,12]. Additionally, they require labeled reporter molecules for signal readout, limiting the methods' adaptability due to the need for designing specific probes for each cell type. Another major drawback is the limited capability of rapid multiplex detection.

Surface-enhanced Raman spectroscopy (SERS) has emerged as a potent label-free detection tool capable of discriminating different bacterial phenotypes due to Raman signatures of their unique molecular composition [15]. However, the efficiency of Raman scattering is typically low, resulting in poor signal-to-noise ratio (SNR) and low spectrum fidelity. This is particularly true in fast spectral collection even with the presence of local field enhancement. It is also challenging to distinguish the molecular fingerprint of one phenotype from others in a bacterial mixture based on subtle differences in their whole-cell Raman spectra. Promising results were achieved by employing metal nanostructures as SERS substrates, which create "hot spots" with strong local field enhancements on surfaces to significantly amplify the Raman scattering effect on analytes [16–18]. By using a tailored plasmonic material and a highly focused laser beam, single-cell SERS with a high SNR were achieved. Statistical analysis techniques and machine learning methods have been applied to successfully identify individual bacterial species in their resident media [19–21], even in mixed cultures [21,22]. In these cases, the laser beam must be directed at individual cells, and Raman spectra are collected on a cell-by-cell basis. To facilitate SERS utility in clinical applications, it is imperative to develop a method that requires minimal sample preparation, technical expertise, and is easy to use.

With the increasing prominence of machine learning and artificial intelligence, these technologies have played a significant role in enhancing experiments and data analysis across various fields, including chemistry, biology, and medical science [16,23]. While traditional machine learning (ML) methods have been applied to analyze Raman spectral data [24,25], deep learning utilizing artificial neural networks is more sophisticated and capable of providing more accurate predictions [26]. In this study, we leverage the power of deep learning to achieve rapid, multiplex detection of three periodontal pathogens from a single spectral measurement of a bacterial mixture. A commercial Raman system was used to collect SERS of a bacterial suspension deposited on a gold-coated substrate. A raw spectrum of the test sample, without background correction or denoising, was sufficient for quick bacterial identification. This approach opens up the possibility of detecting oral bacteria in saliva, which contains locally and systemically derived mediators of periodontal disease. Saliva is a non-invasive biofluid, easily accessible, and simple to collect, making it an ideal diagnostic media for convenient, frequent monitoring, thereby facilitating early diagnosis and timely intervention to improve medical outcomes.

2. Materials and methods

2.1. Bacteria and sample preparation

Three species of oral bacteria, namely *Aggregatibacter actinomycetemcomitans* (ATCC 43718), *Porphyromonas gingivalis* (ATCC 33277), and *Streptococcus mutans* (strain UA159, ATCC 700610), were used in this study. *A. actinomycetemcomitans* was grown in Todd Hewitt Broth (THB, BD, cat #249240, Becton, Dickinson and Company, Sparks, MD, USA) supplemented with 1 % yeast extract (BD, cat #212750), 0.001 % Hemin (Sigma-Aldrich, Inc., St. Louis, MO, USA) & 0.0001 % of Vitamin K1 (Sigma). The cultivation took place in an anaerobic chamber (Forma Scientific Inc., Marjetta, OH, USA) filled with an anaerobic gas mixture (10 % H_2, 5 % CO_2, and 85 % N_2, Linde Gas & Equipment Inc., Burr Ridge, IL, USA) at 37 °C for 48 h. *P. gingivalis* was grown in THB medium supplemented with 0.001 % Hemin & 0.0001 % of Vitamin K1 under anaerobic conditions at 37 °C for 48 h. *S. mutans* was cultivated in Brain Heart Infusion (BD, cat #237500) at 37 °C with 5 % CO_2 overnight. The bacterial cells were harvested and fixed immediately following a previously reported method [27]. In brief, the cells were fixed using a fresh, cold paraformaldehyde solution (4 % in PBS) and left to incubate at room temperature for a minimum of 3 h. The cells were then pelleted by centrifugation at 4,000 rpm for 5 min, washed with PBS three times, and resuspended in PBS to achieve a concentration of 10^7–10^8 cells/mL. For storage, the cell suspension was added to cold absolute ethanol at a 1:1 volume ratio, and the mixture was stored at −20 °C.

For AFM and SERS measurements, a 400 μL fixed bacterial cell suspension was placed in a 1.5 mL tube. After centrifugation at 12,000 rpm for 10 min, the supernatant was discarded, and the pellet was washed three times with DI water. Each pellet was then reconstituted in 50 μL DI water. Subsequently, 2.5 μL of the suspension was deposited onto a gold-coated glass substrate, air-dried in less than an hour, and then subjected to SERS or AFM measurements.

2.2. AFM imaging

AFM imaging of bacterial cells was carried out using an Agilent 5500 Pico-Plus system (Santa Clara, CA). The measurements were conducted in air contact mode using a closed-loop scanner [28–30]. Si_3N_4 probes (#SINI-100, Ted Pella, Redding, CA) with a tip radius below 15 nm and a spring constant of 0.06 N/m were utilized to collect images at a scan rate of 0.5 ln/s. The system's software was used for image analysis. Large-scale images were acquired to examine cell distribution. The cell dimension of each bacterial strain was derived from high-resolution images, and the average value was calculated based on measurements of over 50 cells on images taken at five randomly selected locations for each cell type.

2.3. SERS measurements

Raman spectra were acquired from bacterial cells placed on a glass substrate coated with a 100 nm gold film. The gold deposition was carried out using a Varian 3118 electron beam evaporator. An 80 nm titanium layer was pre-deposited on the glass to improve adhesion of the gold film. AFM imaging revealed the presence of grains measuring 69.0 ± 3.2 nm on the gold-coated substrate to attain surface-enhancement of Raman signals. A Renishaw

Invia Raman Microscope (West Dundee, IL) equipped with a 785 nm laser at 90 mW was used for SERS measurements. The spectra were collected in the range of 300 cm^{-1} – 2000 cm^{-1} using a 50× objective lens. To capture distinctive Raman features for each cell type, high-resolution spectra were recorded with an 80 s exposure and 10 accumulations. Additionally, fast spectra were collected with a 30 s exposure and 1 accumulation. Over 100 fast spectra were collected for each pure species, mixtures of any two species (1:1 ratio), or mixtures of all three species (1:1:1 ratio). These data were subsequently used t for machine learning to establish predictive models for multiplex detection of oral bacteria.

2.4. Machine learning model development

The machine learning workflow, as depicted in Fig. 1, followed a series of steps. Initially, raw SERS spectra collected in 30 s underwent data preprocessing. This involved data shuffling, standardization, principal component analysis (PCA) [31], train-test set split, and data augmentation. Each spectrum contained 1833 data points. Standardization was performed by rescaling each spectrum using the formula $y' = (y - \mu)/\sigma$, where y, μ, and σ represents the spectral intensity, the mean value of y, and the standard error, respectively. Based on prior research [13] and our own ablation studies, PCA was employed to reduce the dimensionality of each spectrum from 1833 data points to 35 data points. The large number of spectra was then divided randomly, with 70 % for training and validation and 30 % for testing. Among the spectra designated for training and validation, 75 % were used for training and 25 % for validation. Multiple data augmentation methods, including jitter, permutation, SuboPtimAl Warped time-series geNEratoR (SPAWNER) [32], Weighted Dynamic Time Warping Barycenter Averaging (wDBA), Random Guided Warping (RGW) [26] and Discriminative Guided Warping (DGW) [26], were applied to enhance the model's regularization and robustness while mitigating biased training and overfitting.

For each oral bacterial cell type, a binary classification model was constructed to detect its presence in the samples. These models consisted of two fully connected layers. We used the Adam optimizer [27] with an exponential decay learning rate schedule and a batch size of 256 for all experiments. Each model underwent training over 1000 epochs using the training data. Subsequently, the models were assessed using the test dataset. The performance of our method and classical machine learning methods were evaluated across accuracy, precision, recall, and F1 score, which are defined as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

$$precision = \frac{TP}{TP + FP},$$

$$recall = \frac{TP}{TP + FN},$$

$$F1\ \ score = \frac{2 \times precision \times recall}{precision + recall},$$

Here, TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively. The results were reported based on five separate stratified sampling splits for training and testing.

3. Results

3.1. AFM analysis of microbial shape and dimension

AFM images were collected from cells of pure bacterial species deposited on gold-coated glass substrates. The same sample preparation procedure was employed for SERS measurements. Consequently, these images depict the spatial distribution and structural characteristics of the cells in SERS measurements. As shown in Fig. 2A–C, *Aa* cells were found either as isolated individual cells or as small cell clusters, whereas *Pg* cells formed large aggregates due to the presence of a unique outer membrane protein known as a crucial factor promoting cell-cell interactions [9]. *Sm* cells organized themselves into densely packed cell sheets, indicating a higher degree of cell aggregation. Being a cariogenic bacterium, *Sm* produces multiple glucan-binding proteins that facilitate local cell accumulation, adhesion, and biofilm formation even in the absence of nutrients [33,34]. This finding aligns with recent reports that *Sm* is one of the most etiological agents to form biofilm at oral cavities [34,35].

With high-resolution images (Fig. 2D–F), the dimensions of individual cells were measured. The cell length and width were found to be 1.09 ± 0.30 μm and 0.56 ± 0.10 μm for *Aa* cells, 0.67 ± 0.20 μm and 0.60 ± 0.20 μm for *Pg* cells, and 0.95 ± 0.20 μm and 0.54 ± 0.10 μm for *Sm* cells, along with the cell length-to-width ratios of 1.9 ± 0.7, 1.1 ± 0.5, and 1.8 ± 0.7, respectively. Supplementary AFM images and cross-sectional analyses can be found in Fig. S1. At the individual cell level, *Pg* cells exhibit a nearly spherical shape and are notably smaller in size. In contrast, *Aa* and *Sm* cells displayed a rod-shaped morphology with similar cell dimension. These observations are consistent with results in previous reports [16,36,37].

3.2. SERS signature of periodontal pathogens

High SNR Raman spectra were collected for each bacterial species at 80 s exposure and 10 accumulations. Fig. 3 shows the normalized spectra with respect to the prominent peak at 934 cm^{-1}, which is attributed to the C-COO^- stretching of phenylaniline and is a shared characteristic among all three cell types. These high SNR spectra exhibit similarities and distinctions in terms of Raman peak frequencies and intensities among different species. These vibrational bands are attributed by proteins, phospholipids, nucleic acids, and polysaccharides. Notably, the common bands at 1065 cm^{-1} and 1128 cm^{-1} correspond to the skeletal stretching of C-C and =C-O-C= in phospholipids, respectively [16,39]. The broad band at approximately 1333 cm^{-1} varies in shape for different cell types and is ascribed to CH_3CH_2 wigging mode in purine bases of nucleic acids [16]. The peak at 1001 cm^{-1} and the peaks around 560 cm^{-1} are distinctive features associated with the symmetric aromatic ring breathing and C–COO^- asymmetric bending of phenylalanine [40]. The Raman shift at 505 cm^{-1} is designated as the S–S disulfide stretch in proteins [17]. Two well-resolved low-frequency bands at 400 cm^{-1} and 462 cm^{-1} are likely a

result of vibrations in polysaccharides and phosphates [38,39,52]. These Raman shifts exhibit variations in band shape and intensity across different cell types. Furthermore, specific peaks unique to each cell type (highlighted in blue) were observed, although many appeared as insignificant shoulders. It is worth noting that amide I, amide II, and amide III vibrations, associated with the protein backbone and carboxylic stretches, fall within the 1220–1660 cm^{-1} region, overlapping with other SERS vibrational bands contributing to spectral complexity. Detailed interpretation of the Raman shifts and peak assignments are provided in Table 1.

To achieve fast detection, SERS spectra were collected at 30 s exposure and 1 accumulation (Fig. 4). Fig. 4A shows five raw spectra of pure *Aa* collected from five randomly selected locations on a substrate. According to locally obtained optical images and correlated AFM images, Raman spectra were typically collected from 9 to 15 cells (Fig. S1). These spectra in Fig. 4A exhibit a significantly reduced SNR compared to the spectrum in Fig. 3A, with some characteristic peaks obscured by the background. Each spectrum also displays distinctions from the other four spectra due to the limited fidelity of SERS (see Discussion section). As >100 randomly collected spectra were examined, the spectral distinctions became increasingly intricate. Similarly collected raw spectra of pure *Pg* and pure *Sm* are shown in Fig. 4B and C, with a comparable level of complexity. Fast spectra were also collected from mixed species of cells (Fig. 4D–G), leading to heightened intricacy due to the interplay of low SNR in the fast spectra and the overlap of the three pathogens' spectra.

3.3. ML-enabled identification of periodontal pathogens

We utilized machine learning techniques with the raw spectral data to create predictive models for the rapid identification of periodontal pathogens. The spectra were collected from samples in seven categories: pure *Aa*, pure *Sm*, pure *Pg*, *Aa-Sm* mixture, *Aa-Pg* mixture, *Sm-Pg* mixture, and *Aa-Sm-Pg* mixture.

Initially, we experimented with classical machine learning methods, including *Support Vector Machine* (*SVM*) [53], K-Nearest Neighbors (*KNN*) [54], *Decision Tree* [55], *Random Forest* [56], *Naive Bayes* [57], and *Logistic Regression* [58]. These methods do not require extensive data sets, but demonstrated suboptimal performance. Consequently, we turned to train with neural network models, which offer the advantage of capturing non-linear relationships in the data, thereby addressing complex problems and providing highly accurate predictions. However, neural network training requires a substantial amount of data, which was lacking in our collected raw spectra – a common issue in practical applications. To mitigate this challenge, we implemented data augmentation as explained in Section 2.4. Three distinctive detection models were developed for each bacteria species. Each model was trained for 1000 epochs with early stopping and used the Adam optimizer with a batch size of 256. For the *Aa* model, we applied a 0.3 dropout rate and a 0.001 initial learning rate with exponential decay. The *Sm* model used a 0.25 dropout rate and a 0.01 initial learning rate with exponential decay, while the *Pg* model was configured with a 0.3 dropout rate and an initial learning rate of 0.00001, subject to exponential decay. The performance of our method was compared to that of classical ML methods by evaluating detection accuracy, precision, recall, and the F1 score.

For singleplex detection, a successful model is expected to accurately identify the presence of a target species (*Aa*, *Sm* or *Pg*) in any sample of the seven groups, regardless of the present of other species. As shown in Fig. 5, our method consistently outperformed the classical ML methods, achieving >97.8 % accuracy, >98.2 % precision, >98.2 % recall and >98.2 % F1 score in predicting the presence of a target species in any sample. Note that the mean values and standard derivations were calculated from 5 randomly shuffled train-test splits, ensuring that our method consistently achieved a detection accuracy of 98.7 % ± 0.5 % for *Aa*, 99.2 % ± 0.7 % for *Pg*, and 97.8 % ± 0.9 % for *Sm*, regardless of how the spectral data was divided for training and testing. In contrast, the classical ML methods demonstrated poorer performance, with accuracy of predicting *Aa*, *Pg* or *Sm* in any given sample varied in the ranges of 71–90 %, 56–87 % or 63–88 %. Differences in prediction accuracy, recall, and F1 score were statistically significant ($p < 0.003$ for any classical method compared to our method). However, the difference in prediction precision varied, with linear SVM and Random Forest showing comparable results ($p > 0.1$ for *Sm* detection).

We explored the application of machine learning to determine pathogen abundance in a sample (evaluation of *Aa* abundance as depicted in Fig. S2). The actual number of cells was determined by AFM imaging of the area where SER data was collected from the focused laser beam. Among the seven methods we investigated, the Gradient Boosting Regression method [59] was able to make predictions that closely matched the actual cell counts (with a prediction error of <2.2 %). Further work is underway to extend this approach to determine pathogen abundance in a mixed sample.

Detecting multiple pathogens simultaneously, as in multiplex detection, presents a greater challenge. It requires the accurate prediction of all three species simultaneously based on a single spectrum pertaining to a bacterial sample. For example, the Random Forest algorithm performed well in singleplex detection, achieving prediction accuracy of 89.5±2.5 % for *Aa*, 86.2±2.2 % for *Pg*, 87.4±3.1 % for *Sm*. However, when it came to multiplex detection, its accuracy dropped to 69.3±5.3 % (Fig. 6). The performance of other classical models was worse. In contrast, our method demonstrated significantly superior performance, achieving a multiplex detection accuracy of 95.6±1.4 % (with a p value of <0.0005 compared to any classical method). The accuracy, precision, recall and F1 score of our method were 1.4–3.3, 1.2–8.1, 1.4–3.6 and 1.3–6.2 times better than those of the classical methods.

We further substantiated the efficacy of our method by constructing confusion matrices. Fig. 7A–C display the confusion matrices that illustrate accuracy in singleplex detection. For instance, when it came to detecting the presence of *Aa* in samples (including pure *Aa*, *Aa-Pg* mixture, *Aa-Sm* mixture, *Aa-Pg-Sm* mixture), the accuracy was an impressive 99.0 %, with an inaccuracy of only 1.0 %. Similarly, when determining the absence of *Aa* in samples (pure *Pg*, pure *Sm*, *Pg-Sm* mixture), the accuracy was 97.3 %, and the inaccuracy was 2.7 %. Fig. 7D shows a single confusion matrix representing multiplex detection accuracy for test samples across all seven groups. The diagonal sections of the matrix represent the detection accuracy (true positive rates). The matrix correlates the prediction results with the actual sample types, considering the presence and absence of all three species. For example, given a spectrum collected from a bacterial sample containing an *Aa-Sm* mixture, the accuracy in identifying that mixture was 93.3 %, with a very low likelihood

of misidentifying it as pure *Aa* (2.0 %), pure *Sm* (2.7 %) or an *Aa-Pg-Sm* mixture (2.0 %). It never misidentified the sample as pure *Pg*, *Aa-Pg* mixture, or *Sm-Pg* mixture (0.0 %). Remarkably, our model's performance in identifying the simultaneous presence of three species in an *Aa-Pg-Sm* mixture was exceptional (99.3 %). When considering all the data in the confusion matrix, the accuracy of detecting the presence and absence of any of the three species in a given sample was 95.6 %. This underscores the exceptional capability of our current method for label-free, multiplex detection of bacteria using raw SERS spectra even with low signal-to-noise ratios.

4. Discussion

SERS is a powerful analytical tool, as the vibrational information is highly specific to chemical signatures within a sample [60,61]. The technique's nondestructive nature, minimum sample preparation requirement, and label free structural fingerprinting capability attracted a broad range of applications in the biomedical field and have driven significant advancements in clinical utility. Notably, whole-cell Raman spectra were utilized to identify bacterial species based on spectral patterns, such as peak intensity, shape, and relative frequency across the entire spectral range, despite the complexity arising from the conservation of many chemical species in different bacterial cells [16,40]. Applying statistical methods, Rebrošová et al. and Vaitiekunait et al. effectively distinguished different bacterial species by utilizing Raman spectra collected from individual colonies of pure species [62,63]. We applied the same strategy to distinguish three periodontal pathogens (*Aa*, *Pg*, and *Sm*) commonly found in saliva.

Detecting multiple analytes within a sample is critical. Multiplex detection from a single spectral measurement was achieved using Raman-coded capture and reporting probes [64]. While effective, the preparation of Raman-active probes was laborious and technically demanding. Nanostructured metal surfaces have been used to significantly enhance Raman signals, enabling the collection of label-free Raman spectra of individual cells with high SNR for species identification. A major barrier is that, the spectrum of each cell must be measured one by one when the laser beam is focused on a target cell. These spectra of individual cells are screened for multiplex analysis. The process is time consuming. Achieving multiple species detection from a single Raman spectrum of a bacterial mixture remains a challenge. A main challenge is the attainment of SERS at high fidelity. During spectrum acquisition, the laser must strike a point where both the "hot spot" and the analyte are present and in proper proximity. The random and uncontrolled positioning of "hot spots" on a rough substrate hinders the ability to accurately determine the strength, polarization, and spatial distribution of these local fields, leading to significant variations in the observed spectra. Rapidly collected raw spectra were more intricate due to low SNRs (Fig. 4). Additionally, Raman spectra of different bacterial cell types, such as *Aa*, *Pg*, and *Sm*, exhibit subtle differences, resulting in substantial spectral overlaps.

To overcome these challenges, we applied machine learning techniques to the raw SERS spectra obtained from each of the three species and their mixtures. These spectra were simply collected from a drop of bacterial suspension deposited on a gold coated substrate and were used without background correction or denoising. Machine learning

assisted multiplex bacterial identification with SERS spectra was reported by other groups [19,20,22]. These studies utilized single-cell spectra of pure species to produce a single classification while to represent a multi-class classification task. That is, for any given sample, the reported models were tasked to identify one species from multiple possible species. Our method is distinctive. Our dataset included Raman spectra of all possible combination of bacterial species, tasking the model to predict the presence or absence of each species in any given sample based on a single Raman spectrum. This means that 1) the number of species in each test sample was unknown; 2) the presence of any species in each test sample was unknown; 3) a sample was classified correctly only when all three species were predicted correctly. For samples containing pure *Aa*, pure *Pg*, pure *Sm*, mixtures of *Aa-Pg*, *Aa-Sm*, *Pg-Sm* or *Aa-Pg-Sm*, our advanced model achieved an accuracy of 95.6 % in detecting the presence or absence of any of the three species.

We also attempted pathogen identification in samples containing bacteria spiked in sterile filtered saliva (Fig. S3). Our preliminary result indicated that singleplex detection had prediction accuracies of >88 % for *Aa*, >85 % for *Pg*, and >99 % for *Sm*, whereas multiplex detection had a lower accuracy of 80.6 %. The reduced detection accuracy is primarily attributed to the complexity of the saliva matrix, which contains other microbial species, proteins, enzymes, metabolites, among others, leading to more intricate Raman spectra. Additionally, 10 s spectra were used in this exploratory work to expedite the detection process. With established prediction models, considering the time required for data input and computation, the results for detecting target species from 500 samples/spectra can be delivered in 15 s. Therefore, the speed of detection depends on the time needed for spectral acquisition in addition to biopsy collection and processing. However, fast spectral acquisition has a negative impact on SNR, affecting the performance of ML models. To address these issues, we are actively exploring the integration of alternative neural network architectures to enhance prediction models for accurate multiplex detection of periodontal pathogens in saliva. Saliva collection is straightforward, non-invasive, and can be performed independently or at home without requiring specialized instruments or trained personnel. The presence of various biomarkers in saliva offers significant advantages for non-invasive and convenient sampling, promoting patient compliance with frequent monitoring and tracking of outcomes.

5. Conclusion

Taken together, the advanced deep learning tactic enabled the efficient application of SERS in swiftly and accurately detecting multiple periodontal pathogens within a mixture of bacteria. The sensing modality can be extended to identify various bacteria and biological substances. The same strategy can be employed in developing predictive models for detecting various biomarkers in saliva, a conveniently obtainable liquid biopsy. Our initial findings have demonstrated the practicality of this approach and inspired further investigation.

Aiming to establish a versatile and easily adaptable approach for label-free, rapid, and multiplex detection of pathogens, we have designed a platform with microchambers integrated into a microfluidic chip, tailored for SERS-based label-free detection. These

microchambers facilitate the compartmentalization of microbes. Ideally, oral bacteria are confined to one cell per compartment. In practice, the confinement ranged from zero to a few cells per compartment. The machine learning enabled multiplex detection method developed here promises pragmatic and reliable pathogen detection. It also serves as a foundation for creating an on-chip salivary sensor, which will enable regular and progressive monitoring of periodontal pathogens. This, in turn, will assist clinicians in making informed treatment decisions at the time of consultation and enable earlier clinical interventions to reverse the course of the disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This research was funded by the National Institute of Dental and Craniofacial Research and the Office of the Director of the National Institutes of Health under Award Number R01DE031832.

Data availability

Data will be made available on request.

References

[1]. Nazir M, Al-Ansari A, Al-Khalifa K, Alhareky M, Gaffar B, Almas K, Global prevalence of periodontal disease and lack of its surveillance, Sci. World J. 2020 (2020), 2146160.

[2]. Metwalli KH, Khan SA, Krom BP, Jabra-Rizk MA, Streptococcus mutans, Candida albicans, and the human mouth: a sticky situation, PLoS Pathog. 9 (2013), e1003616. [PubMed: 24146611]

[3]. Preshaw PM, Detection and diagnosis of periodontal conditions amenable to prevention, BMC Oral Health 15 (2015) S5. [PubMed: 26390822]

[4]. Deo PN, Deshmukh R, Oral microbiome: unveiling the fundamentals, J. Oral Maxillofac. Pathol. 23 (2019) 122–128.

[5]. Willis JR, Gabaldón T, The human oral microbiome in health and disease: from sequences to ecosystems, Microorganisms 8 (2020).

[6]. Raja M, Ummer F, Dhivakar CP, Aggregatibacter actinomycetemcomitans - a tooth killer? J. Clin. Diagn. Res. 8 (2014), Ze13–16.

[7]. Wilson M, Henderson B, Virulence factors of Actinobacillus actinomycetemcomitans relevant to the pathogenesis of inflammatory periodontal diseases, FEMS Microbiol. Rev. 17 (1995) 365–379. [PubMed: 8845187]

[8]. Kumari S, Samara M, Ampadi Ramachandran R, Gosh S, George H, Wang R, Pesavento RP, Mathew MT, A review on saliva-based health diagnostics: biomarker selection and future directions, Biomed. Mater. Devices (New York, N. Y.) (2023) 1–18.

[9]. How KY, Song KP, Chan KG, Porphyromonas gingivalis: an overview of periodontopathic pathogen below the gum line, Front. Microbiol. 7 (2016) 53. [PubMed: 26903954]

[10]. Zemanick ET, Wagner BD, Sagel SD, Stevens MJ, Accurso FJ, Harris JK, Reliability of quantitative real-time PCR for bacterial detection in cystic fibrosis airway specimens, PLoS One 5 (2010), e15101. [PubMed: 21152087]

[11]. Könönen E, Gursoy M, Gursoy UK, Periodontitis: a multifaceted disease of tooth-supporting tissues, J. Clin. Med. 8 (2019) 1135. [PubMed: 31370168]

[12]. Barghouthi SA, A universal method for the identification of bacteria based on general PCR primers, Indian J. Microbiol. 51 (2011) 430–444.

[13]. Ho C-S, Jean N, Hogan CA, Blackmon L, Jeffrey SS, Holodniy M, Banaei N, Saleh AAE, Ermon S, Dionne J, Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning, Nat. Commun. 10 (2019) 4927. [PubMed: 31666527]

[14]. Miniaturized Biosensing Devices, 1 ed., 2022.

[15]. Xu L-J, Lei Z-C, Li J, Zong C, Yang CJ, Ren B, Label-free surface-enhanced Raman spectroscopy detection of DNA with single-base sensitivity, J. Am. Chem. Soc. 137 (2015) 5149–5154. [PubMed: 25835155]

[16]. Witkowska E, Łasica AM, Niciński K, Potempa J, Kamińska A, In search of spectroscopic signatures of periodontitis: a SERS-based magnetomicrofluidic sensor for detection of Porphyromonas gingivalis and Aggregatibacter actinomycetemcomitans, ACS Sensors 6 (2021) 1621–1635. [PubMed: 33792284]

[17]. Shin HJ, Lee JH, Kim YD, Shin I, Sim T, Lim D-K, Raman-based in situ monitoring of changes in molecular signatures during mitochondrially mediated apoptosis, ACS Omega 4 (2019) 8188–8195. [PubMed: 31459907]

[18]. Bahns JT, Yan F, Qiu D, Wang R, Chen L, Hole-enhanced Raman scattering, Appl. Spectrosc. 60 (2006) 989–993. [PubMed: 17002823]

[19]. Rho E, Kim M, Cho SH, Choi B, Park H, Jang H, Jung YS, Jo S, Separation-free bacterial identification in arbitrary media via deep neural network-based SERS analysis, Biosens. Bioelectron. 202 (2022), 113991. [PubMed: 35078144]

[20]. Kloss S, Kampe B, Sachse S, Rösch P, Straube E, Pfister W, Kiehntopf M, Popp J, Culture independent Raman spectroscopic identification of urinary tract infection pathogens: a proof of principle study, Anal. Chem. 85 (2013) 9610–9616. [PubMed: 24010860]

[21]. Shang L, Xu L, Wang Y, Liu K, Liang P, Zhou S, Chen F, Peng H, Zhou C, Lu Z-M, Li B, Rapid detection of beer spoilage bacteria based on label-free SERS technology, Anal. Methods 14 (2022).

[22]. M Y, Chawla K, Bankapur A, Acharya M, D'Souza JS, Chidangil S, A micro-Raman and chemometric study of urinary tract infection-causing bacterial pathogens in mixed cultures, Anal. Bioanal. Chem. 411 (2019) 3165–3177. [PubMed: 30989268]

[23]. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D, Improved protein structure prediction using potentials from deep learning, Nature 577 (2020) 706–710. [PubMed: 31942072]

[24]. Novelli-Rousseau A, Espagnon I, Filiputti D, Gal O, Douet A, Mallard F, Josso Q, Culture-free antibiotic-susceptibility determination from single-bacterium Raman spectra, Sci. Rep. 8 (2018) 3957. [PubMed: 29500449]

[25]. Guo S, Heinke R, Stöckel S, Rösch P, Popp J, Bocklitz T, Model transfer for Raman-spectroscopy-based bacterial classification, J. Raman Spectrosc. 49 (2018) 627–637.

[26]. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H, State-of-the-art in artificial neural network applications: a survey, Heliyon 4 (2018), e00938. [PubMed: 30519653]

[27]. Tsuruda K, Shimazu A, Sugai M, Detection of a single bacterial cell using a 16S ribosomal RNA-specific oligonucleotide probe designed to investigate periodontal pathogens, Oral Microbiol. Immunol. 24 (2009) 133–140. [PubMed: 19239640]

[28]. Wang R, Krishnamurthy SN, Jeong J-S, Driks A, Mehta M, Gingras BA, Fingerprinting species and strains of Bacilli spores by distinctive coat surface morphology, Langmuir 23 (2007) 10230–10234. [PubMed: 17722943]

[29]. Guo A, Shieh YC, Wang RR, Features of material surfaces affecting virus adhesion as determined by nanoscopic quantification, Colloids Surf. A Physicochem. Eng. Asp. 602 (2020), 125109.

[30]. Chi N, Lozo S, Rathnayake RAC, Botros-Brey S, Ma Y, Damaser M, Wang RR, Distinctive structure, composition and biomechanics of collagen fibrils in vaginal wall connective tissues associated with pelvic organ prolapse, Acta Biomater. 152 (2022) 335–344. [PubMed: 36055614]

[31]. Abdi H, Williams LJ, Principal component analysis, WIREs Comput. Stat. 2 (2010) 433–459.

[32]. Kamycki K, Kapuscinski T, Oszust M, Data augmentation with suboptimal warping for time-series classification, Sensors 20 (2020) 98.

[33]. Banas JA, Fountain TL, Mazurkiewicz JE, Sun K, Margaret Vickerman M, Streptococcus mutans glucan-binding protein-a affects Streptococcus gordonii biofilm architecture, FEMS Microbiol. Lett. 267 (2007) 80–88. [PubMed: 17166223]

[34]. Wen ZT, Burne RA, Functional genomics approach to identifying genes required for biofilm development by *Streptococcus mutans*, Appl. Environ. Microbiol. 68 (2002) 1196–1203. [PubMed: 11872468]

[35]. Krzyściak W, Jurczak A, Kościelniak D, Bystrowska B, Skalniak A, The virulence of Streptococcus mutans and the ability to form biofilms, Eur. J. Clin. Microbiol. Infect. Dis. 33 (2014) 499–515. [PubMed: 24154653]

[36]. Nørskov-Lauritsen N, Claesson R, Birkeholm Jensen A, Åberg CH, Haubek D, Aggregatibacter Actinomycetemcomitans: clinical significance of a pathobiont subjected to ample changes in classification and nomenclature, Pathogens (Basel, Switzerland) 8 (2019).

[37]. Chapter 4 - subgingival microbes, in: Zhou X, Li Y (Eds.), Atlas of Oral Microbiology, Academic Press, Oxford, 2015, pp. 67–93.

[38]. Noothalapati H, Sasaki T, Kaino T, Kawamukai M, Ando M, Hamaguchi HO, Yamamoto T, Label-free chemical imaging of fungal spore walls by Raman microscopy and multivariate curve resolution analysis, Sci. Rep. 6 (2016), 27789. [PubMed: 27278218]

[39]. Movasaghi Z, Rehman S, Rehman IU, Raman spectroscopy of biological tissues, Appl. Spectrosc. Rev. 42 (2007) 493–541.

[40]. Madzharova F, Heiner Z, Kneipp J, Surface enhanced hyper-Raman scattering of the amino acids tryptophan, histidine, phenylalanine, and tyrosine, J. Phys. Chem. C 121 (2017) 1235–1242.

[41]. Cheng WT, Liu MT, Liu HN, Lin SY, Micro-Raman spectroscopy used to identify and grade human skin pilomatrixoma, Microsc. Res. Tech. 68 (2005) 75–79. [PubMed: 16228983]

[42]. Farquharson S, Shende C, Inscore FE, Maksymiuk P, Gift A, Analysis of 5-fluorouracil in saliva using surface-enhanced Raman spectroscopy, J. Raman Spectrosc. 36 (2005) 208–212.

[43]. Stone N, Kendall C, Smith J, Crow P, Barr H, Raman spectroscopy for identification of epithelial cancers, Faraday Discuss. 126 (2004) 141–157 (discussion 169–183). [PubMed: 14992404]

[44]. Ruiz-Chica AJ, Medina MA, Sánchez-Jiménez F, Ramírez FJ, Characterization by Raman spectroscopy of conformational changes on guanine–cytosine and adenine–thymine oligonucleotides induced by aminooxy analogues of spermidine, J. Raman Spectrosc. 35 (2004) 93–100.

[45]. Fogarty SW, Patel II, Martin FL, Fullwood NJ, Surface-enhanced Raman spectroscopy of the endothelial cell membrane, PLoS One 9 (2014), e106283. [PubMed: 25188340]

[46]. Jeffers RB, Cooper JB, FT-surface-enhanced Raman scattering of phenylalanine using silver-coated glass fiber filters, Spectrosc. Lett. 43 (2010) 220–225.

[47]. Huang Z, McWilliams A, Lui H, McLean DI, Lam S, Zeng H, Near-infrared Raman spectroscopy for optical diagnosis of lung cancer, Int. J. Cancer 107 (2003) 1047–1052. [PubMed: 14601068]

[48]. Farquharson S, Gift A, Shende C, Inscore F, Ordway B, Farquharson C, Murren J, Surface-enhanced Raman spectral measurements of 5-fluorouracil in saliva, Molecules 13 (2008) 2608–2627. [PubMed: 18946423]

[49]. Schulz H, Baranska M, Identification and quantification of valuable plant substances by IR and Raman spectroscopy, Vib. Spectrosc. 43 (2007) 13–25.

[50]. Krafft C, Neudert L, Simat T, Salzer R, Near infrared Raman spectra of human brain lipids, Spectrochim. Acta A Mol. Biomol. Spectrosc. 61 (2005) 1529–1535. [PubMed: 15820887]

[51]. Jyothi Lakshmi R, Kartha VB, Murali Krishna C, JG RS, Ullas G, Uma Devi P, Tissue Raman spectroscopy for the study of radiation damage: brain irradiation of mice, Radiat. Res. 157 (2002) 175–182. [PubMed: 11835681]

[52]. Ishimaru Y, Oshima Y, Imai Y, Iimura T, Takanezawa S, Hino K, Miura H, Raman spectroscopic analysis to detect reduced bone quality after sciatic neurectomy in mice, Molecules 23 (2018) 3081. [PubMed: 30477282]

[53]. Cortes C, Vapnik V, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[54]. Hart T.C.a.P., Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21–27.

[55]. Loh W-Y, Classification and regression trees, WIREs Data Min. Knowl. Disc. 1 (2011) 14–23.

[56]. Breiman L, Random Forests, Mach. Learn. 45 (2001) 5–32.

[57]. Zhang H, The optimality of naive Bayes, Aa 1 (2004) 3.

[58]. Hosmer DW Jr., Lemeshow S, Sturdivant RX, Applied Logistic Regression, John Wiley & Sons, 2013.

[59]. Friedman JH, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[60]. Pilot R, Signorini R, Durante C, Orian L, Bhamidipati M, Fabris L, A review on surface-enhanced Raman scattering, Biosensors 9 (2019).

[61]. Plou J, Valera PS, García I, de Albuquerque CDL, Carracedo A, Liz-Marzán LM, Prospects of surface-enhanced Raman spectroscopy for biomarker monitoring toward precision medicine, ACS Photonics 9 (2022) 333–350. [PubMed: 35211644]

[62]. Rebrošová K, Šiler M, Samek O, Růžička F, Bernatová S, Holá V, Ježek J, Zemánek P, Sokolová J, Petráš P, Rapid identification of staphylococci by Raman spectroscopy, Sci. Rep. 7 (2017), 14846. [PubMed: 29093473]

[63]. Vaitiekūnaitė D, Snitka V, Differentiation of closely related oak-associated gram-negative bacteria by label-free surface enhanced Raman spectroscopy (SERS), Microorganisms 9 (2021) 1969. [PubMed: 34576865]

[64]. Azhar U, Ahmed Q, Ishaq S, Alwahabi ZT, Dai S, Exploring sensitive label-free multiplex analysis with Raman-coded microbeads and SERS-coded reporters, Biosensors 12 (2022) 121. [PubMed: 35200381]

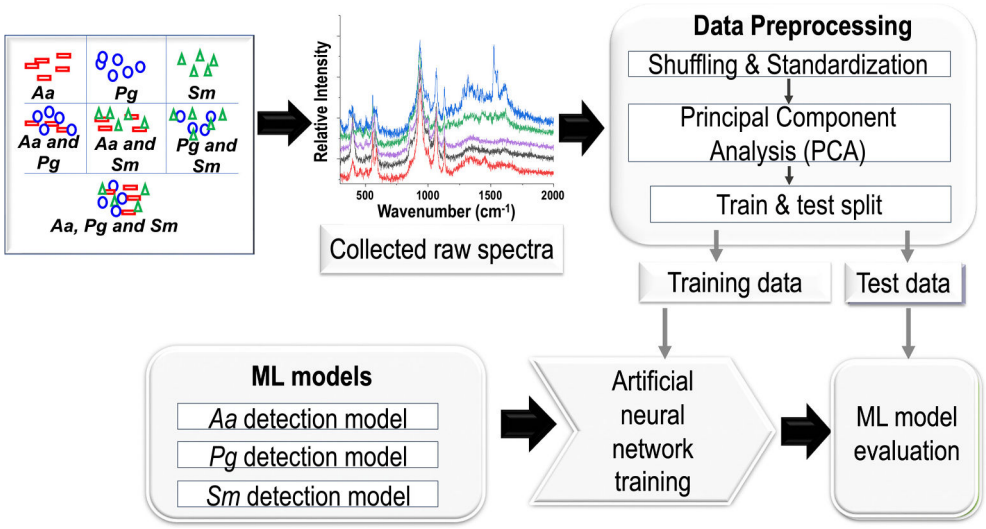

Fig. 1.
Workflow for machine learning model development.

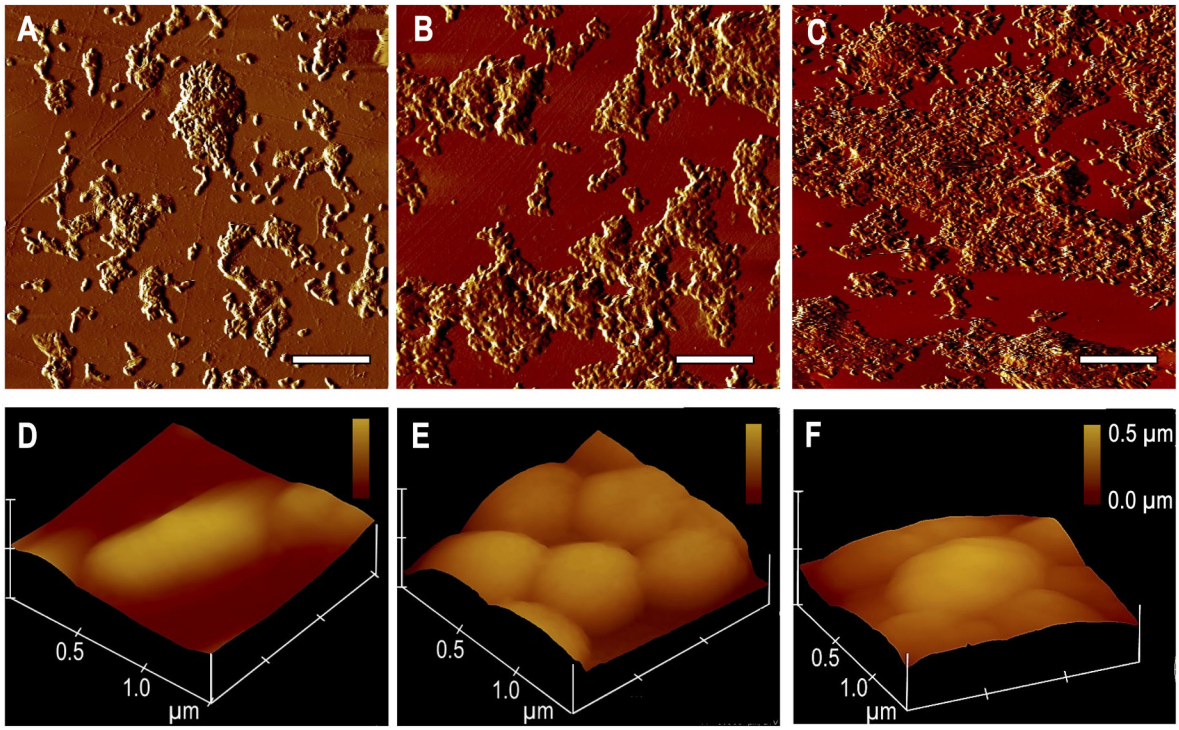

Fig. 2.

AFM images of oral bacteria on gold-coated substrates. A–C) Large-scale images illustrating the distribution of *Aa* (A), *Pg* (B) and *Sm* (C) cells. Scale bar: 10 mm. D–F) High-resolution 3D images displaying the morphology of individual *Aa* (D), *Pg* (E) and *Sm* (F) cells.

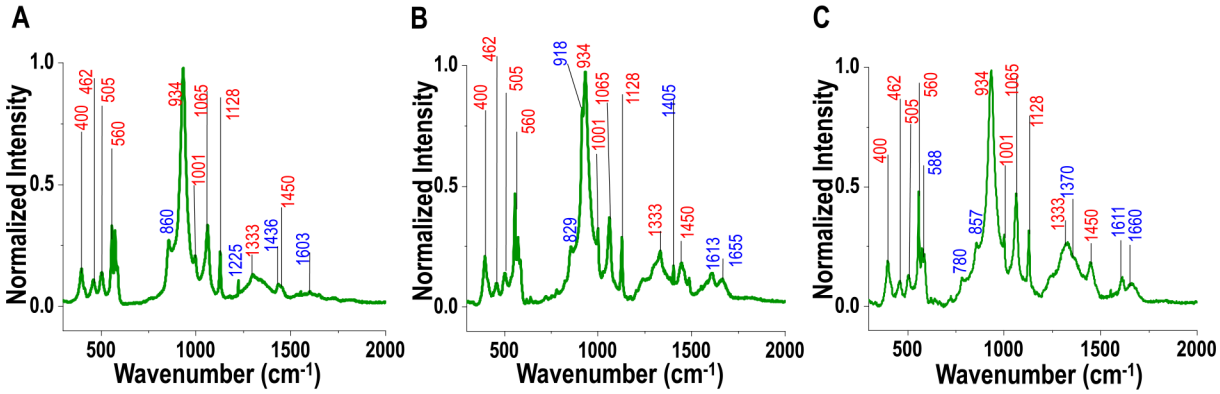

Fig. 3.

High SNR Raman spectra of *Aa* (A), *Pg* (B), and *Sm* (C) cells collected on gold-coated substrates at 80 s exposure and 10 accumulations. Peaks labeled in blue are characteristic for each species, and those in red are common among the three species. Peak assignments are shown in Table 1.

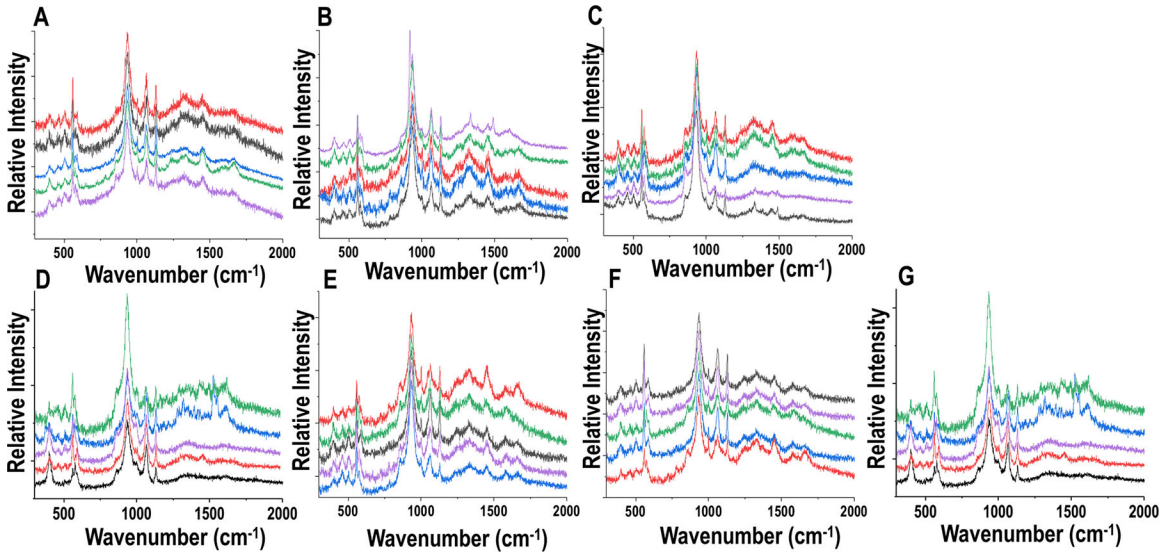

Fig. 4.

Fast SERS spectra collected at 30 s and 1 accumulation. A–C) Raw spectra of pure species of *Aa* (A), *Pg* (B), and *Sm* (C) cells collected at random locations of sample surfaces. D–G) Raw spectra of mixtures of *Aa-Pg* (D), *Aa-Sm* (E), *Pg-Sm* (F), and *Aa-Pg-Sm* (G) cells collected at random locations of sample surfaces.

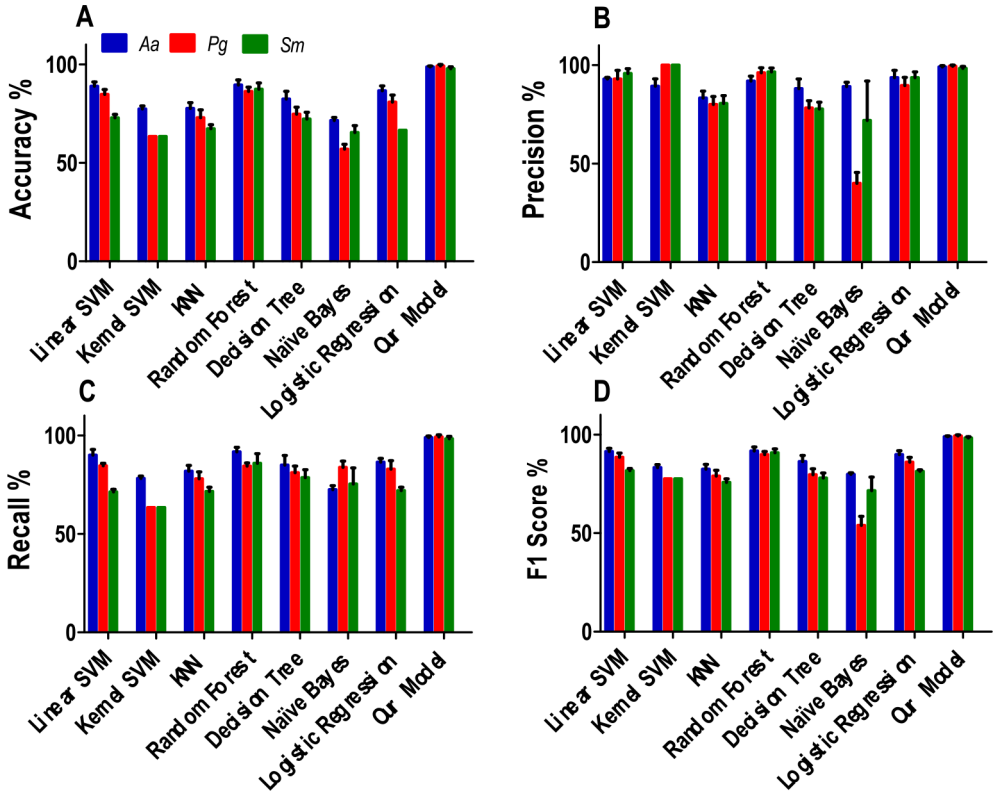

Fig. 5.

Evaluation of eight machine learning methods' overall performance for singleplex detection of *Aa*, *Pg* or *Sm* in various samples. Samples of all seven groups, including pure species, mixture of two species, or mixture of three species, were counted in the evaluation. Performances were evaluated across accuracy (A), precision (B), recall (C), and F1 score (D).

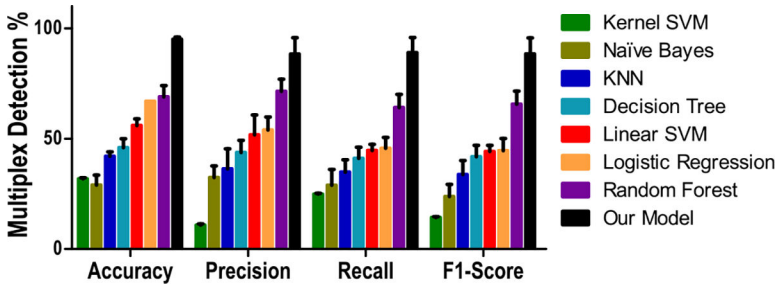

Fig. 6.

Comparison of eight machine learning methods' performance in multiplex detection. Performances were evaluated in predicting the presence or absence of all three species simultaneously based on a single spectrum. Samples of all seven groups were counted in the evaluation.

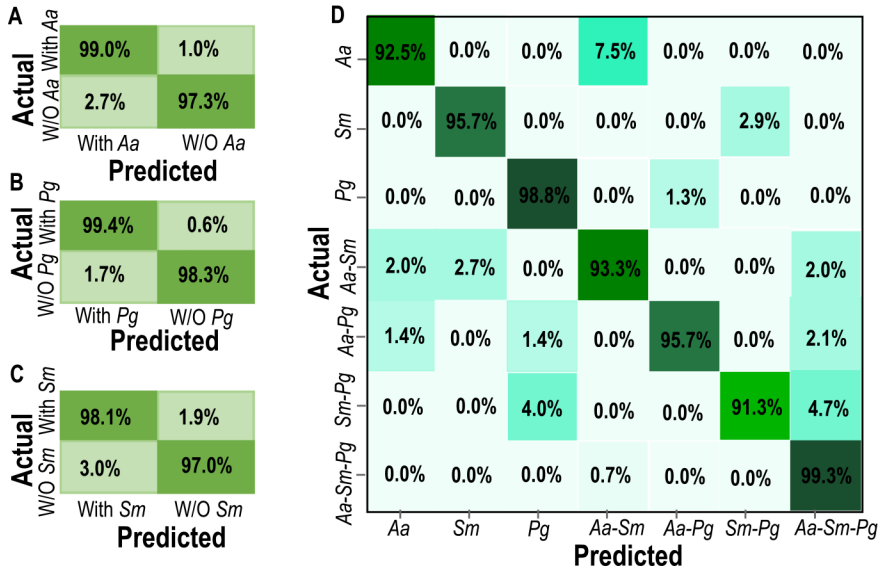

Fig. 7.

Confusion matrices on singleplex (A–C) and multiplex (D) detection of *Aa*, *Pg* and *Sm* using our method. Samples of all seven groups were counted in the evaluation. (A–C) Probabilities of true positive, false positive, true negative and false negative in identifying the presence of *Aa* (A), *Pg* (B) or *Sm* (C) in a test sample. (D) Cross-validation of cell species identification.

Table 1

SERS spectral interpretation and peak assignments.

Raman shift/cm^{-1}	Peak assignment	Species specific	Reference #
400	Polysaccharides	All 3	[38]
462	Phosphates	All 3	[39]
505	–S–S– stretching	All 3	[17]
560	C-COO$^-$ asymmetric bending of phenylalanine	All 3	[40]
588	Symmetric stretching in PO_4^{3-}	*Sm*	[41]
780	Ring breathing modes in DNA/RNA bases	All 3	[42,43]
829	Phosphodiester	*pg*	[44]
856–860	Tyrosine C-C stretching	*Aa, Sm*	[39,41]
918	C-C stretch of proline ring/glucose/lactic acid	*Pg*	[45]
934	C-COO$^-$ phenylalanine	All 3	[40]
1001	Symmetric aromatic ring breathing phenylalanine	All 3	[40,46]
1065	Skeletal stretching of C-C in phospholipids	All 3	[39]
1128	=C-O-C= in phospholipids	All 3	[16]
1225	PO^{-2} in nucleic acids	*Aa*	[47,48]
1243	Amide III	*Sm*	[49]
1333	CH_3CH_2 wagging mode in purine bases of DNA	All 3	[16]
1370	Saccharide	*Sm*	[50]
1405	COO^- in protein	*Pg*	[51]
1436	CH_2 scissoring	*Aa*	[39]
1450	CH_2 bending	*Pg*	[39]
1611–1613	Cytosine (NH_2), tyrosine	*Pg, Sm*	[39]
1655–1660	Amide I	*Pg, Sm*	[39]