Machine learning identification of periodontitis patients and their subgingival microbial features

Natalia Sánchez-Lopera[1], Manuela Restrepo-Posada[1], Sthefania Gutiérrez-González[2], David Ortega-Valderrama[2], Sandra Amaya-Sánchez[2], Adolfo Contreras[2], Andrés Felipe Gutiérrez[3], Leonardo A. Pachon[3,4], Javier Enrique Botero[1]

[1]Universidad de Antioquia, Facultad de Odontología, Medellín-Colombia. [2]Universidad del Valle, Escuela de Odontología, Cali-Colombia.[3]Universidad de Antioquia, Institute of Physics, Medellín-Colombia. [4]Guane Enterprises, Unidad I+D+I, Medellín-Colombia.

Abstract

Aim: The objective of this study was to evaluate a machine learning model to identify periodontitis patients and their cultivable subgingival microbial features.

Methods: We analyzed the composition of the subgingival microbiota in health, gingivitis and periodontitis using machine learning. A total of 1026 microbiological records comprising 73 healthy, 205 gingivitis and 748 periodontitis culture samples were analyzed for associated important features and patterns of clustering.

Results: The most frequent microorganism was *Fusobacterium spp.* followed by *Prevotella intermedia/nigrescens, Porphyromonas gingivalis, Eikenella corrodens, Tannerella forsythia*, enteric rods and *Aggregatibacter actinomycetemcomitans*. The frequency of these microorganisms was higher in periodontitis (p≤0.05). The machine learning algorithm efficiently distinguished between health and periodontitis with good performance. Age, *P. gingivalis, Fusobacterium spp.* and *P. intermedia* had a high and positive impact on the prediction of periodontitis cases followed by the presence of *A. actinomycetemcomitans, T. forsythia* and *P. micra*. There was a discernible increase in the cultivable subgingival microbiota counts from healthy to gingivitis and to periodontitis.

Conclusions: Machine learning was able to discriminate between health and periodontitis with good performance. In addition, *P. gingivalis, Fusobacterium* spp. and *P. intermedia* were important determinants in the prediction of periodontitis cases.

Keywords: subgingival microbiota; culture, periodontitis; machine learning.

Introduction

The periodontitis associated subgingival biofilm is dominated by gram-negative anaerobic rods and spirochetes, followed by facultative rods (van Winkelhof *et al.*, 2016). Of these, only few microorganisms that are cultivable, have been strongly associated with periodontitis which include *Porphyromonas gingivalis, Tannerella forsythia, Treponema denticola, Aggregatibacter actinomycetemcomitans, Fusobacterium nucleatum* and *Prevotella* species (Mdala *et al.*, 2013, Marín *et al.*, 2019). Other microorganisms of clinical relevance recovered from subgingival biofilm include *Eikenella corrodens, Selenomonas sputigena, Treponema socranskii, Campylobacter rectus*, *Capnocytophaga* spp., *Dialister pneumosintes* and *Parvimonas micra* have been also associated with periodontitis (Heller *et al.*, 2011, Hiranmayi *et al.*, 2017, Pardo-Castaño *et al.*, 2020). In addition, non-oral resident gram-negative facultative rods, such as Enterobacteriaceae, Pseudomonaceae and Acinetobacter, have been recovered in high proportions from subgingival samples from periodontitis lesions as compared to healthy subjects (Botero *et al.*, 2007, van Winkelhoff *et al.*, 2016). In contrast, bacteria of the genera *Actinomyces* and *Streptococcus* have been associated with periodontal

Correspondence to: Javier Enrique Botero
E-mail: drjavo@yahoo.com

health (Feres *et al*., 2021). In recent years, next generation sequencing studies revealed that microorganisms of the genera *Porphyromonas, Treponema, Fusobacterium, Tannerella* and *Filifactor* are present in higher proportions in periodontitis. Furthermore, these periodontal pathogens are significantly reduced by periodontal therapy, and this is associated with a good clinical response. In fact, high throughput molecular microbial studies not only expanded the knowledge of the diversity of the subgingival microbiota but confirmed the clinical relevance of classic periodontal pathogens that were studied by culture and checkerboard DNA-DNA hybridization in suppression studies (Feres *et al*., 2021).

It is estimated that only 50% of the microorganisms that reside in different niches in the mouth are cultivable (Harper-Owen *et al*., 1999). Recent efforts in promoting culture-based analyses by Lagier *et al.* (2017), termed ¨culturomics¨, have expanded the repertoire of cultured microorganisms including amoeba and giant viruses. They reported the culture of 329 new bacterial species and 327 isolated for the first time from human samples (Lagier *et al*., 2017). Recently, high throughput molecular microbial analyses have allowed the study of the oral microbiome exposing a high degree of microbial diversity (Carda-Diéguez *et al*., 2019, Feres *et al*., 2021). Nevertheless, these modern molecular techniques do not offer consistent proof of microbial viability and depending on the platform of analysis, PCR bias, data error and expensive / extensive bioinformatics analyses may result in significant limitations and yet, our current approach for treating periodontitis patients is still the same (Rezasoltani *et al*., 2020, Feres *et al*., 2021). In the biomedical field, microbial culture is considered the ¨gold standard¨ for the detection of clinically relevant pathogens and its association with the clinical diagnosis of certain diseases such as bacterial meningitis (Laupland and Valiquette 2013). Several advantages of bacterial culture over molecular techniques are that it allows for the recovery of viable cells which are a source for antimicrobial testing, complete subtyping, expression of antigens and further genomic analysis (Teles *et al*., 2013). Therefore, culture-based analyses are complementary to high throughput molecular methods for the study of the human subgingival microbiota.

Machine learning is a method of data analysis that involves the study of computer algorithms that improve automatically with experience. Machine learning creates mathematical models with the information extracted from the dataset (training data) and then makes predictions or decisions without being programmed to do. This form of data analysis has been used to solve different problems in the biomedical field including metagenomics, taxonomic profiling, and gene prediction (Chen *et al*., 2018). To the extent of our knowledge, machine learning analysis has not been used to study the cultivable subgingival microbiota in periodontitis patients. Therefore, the objective of this study was to evaluate a machine learning model to identify periodontitis patients and their cultivable subgingival microbial features.

Materials and Methods

Study design and sample

A retrospective analysis of the results from subgingival microbial cultures from periodontitis, gingivitis and healthy patients was carried out between 2003-2019. The databases from two oral microbiology laboratories, Universidad del Valle (Cali-Colombia) and Universidad de Antioquia (Medellín-Colombia), were screened for the inclusion criteria for healthy, gingivitis and periodontitis subjects. These laboratories serve as reference centers for their respective dental school clinics that provide health services to the community. The study was approved by the institutional review board (18-2019) and conducted according to the Declaration of Helsinki of 1975, as revised in 2013.

Selection criteria

The identification of potential results was based on the reported diagnosis at the time of subgingival sampling registered in the database. The healthy group included all results identified as healthy subjects with probing depths ≤3mm, no periodontal attachment loss and no bleeding on probing at sampled sites. The gingivitis group included all diagnoses reported as marginal gingivitis and plaque induced gingivitis and presented probing depths ≤3mm with positive bleeding on probing but no periodontal attachment loss at sampled sites. The periodontitis group included all forms of used terms within the date range which included: adult periodontitis, chronic periodontitis, aggressive periodontitis, juvenile periodontitis, rapidly progressing periodontitis, and refractory periodontitis. Periodontitis patients presented at least 2 non-adjacent sampled pockets depths ≥5 mm, positive bleeding on probing and periodontal attachment loss. All other forms of diagnosis such as endodontic lesions, abscesses, peri-implant lesions, osteomyelitis, and necrotizing forms of periodontal disease were excluded. All reports that included saliva samples from healthy, gingivitis and periodontitis subjects or were incomplete, were also excluded from the analysis.

Microbial detection

Microbial detection was performed by means of anaerobic cultures of subgingival plaque samples as described by Botero *et al.* (2007) and Martínez-Pabon *et al.* (2010). Briefly, subgingival plaque samples from the 5 deepest sites were taken using sterile paper points inserted to the bottom of the sulcus / pocket for 30 seconds and pooled in a vial containing Viability Medium

Götenborg Anaerobical (VMGA) III transport medium. All samples were processed within 24 hours and incubated in CO_2 using TSBV agar (trypticase-soy with serum, bacitracin, and vancomycin) as selective media for *A. actinomycetemcomitans*, and in anaerobic culture jars using Brucella blood agar (supplemented with 5% defibrinated sheep blood, hemin and K_1 vitamin). Microbial identification of *Campylobacter* spp., *Eubacterium* spp., *Fusobacterium* spp., *Capnocytophaga* spp., *D. pneumosintes*, *A. actinomycetemcomitans*, *P. gingivalis*, *Prevotella intermedia/nigrescens, Prevotella melaninogenica, T. forsythia, P. micra* and *E. corrodens* were based on colony morphology and using standard biochemical tests (catalase, CAAM, 4-Methylumbelliferyl-ß-D Glucuronide) and commercial micromethod system (RapID ANA II, Remel, Norcross, GA). Gram-negative enteric rods were subcultured and colony purified on MacConkey and cetrimide agar plates and identified using a standardized biochemical test (API 20E, BioMerieux, Marcy l'Etoile, France).

Since the data for this study was retrieved from two laboratories, it was tabulated separately due to different quantification strategies. The total viable counts were enumerated and expressed either as a percentage or colony forming units (CFU) in Cali and Medellín, respectively.

Data analysis

Age, sex (male, female), diagnosis (healthy, gingivitis, periodontitis) and microbial culture detection information were collected. The first analysis approach was to present the information using conventional statistical tests. Age is presented as the mean (95% confidence interval) and differences were determined with the ANOVA test. Counts of each microorganism were converted to dichotomic results (+/-) and the frequency detection established. Categorical variables were analyzed with the chi^2 test. Differences were considered statistically significant when $p \leq 0.05$. All analyses were conducted in a statistical software (IBM Corp. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp).

The second analysis approach involved machine learning models (Figure 1). The problem was cast as a classification problem among different classes and the data science cycle comprised three main standard stages: (I) preprocessing, (II) modelling and (III) evaluation, each of which are discussed below.

(I) Preprocessing: Since there were no missing values in the dataset, there was no need for implementing any imputation methodology. The dataset was unbalanced (few samples for two classes); however, based on the clinical criterion, no data augmentation methodology was implemented. For the data exploration phase, the dataset was normalized according to the standard scaler methodology implemented in the Scikit Learn Library (Scikit) and projected onto a two-dimensional manifold on the basis of the Uniform Manifold Approximation and Projection (UMAP) dimension reduction methodology and the Density-Based Spatial Clustering of Applications with Noise Algorithm segmentation methodology (DBSCAN). Since classes were unbalanced, the classification problem was designed to distinguish among healthy and not healthy patients. The index of the cluster obtained by means of the DBSCAN methodology was utilized as an additional variable in the dataset for the classification problem, albeit no quantifiable impact of the results was observed (Figure 1).

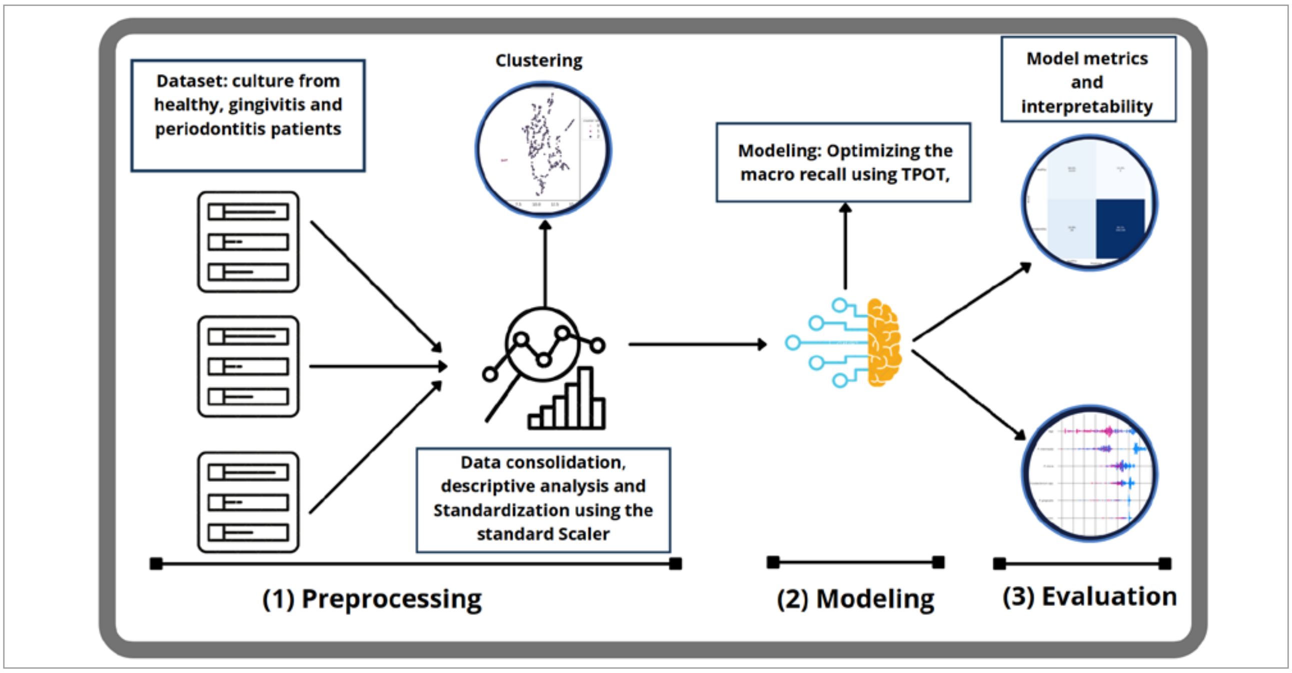

Dataset: culture from healthy, gingivitis and periodontitis patients
Clustering
Model metrics and interpretability
Modeling: Optimizing the macro recall using TPOT,
Data consolidation, descriptive analysis and Standardization using the standard Scaler
(1) Preprocessing
(2) Modeling
(3) Evaluation

Figure 1. Workflow of machine learning steps and analysis. The algorithm development comprised 3 stages: (I) preprocessing, (II) modelling and (III) evaluation.

(II) Modeling: The uncertainty on which classification methodologies should be implemented was reduced by means of the TPOT Auto Machine Learning Library (TPOT) which, once a figure-of-merit is defined, provides the best methodology (among all the one implemented in the Scikit Learn Library) and its best parameters for a particular dataset. For the present case, the selected figure-of-merit was the macro recall (average of recall per class). According to the TPOT Library, the best data processing pipeline was: exported_pipeline = make_pipeline (StackingEstimator(estimator=GaussianNB()), PolynomialFeatures(degree=2, include_bias=False, interaction_only=False), BernoulliNB(alpha=0.1, fit_prior=True)).

(III) Evaluation: To evaluate the performance of the model, the cross-validation methodology implemented in the Scikit Learn Library was utilized and presented in receiver operating characteristics (ROC) curves (Figure 1). Then, a tree-explainer was used to compute local explanations (SHapley Additive exPlanations-SHAP values) based on the associated features from each case. The mean absolute SHAP values were calculated, and the overall feature importance presented in a graph. For a final analysis, to determine patterns in the composition of the subgingival microbiota, clustering methods, supervised learning, and model interpretability were applied using Uniform Manifold Approximation and Projection (UMAP) and Density-Based Spatial Clustering of Applications with Noise Algorithm (DBSCAN) clustering analysis. Python 3.9.5 was used for machine learning modeling and analysis.

Results

A total of 1725 reports were screened, and 1026 subgingival samples were considered for analysis after exclusions. Of these, 73 (7.11%) healthy 748, 205 (19.9%) gingivitis and (73%) were periodontitis subjects. Healthy and gingivitis subjects were significantly younger than periodontitis subjects (mean age 40.5 years old) and the majority were female in each group ($p \leq 0.05$). Fewer than 3% of subjects in each group reported smoking (Table 1).

The most frequent microorganism was *Fusobacterium* spp. followed by *P. intermedia/nigrescens, P. gingivalis, E. corrodens, T. forsythia*, enteric rods and *A. actinomycetemcomitans*. In the same manner, the frequency of these microorganisms was lower in healthy subjects, increased in gingivitis and presented the highest values in periodontitis subjects ($p \leq 0.05$). In contrast, the frequency detection of *Campylobacter* spp., *P. micra, Eubacterium* spp., *D. pneumosintes*, yeasts, β-hemolytic streptococci, *P. melaninogenica* and *Capnocytophaga* spp. were similar between groups (Table 2).

Table 1. Demographic characteristics of the study sample

Variable		Healthy	Gingivitis	Periodontitis	p value
Number of subjects		73	205	748	NA
Age; mean (95% CI)		28.6 (26.5-30.6)	29.1 (26.8-31.3)	40.5 (39.4-41.7)	a**
Sex	Female	45 (61.6%)	179 (87.3%)	500 (66.8%)	b***
	Male	28 (38.4%)	26 (12.4%)	248 (33.2%)	

NA: not applicable. NS: not significant. (a) ANOVA test. (b) Chi^2. ***p≤0.0001. p**≤0.001. p*≤0.05.

Table 2. Frequency detection of cultured microorganisms.

Microorganism	Healthy	Gingivitis	Periodontitis	p value
Fusobacterium spp.	25 (34.2%)	133 (64.8%)	593 (79.2%)	a***
P. intermedia/nigrescens	9 (12.3%)	77 (37.5%)	408 (54.5%)	a***
P. gingivalis	1 (1.36%)	61 (29.7%)	405 (54.1%)	a***
E. corrodens	6 (8.21%)	50 (24.3%)	219 (29.2%)	a**
T. forsythia	5 (6.84%)	23 (11.2%)	195 (26.0%)	a***
Enteric rods	10 (13.6%)	22 (10.7%)	160 (21.3%)	a*
A. actinomycetemcomitans	2 (2.73%)	29 (14.1%)	127 (16.9%)	a*
Campylobacter spp.	10 (13.6%)	37 (18.04%)	118 (15.7%)	NS
P. micra	7 (9.58%)	27 (13.1%)	104 (13.9%)	NS
Eubacterium spp.	3/43 (6.9%) b	25/171 (14.6%) b	112/619 (18.0%) b	NS
D. pneumosintes	0/43 b	25/171 (14.6%) b	91/619 (14.7%) b	NS
Yeasts	1/43 (2.32%) b	24/171 (14.03%) b	73/619 (11.7%) b	NS
β-Hemolytic streptococci	1/43 (2.32%) b	6/171 (3.5%) b	25/619 (4.03%) b	NS
P. melaninogenica	0/30 b	2/34 (5.8%) b	5/129 (3.8%) b	NS
Capnocytophaga spp.	0/30 b	5/34 (14.7%) b	5/129 (3.8%) b	NS

NS: not significant. (a) Chi^2. (b) Frequency detection calculated on the available samples where indicated. ***p≤0.0001. p**≤0.001. p*≤0.05.

Figure 2 shows the confusion matrix from the algorithm training used in machine learning. We trained the model to classify the cases into healthy and periodontitis according to the demographic and microbial variables (frequency). Gingivitis cases were not considered in the model due to its similarity with periodontitis in terms of the variables. After the algorithm was deployed, it was able to discriminate with good performance between healthy cases (84.6%) and periodontitis (89.2%). The distinction of healthy cases was lower than periodontitis because the number of samples from healthy individuals was very limited. Furthermore, after the algorithm was trained, it was tested with several subsamples which yielded a good performance for the classification of healthy and periodontitis cases (AUC 0.94 ± 0.00; Figure 3). Gingivitis cases were excluded from the analysis since the prediction algorithm was not able to differentiate them from periodontitis cases. Important features that help predict periodontitis are presented in Figures 4 and 5. After the prediction algorithm was modeled, trained, tested, and applied to the sample, SHAP Values (SHapley Additive exPlanations) were calculated to describe the prediction made by the machine and assess the impact of important features. The y-axis indicates the variable name and in order of importance from top to bottom. The x-axis indicates the SHAP value with positive or negative associations. The color indicates the value of the analyzed feature (variable) as high or low. SHAP values for age, *P. gingivalis*, *Fusobacterium* spp. and *P. intermedia* had a high and positive impact on the prediction of periodontitis cases followed by the presence of *A. actinomycetemcomitans*, *T. forsythia* and *P. micra*. This means that the higher the age and counts of these microorganisms, the higher the association with periodontitis. The opposite occurred for healthy cases, in which the age and counts of microorganisms were lower, and this had a high association with health.

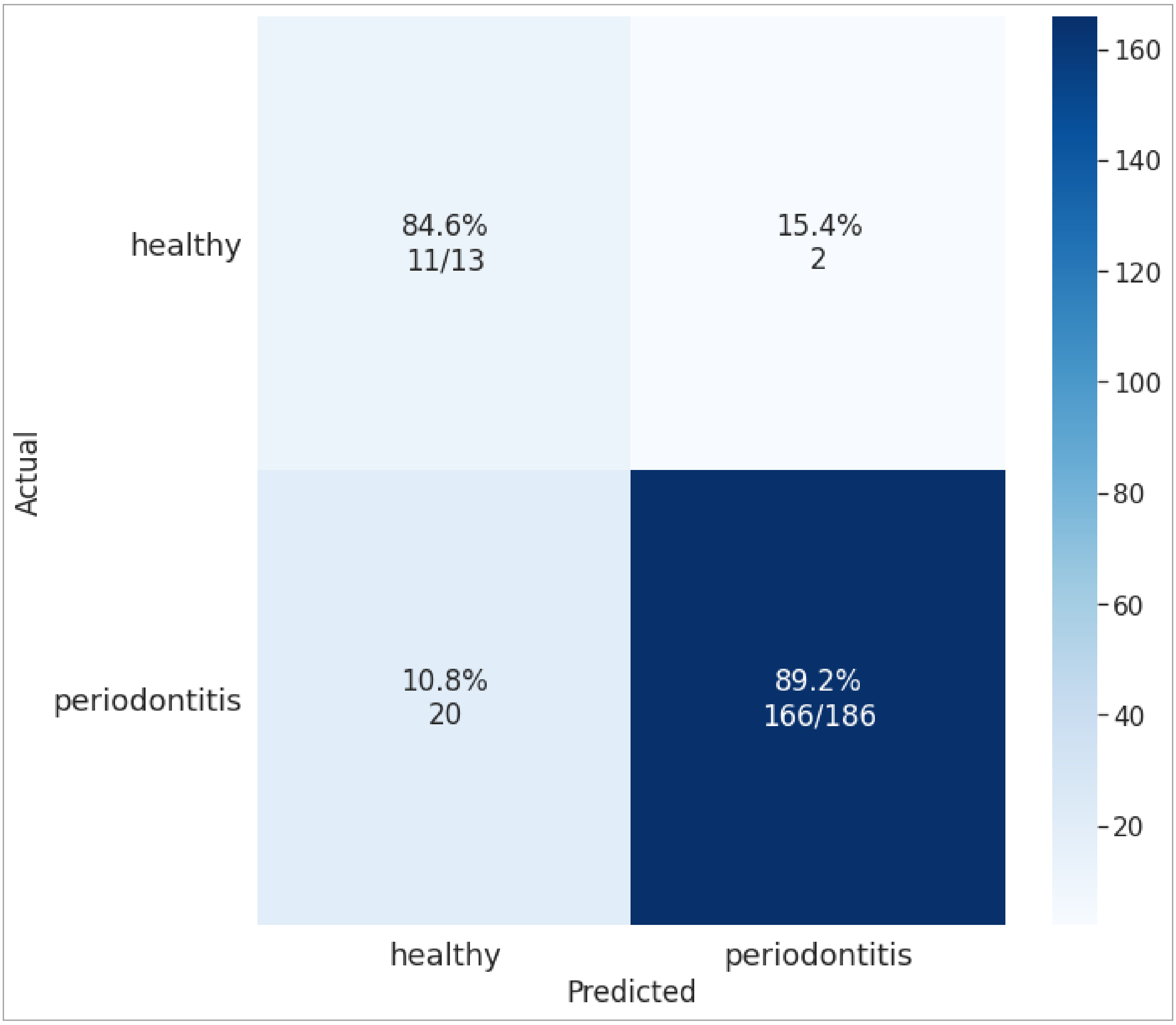


Figure 2. Confusion matrix for the prediction of periodontitis and healthy cases. After the algorithm was deployed it was able to discriminate with good performance between healthy cases (84.6%) and periodontitis (89.2%).

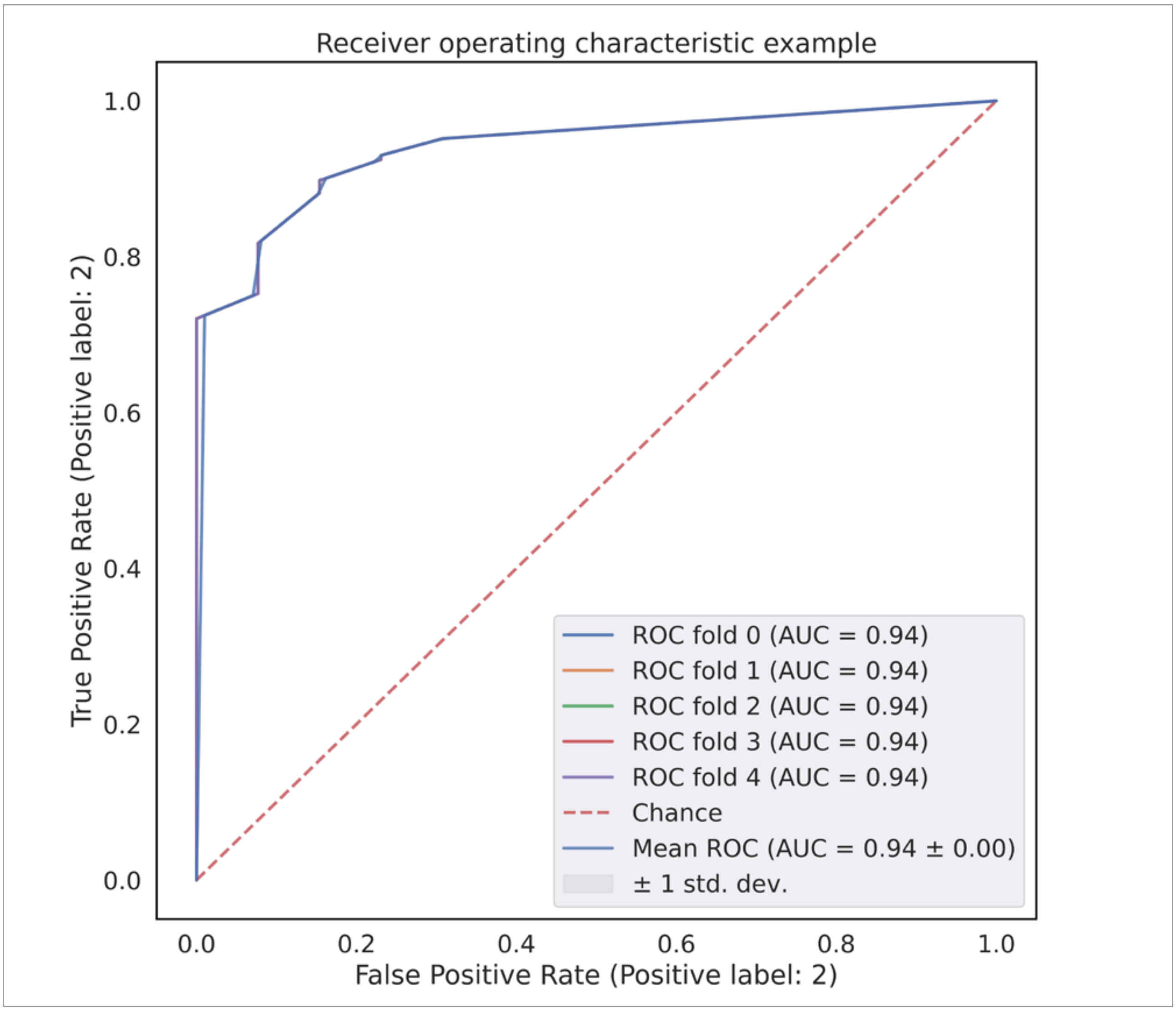


Figure 3. Receiver operating characteristic analysis of the algorithm used in machine learning. After the algorithm was trained, it was tested with several subsamples which yielded a good performance for the classification of healthy and periodontitis cases (AUC 0.94 ± 0.00).

The composition of the subgingival microbiota from healthy, gingivitis and periodontitis individuals are depicted in supplementary material 1. There is a discernible difference in the areas corresponding to an increase in the cultivable subgingival microbiota from healthy to gingivitis and to periodontitis. While gingivitis and periodontitis showed similar microbiotas, the proportions of *P. gingivalis, P. intermedia, T. forsythia, E. corrodens* and *D. pneumosintes* were higher in periodontitis subjects. *Fusobacterium* spp., *P. micra* and enteric rods were in similar proportions in all groups. The transition from health to gingivitis and then to periodontitis was characterized by increasing counts of *P. gingivalis* while *D. pneumosintes* was exclusive of gingivitis and periodontitis cases.

The DBSCAN returned 3 valid clusters based on the similarities between the subgingival microbial profiles of the subjects included in each cluster (supplementary material 2, 3 and 4). The highest density cluster (cluster 2) included 4.4% of healthy, 26.2% of gingivitis and 69.4% of periodontitis subjects. Then followed cluster 0 which contained 7.5% of healthy, 7% of gingivitis and 85.5% of periodontitis subjects. The lowest density cluster (cluster 1) comprised 4.5% of healthy, 8.6% of gingivitis and 86.9% of periodontitis subjects. The subgingival microbial profile of cluster 2 was composed by higher counts of *Fusobacterium* spp (4.9%), *P. gingivalis* (3.1%), *P. intermedia* (3%) and *T. forsythia* (1%). Cluster 0 presented a subgingival microbial profile composed by *Fusobacterium* spp (3.8%), *P. intermedia* (3.3%), enteric rods (2.7%), *P. gingivalis* (2.4%) and *D. pneumosintes* (1%). In contrast, the subgingival microbial profile of cluster 1 was dominated by the highest counts of enteric rods (67.5%) followed by *Fusobacterium* spp. (2%) and *P. gingivalis* (1.6%) (supplementary material 4).

P. gingivalis, T. forsythia and *P. intermedia* were considered the independent variables in the model and the effect on other microorganisms was studied in

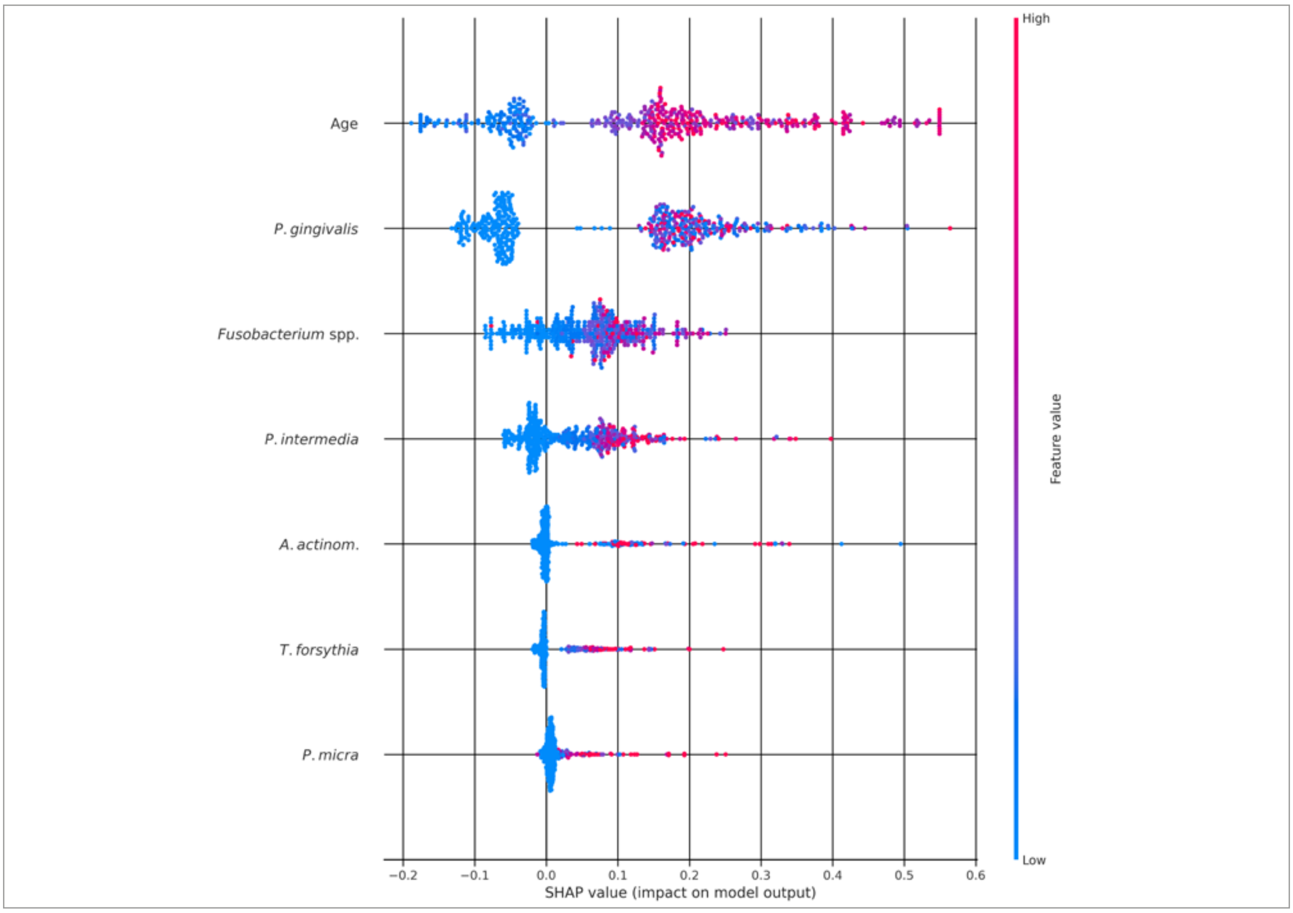


Figure 4. Feature analysis of the prediction of periodontitis cases. SHAP: SHapley Additive exPlanations.

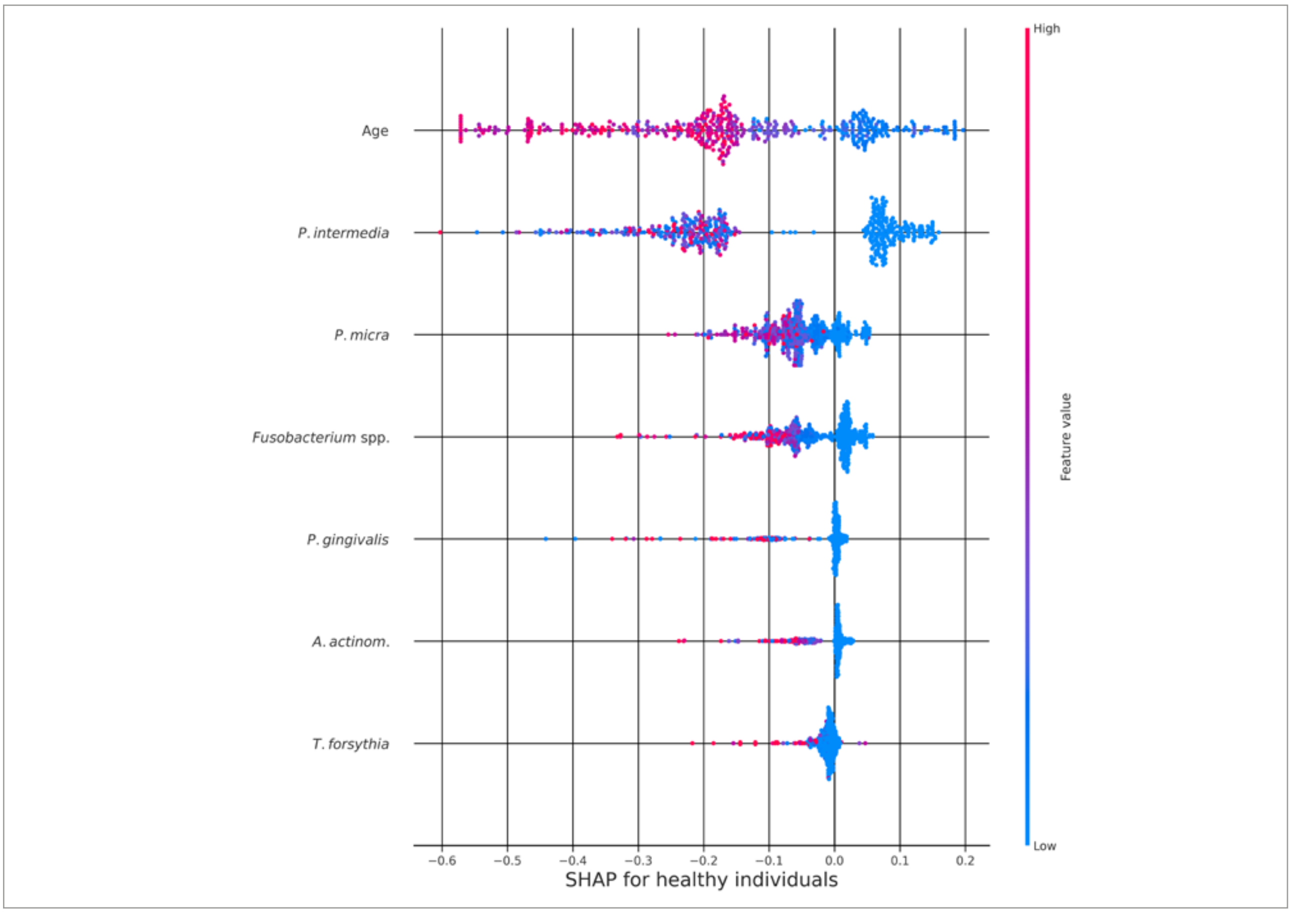


Figure 5. Feature analysis of the prediction of healthy cases. SHAP: SHapley Additive exPlanations.

periodontitis samples (supplementary material 5, 6 and 7). The most abundant microorganisms in terms of cultured counts were *P. gingivalis* and *P. intermedia* when present together. It was observed that with increasing counts of *P. gingivalis* a decrease in the counts of *P. intermedia, E. corrodens* and *T. forsythia* was observed (supplementary material 5). Similarly, with increasing counts of *T. forsythia*, a decrease in the counts of *P. gingivalis* and *P. intermedia* occurred (supplementary material 6). In contrast, when *P. intermedia* counts increased, there was no clear pattern and was related with higher counts of *P. gingivalis*, *T. forsythia* and *E. corrodens* (supplementary material 7). *A. actinomycetemcomitans* counts were very low as compared to other microorganisms and no distinct association was observed.

Discussion

This study showed that machine learning was able to discriminate health from periodontitis cases. In addition, the model was able to identify the most prominent microbial features of the cultivable subgingival microbiota that are relevant to health, gingivitis and periodontitis.

It has been previously confirmed that periodontitis increases with age (Billings *et al*., 2018). In this study, healthy and gingivitis patients were significantly younger than periodontitis patients (mean age 40.5 years old). Age is not a risk factor by itself for periodontitis, but rather signifies the natural history of individuals, a history that include variable susceptibility and exposures to multiple etiological factors. However, periodontitis can occur at any age and this just illustrate the complexity of periodontal disease. During our analyses, age was an important feature for the prediction of periodontitis by the machine and this agrees with current knowledge of periodontal disease (Billings *et al*., 2018). Interesting was the finding that the machine was not able to distinguish gingivitis from periodontitis based on the microbiology even though gingivitis subjects were significantly younger. This may reflect the transition that the subgingival environment is going through during the development of periodontal disease. Pathological and microbiological changes that occur during gingivitis mark the establishment of periodontal disease that are necessary for the progression to periodontitis. However, not all cases of gingivitis result in periodontitis due to the multifactorial nature of the disease. On the other hand, pathogens such as *P. gingivalis, A. actinomycetemcomitans* and *T. forsythia* are rarely detected in young periodontally healthy individuals, even with molecular detection techniques (Lombardo *et al*., 2021). Changes in lifestyle determinants, oral hygiene, and plaque retentive factors through life influence shifts in the subgingival microbiota from health to gingivitis and then to periodontitis.

The microbial features of health and periodontitis were analyzed. First, a frequency detection analysis determined which microorganisms were more prevalent in each condition. Then, a machine learning algorithm was trained to identify healthy and periodontitis cases based on the features provided. As a result, the machine was able to distinguish healthy and periodontitis cases with good performance. A previous study by Kim *et al.* (2020) found similar prediction accuracy in saliva samples using the copy numbers of nine pathogens. In contrast, gingivitis cases were not easily distinguished from periodontitis cases by the machine and perhaps represents the continuum from health to disease in the subgingival microbial environment. The outcome determined that in addition to higher age, the most important microbiological features for periodontitis subjects were increasing counts of *P. gingivalis, Fusobacterium* spp. and *P. intermedia*. Although the counts of *A. actinomycetemcomitans*, *T. forsythia* and *P. micra* were important in the prediction model, their association was lower but still predictive of periodontitis. In contrast, healthy subjects were significantly younger, and the subgingival microbiota was readily identifiable by the machine by low counts and frequency. Previous culture-based studies showed comparable results (Sato *et al.*, 1993).

This is the first study that attempts to analyze the profiles of the cultured subgingival microbiota in periodontitis using machine learning. Clustering algorithms play an important role in machine learning applications. Any clustering method is an algorithm capable of dividing the input data into subsets in such a way that data sharing some similarity is clustered together. Typically, clustering methods are extensively used in unlabeled datasets to find associations in the input data. This feature is also useful in scenarios where few labeled samples are available. Using a clustering algorithm, a particular tag is assigned to each cluster. Such tags are used to help a supervised algorithm in finding an appropriate decision boundary in a classification task. Furthermore, to apply a cluster-then-label approach, it is necessary to perform a dimensionality reduction of the dataset. Most clustering algorithms depend on a metric measure, points belonging to a particular cluster are "close" in terms of such measure. In high dimensional spaces, the problem of dimensionality makes the concept of proximity not qualitatively meaningful (Aggarwal *et al*., 2001). To overcome such difficulty, several dimensionality reduction methods are used in machine learning tasks and can be divided into two main categories (McInnes *et al*., 2018): 1) preservation of local distances, and 2) preservation of global distances. In the first category, algorithms such as multiple correspondence analysis (MCA) or the principal component analysis (PCA) are widely used. Both methods generate projections into a set of axes that preserves the variance in the original dataset. In the second category, T-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and

Projection (UMAP), are state of the art algorithms for dimensionality reduction, preserving the global structure of the original dataset, showing high performance in several supervised and unsupervised tasks (McInnes *et al.*, 2018, Maaten and Hinton 2008). The motivation of using a cluster-then-label approach is to imprint features that can help to improve the performance of a machine learning classifier by considering relationships in the abundance of certain pathogens of the subgingival microbiota. In the case of the subgingival microbiota, this imbalance can lead to different inflammatory changes in the periodontal tissues. To perform this goal, the dimensionality reduction is performed on the entire dataset, without the diagnosis information. The final stage of the cluster-then-label approach is the choosing of a clustering algorithm. In the case of the present work, a density-based approach, the Density-Based Spatial Clustering of Applications with Noise Algorithm (DBSCAN) was used due to the resilience in presence of noisy data (Campello *et al.*, 2013). The DBSCAN algorithm returned 3 clusters that presented similarities in the subgingival microbiota. While the highest density cluster consisted mainly of periodontal pathogens *Fusobacterium* spp., *P. gingivalis*, *P. intermedia* and *T. forsythia* in most periodontitis patients, the lowest density cluster was dominated by the highest levels of enteric rods and only presented in a small group of subjects with periodontitis and few healthy subjects which is a rare finding. *Fusobacterium* spp. has been shown to play an important coaggregation role for the colonization of *P. gingivalis* and *T. forsythia* (Saito *et al.*, 2008, Thurnheer *et al.*, 2019). In addition, studies have shown higher frequency and counts of uncommon subgingival microorganisms such as enteric rods in periodontitis patients that complement the dysbiotic oral ecology (Chen *et al.*, 1997, Herrera *et al.*, 2008, Ranganathan *et al.*, 2017). It is possible that the presence of high counts of enteric rods represent a transitory colonization in different oral surfaces such as saliva, tongue, and oral mucosa. Our clustering analysis showed that similar patterns of subgingival periodontal pathogens colonization can occur in some healthy, gingivitis and periodontitis subjects, indicating that the simple presence of periodontal pathogens does not necessarily mean periodontitis and in turn, shows how complex the interactions between the biofilm (i.e.: co-aggregation, microbial succession) and host (i.e.: immune response, smoking) are for the clinical development of periodontitis (Teles *et al.*, 2013). However, the difference between health and periodontitis was clearly determined by the clustering algorithm by the increase in the proportions of the identified species, which is in agreement with previous studies (Abusleme *et al.*, 2013). The classic role of *P. gingivalis*, *Fusobacterium* spp., *T. forsythia* and *P. intermedia* in periodontitis was further corroborated by our results.

A shift in the microbial counts and frequency from health to disease was identified by the machine. Inflammatory changes that occur within periodontal tissues in response to an imbalance in the subgingival microbiota results in the formation of the periodontal pocket (Meuric *et al.*, 2017). This clinical feature has been associated with changes in the subgingival microbiota and therefore have an impact on the cultivable counts (Pérez-Chaparro *et al.*, 2018). Although a correlation between probing depth and cultivable counts was not determined, periodontitis patients by selection had increased probing depths (>5mm) as compared to healthy samples (<3mm) and therefore suggest an altered subgingival environment. The definition of a cultivable healthy and disease microbiota is still difficult. Further microbiological studies are required for a better understanding of the subgingival microbiota and its relationship with the host.

This study presented a novel approach to the analysis of the culturable subgingival microbiota. But it is important to understand that the main limitation of culture-based analysis is the number of microorganisms that can be studied as compared to high-throughput genomic methods. Consequently, diversity was limited, and it only focused on the most cultured subgingival microorganisms including the most recognized periodontal pathogens and enteric rods which are not commonly analyzed in high-throughput genomic methods. Nonetheless, machine learning analysis showed that with a limited group of microorganisms, it was possible to make associations between clusters of microorganisms and periodontitis based on the most prominent features and subgingival microbial profiles. The results are comparable to previous studies (Pérez-Chaparro *et al.*, 2018) and highlight the usefulness of culture-based approaches to study the implications of the subgingival microbiota in periodontitis.

This study has some limitations. Since it was a retrospective study of samples from patients with diverse periodontal conditions, it used the data stored in each laboratory. This information was restricted to identification of the patient, diagnosis, microbial identification, and sampled sites clinical parameters. Therefore, complete periodontal parameters regarding stage/grade of periodontitis, known risk factors such as diabetes and smoking were not available. Future prospective studies using machine learning that include greater demographic, microbiological and clinical variables would produce a more experienced model that could help explain the interactions between the microbiota and the host. The next step would be the use of artificial intelligence to develop systems that accurately predict the risk of periodontitis and response to therapy based on these interactions. This could have great implications in the development of chair-side tests for the early detection of patients at risk.

Conclusions

Machine learning was able to discriminate between health and periodontitis with good performance. In addition, *P. gingivalis*, *Fusobacterium* spp. and *P. intermedia* were important determinants in the prediction of periodontitis cases. Culture-based analysis of the subgingival microbiota could help identify individuals at risk of developing periodontitis and remains essential for the study of the subgingival microbiota.

References

Abusleme L, Dupuy AK, Dutzan N, *et al*. The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *ISME Journal* 2013; **7**: 1016-1025.

Aggarwal CC, Hinneburg A, & Keim DA. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: Van den Bussche J., Vianu V. (eds) Database Theory — ICDT 2001. ICDT 2001. Lecture Notes in Computer Science, vol 1973. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44503-X_27.

Billings M, Holtfreter B, Papapanou PN, Mitnik GL, Kocher T, Dye BA. Age-dependent distribution of periodontitis in two countries: Findings from NHANES 2009 to 2014 and SHIP-TREND 2008 to 2012. *Journal of Periodontology* 2018;**89**: S140-S158.

Botero JE, Contreras A, Lafaurie G, Jaramillo A, Betancourt M, Arce RM. Occurrence of periodontopathic and superinfecting bacteria in chronic and aggressive periodontitis subjects in a Colombian population. *Journal of Periodontology* 2007; **78**: 696-704.

Campello, R J, Moulavi D, & Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Carda-Diéguez M, Bravo-González LA, Morata IM, Vicente A, Mira A. High-throughput DNA sequencing of microbiota at interproximal sites. *Journal of Oral Microbiology* 2019; **12**: 1687397.

Chen C, Ashimoto A, Sangsurasak S, Flynn MJ, Slots J. Oral food consumption and subgingival microorganisms: subgingival microbiota of gastrostomy tube-fed children and healthy controls. *Journal of Periodontology* 1997; **68**: 1163-1168.

Chen WP, Chang SH, Tang CY, Liou ML, Tsai SJ, Lin YL. Composition Analysis and Feature Selection of the Oral Microbiota Associated with Periodontal Disease. *Biomed Research International* 2018;2018: 3130607.

Duran-Pinedo AE, Paster B, Teles R, Frias-Lopez J. Correlation network analysis applied to complex biofilm communities. *PLoS One* 2011; **6**: e28438.

Feres M, Retamal-Valdes B, Gonçalves C, Cristina Figueiredo L, Teles F. Did Omics change periodontal therapy?. *Periodontology 2000* 2021; **85**: 182-209.

Harper-Owen R, Dymock D, Booth V, Weightman AJ, Wade WG. Detection of unculturable bacteria in periodontal health and disease by PCR. *Journal of Clinical Microbiology* 1999; **37**: 1469-1473.

Heller D, Varela VM, Silva-Senem MX, Torres MC, Feres-Filho EJ, Colombo AP. Impact of systemic antimicrobials combined with anti-infective mechanical debridement on the microbiota of generalized aggressive periodontitis: a 6-month RCT. *Journal Clinical Periodontology* 2011; **38**: 355-364.

Herrera D, Contreras A, Gamonal J, Oteo A, Jaramillo A, Silva N, Sanz M, Botero JE, León R. Subgingival microbial profiles in chronic periodontitis patients from Chile, Colombia and Spain.*Journal of Clinical Periodontology* 2008; **35**: 106-113.

Hiranmayi KV, Sirisha K, Ramoji Rao MV, Sudhakar P. Novel Pathogens in Periodontal Microbiology. *Journal of Pharmacy and Bioallied Sciences* 2017; **9**: 155-163.

Kim EH, Kim S, Kim HJ, Jeong HO, Lee J, Jang J, Joo JY, Shin Y, Kang J, Park AK, Lee JY, Lee S. Prediction of Chronic Periodontitis Severity Using Machine Learning Models Based On Salivary Bacterial Copy Number. *Frontiers in Cellular Infection Microbiology* 2020; **10**: 571515.

Lagier JC, Drancourt M, Charrel R, *et al*. Many More Microbes in Humans: Enlarging the Microbiome Repertoire. *Clinical Infectious Diseases* 2017; **65**: S20-S29.

Laupland KB, Valiquette L. The changing culture of the microbiology laboratory. *Canadian Journal of Infectious Diseases and Medica Microbiology* 2013; **24**:125-128.

Lombardo L, Palone M, Scapoli L, Siciliani G, Carinci F. Short-term variation in the subgingival microbiota in two groups of patients treated with clear aligners and vestibular fixed appliances: A longitudinal study. *Orthodontic and Craniofacial Research* 2021; **24**: 251-260.

Maaten, LVD, & Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008; **9**: 2579-2605.

Marín MJ, Ambrosio N, O'Connor A, Herrera D, Sanz M, Figuero E. Validation of a multiplex qPCR assay for detection and quantification of Aggregatibacter actinomycetemcomitans, Porphyromonas gingivalis and Tannerella forsythia in subgingival plaque samples. A comparison with anaerobic culture. *Archives of Oral Biology* 2019; **102**: 199-204.

Martínez-Pabón MC, Isaza-Guzmán DM, Mira-López NR, García-Vélez C, Tobón-Arroyave SI. Screening for subgingival occurrence of gram-negative enteric rods in periodontally diseased and healthy subjects. *Archives of Oral Biology* 2010; **55**: 728-736.

McInnes L, Healy J, Saul N, & Großberger L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 2018; **3**: 861.

Mdala I, Olsen I, Haffajee AD, Socransky SS, de Blasio BF, Thoresen M. Multilevel analysis of bacterial counts from chronic periodontitis after root planing/scaling, surgery, and systemic and local antibiotics: 2-year results. *Journal of Oral Microbiology* 2013;**5**: 20939.

Meuric V, Le Gall-David S, Boyer E, Acuña-Amador L, Martin B, Fong SB, Barloy-Hubler F, Bonnaure-Mallet M. Signature of Microbial Dysbiosis in Periodontitis. *Applied Environmental Microbiology* 2017; **83**: e00462-17.

Pardo-Castaño C, Vásquez D, Bolaños G, Contreras A. Strong antimicrobial activity of collinin and isocollinin against periodontal and superinfectant pathogens in vitro. *Anaerobe* 2020; **62**: 102163.

Pérez-Chaparro PJ, McCulloch JA, Mamizuka EM, Moraes ADCL, Faveri M, Figueiredo LC, Duarte PM, Feres M. Do different probing depths exhibit striking differences in microbial profiles? *Journal of Clinical Periodontology* 2018; **45**: 26-37.

Ranganathan AT, Sarathy S, Chandran CR, Iyan K. Subgingival prevalence rate of enteric rods in subjects with periodontal health and disease. *Journal of the Indian Society of Periodontology* 2017; **21**: 224-228.

Rezasoltani S, Ahmadi Bashirzadeh D, Nazemalhosseini Mojarad E, Asadzadeh Aghdaei H, Norouzinia M, Shahrokh S. Signature of Gut Microbiome by Conventional and Advanced Analysis Techniques: Advantages and Disadvantages. *Middle East Journal of Digestive Diseases* 2020; **12**: 5-11.

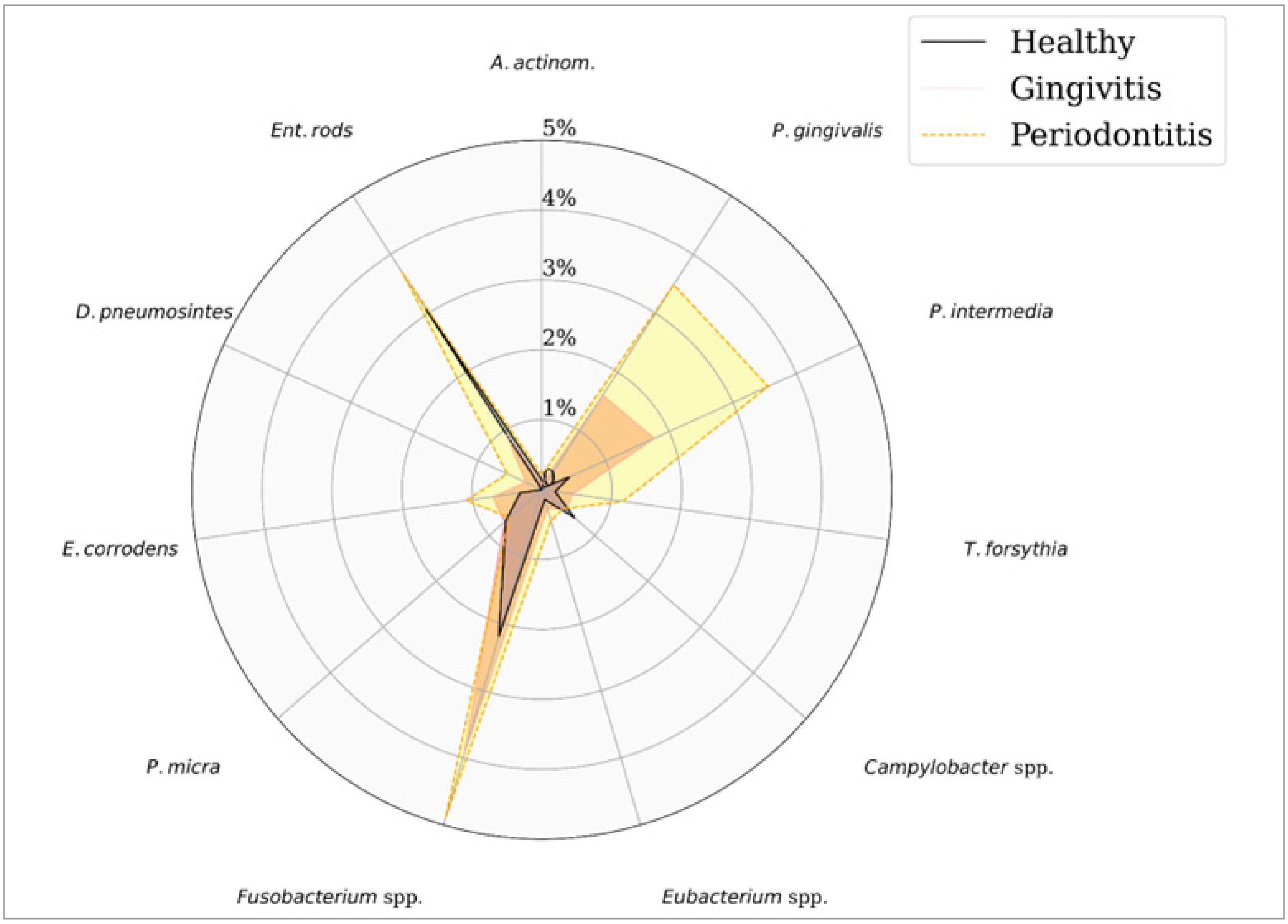
Saito Y, Fujii R, Nakagawa KI, Kuramitsu HK, Okuda K, Ishihara K. Stimulation of Fusobacterium nucleatum biofilm formation by Porphyromonas gingivalis. *Oral Microbiology and Immunology* 2008; **23**: 1-6.

Sato K, Yoneyama T, Okamoto H, Dahlén G, Lindhe J. The effect of subgingival debridement on periodontal disease parameters and the subgingival microbiota. *Journal of Clinical Periodontology* 1993; **20**:359-365.

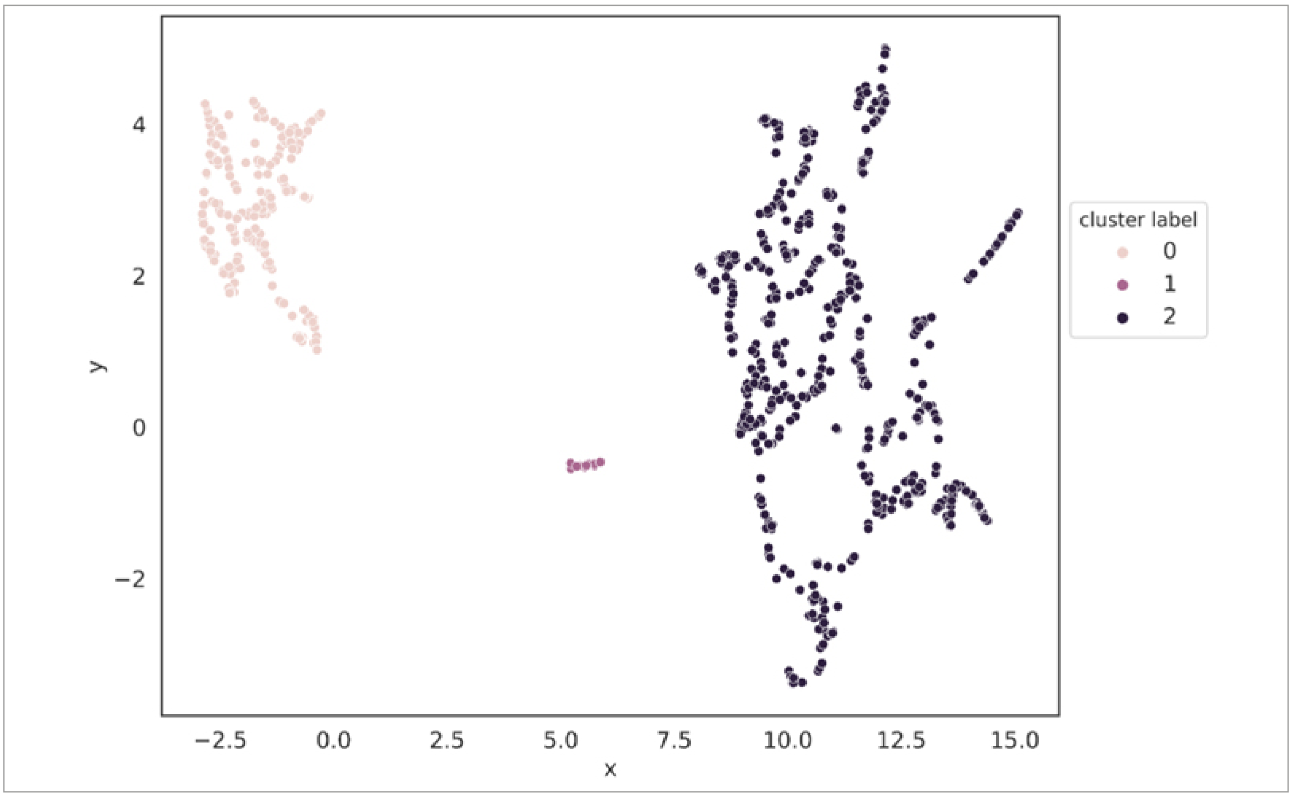
Teles R, Teles F, Frias-Lopez J, Paster B, Haffajee A. Lessons learned and unlearned in periodontal microbiology. *Periodontology 2000* 2013; **62**: 95-162.

Thurnheer T, Karygianni L, Flury M, Belibasakis GN. Fusobacterium Species and Subspecies Differentially Affect the Composition and Architecture of Supra- and Subgingival Biofilms Models. *Frontiers in Microbiology* 2019; **10**: 1716.

van Winkelhoff AJ, Rurenga P, Wekema-Mulder GJ, Singadji ZM, Rams TE. Non-oral gram-negative facultative rods in chronic periodontitis microbiota. *Microbial Pathogenesis* 2016; **94**: 117-122.



Supplementary material 1. Mean bacterial counts from healthy, gingivitis and periodontitis cases. There is a discernible difference in the areas corresponding to an increase in the subgingival microbiota from healthy to gingivitis and to periodontitis.

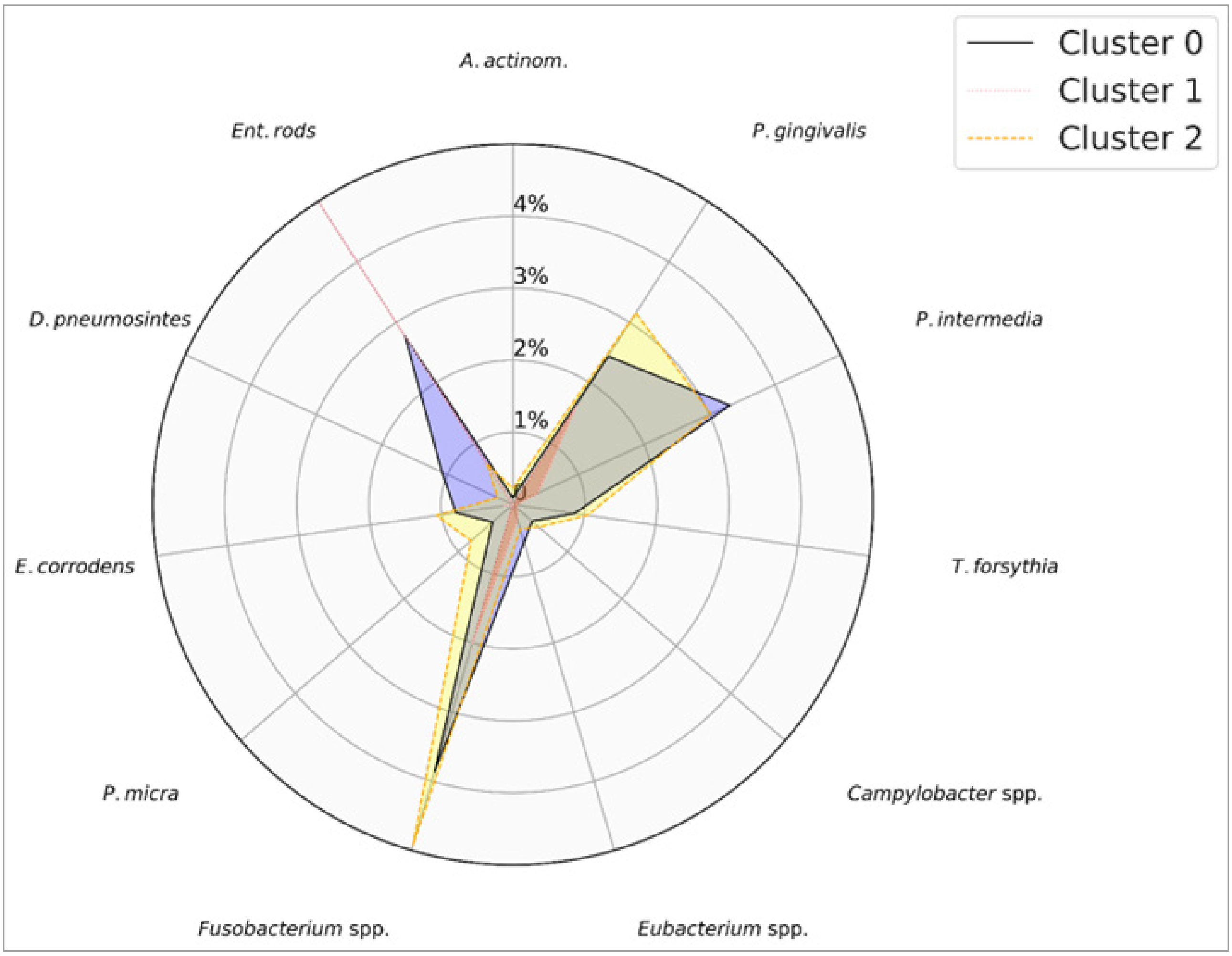


Supplementary material 2. Density-Based Spatial Clustering of Applications with Noise Algorithm (DBSCAN). Three clusters were identified based on the similarities of the composition of the subgingival microbiota.

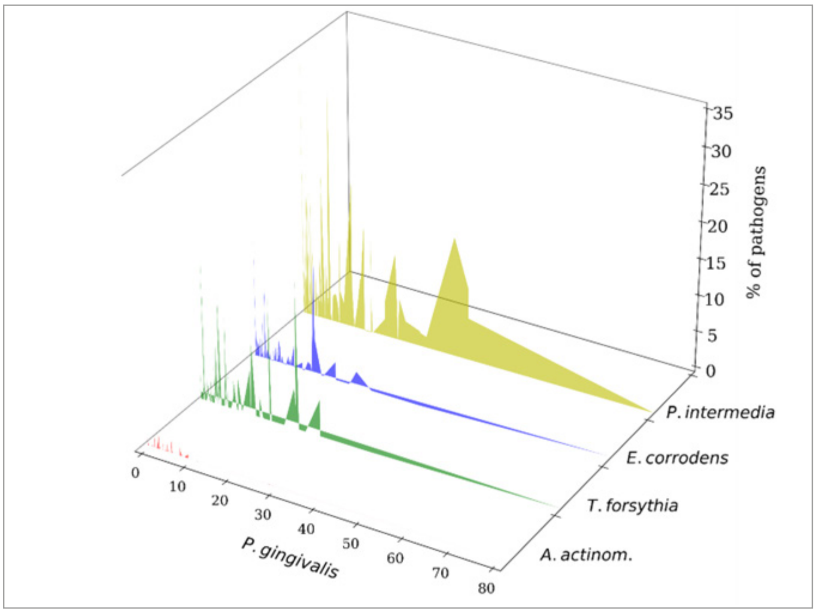
Supplementary material 3. Number of healthy, gingivitis and periodontitis subjects according to DBSCAN clustering analysis.

	Cluster 0	Cluster 1	Cluster 2
Healthy	17 (7.5%)	1 (4.5%)	25 (4.4%)
Gingivitis	16 (7.0%)	2 (8.6%)	153 (26.2%)
Periodontitis	194 (85.5%)	20 (86.9%)	405 (69.4%)

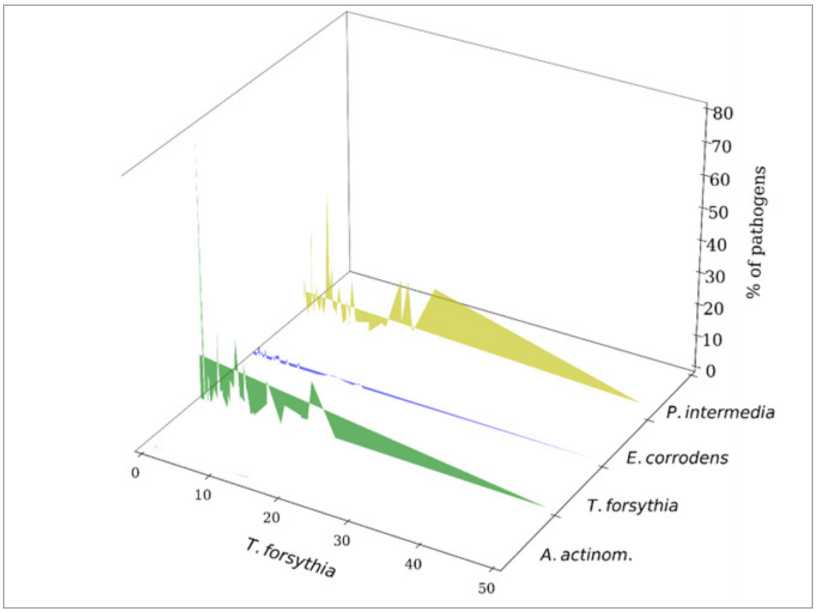
DBSCAN: Density-Based Spatial Clustering of Applications with Noise Algorithm.



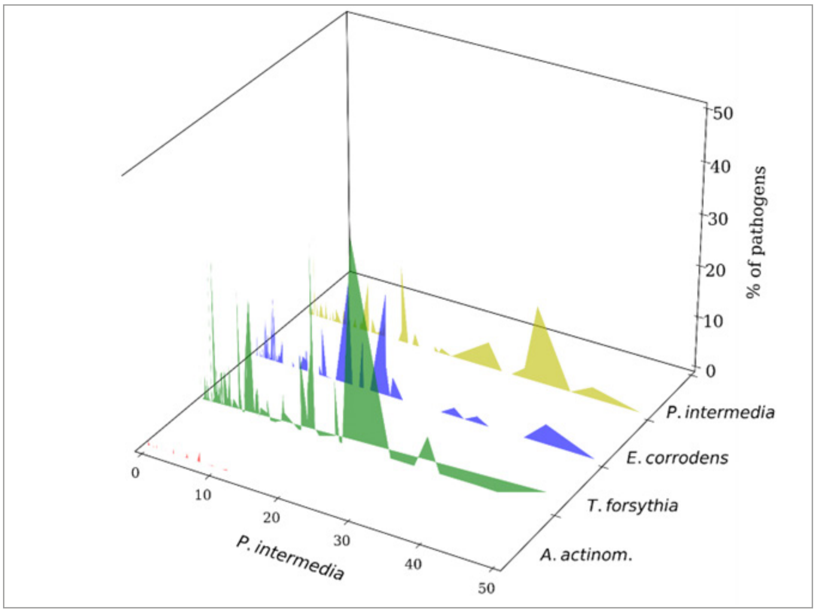
Supplementary material 4. Radar plot of identified subgingival profiles by dimensional reduction analysis and similarity. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise Algorithm) returned 3 valid clusters based on the similarities between the subgingival microbial profiles of the subjects included in each cluster.



Supplementary material 5. Relationship of P. gingivalis with other periodontal pathogens. Counts of P. gingivalis were considered the independent variable in the model and the effect on other microorganisms was studied in periodontitis samples.



Supplementary material 6. Relationship of T. forsythia with other periodontal pathogens. Counts of T. forsythia were considered the independent variable in the model and the effect on other microorganisms was studied in periodontitis samples.



Supplementary material 7. Relationship of P. intermedia with other periodontal pathogens. Counts of P. intermedia were considered the independent variable in the model and the effect on other microorganisms was studied in periodontitis samples.