

¹

Doctoral Thesis

²

Microbiota in Human Diseases

³

Jaewoong Lee

⁴

Department of Biomedical Engineering

⁵

Ulsan National Institute of Science and Technology

⁶

2025

⁷

Microbiota in Human Diseases

⁸

Jaewoong Lee

⁹

Department of Biomedical Engineering

¹⁰

Ulsan National Institute of Science and Technology



CHURCH OF THE FLYING SPAGHETTI MONSTER

February 09, 2021

Letter of Good Standing

Dear Sir or Madam:

I am pleased to verify that _____

JAEWOONG LEE

is an ordained minister of the Church of the Flying Spaghetti Monster and recognized
within our organization as a member in good standing.

We hereby consent to this minister performing ceremonies and request that they are
granted all privileges and respect appropriate to a spiritual leader.

Any questions can be directed to the undersigned.

A handwritten signature in black ink that reads "Bobby Henderson".

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson



CHURCH OF THE FLYING SPAGHETTI MONSTER

February 09, 2021

Letter of Good Standing

Dear Sir or Madam:

I am pleased to verify that _____

JAEWOONG LEE

is an ordained minister of the Church of the Flying Spaghetti Monster and recognized
within our organization as a member in good standing.

We hereby consent to this minister performing ceremonies and request that they are
granted all privileges and respect appropriate to a spiritual leader.

Any questions can be directed to the undersigned.

A handwritten signature in black ink that reads "Bobby Henderson".

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson

13

Abstract

14 (Microbiome)

15 (PTB) Section 2 introduces...

16 (Periodontitis) Section 3 describes...

17 (Colon) Setion 4...

18 (Conclusion)

19

20 **This doctoral dissertation is an addition based on the following papers that the author has already
21 published:**

- 22 • Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).
23 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*,
24 13(1), 21105.

Contents

26	1	Introduction	2
27	2	Predicting preterm birth using random forest classifier in salivary microbiome	8
28	2.1	Introduction	8
29	2.2	Materials and methods	10
30	2.2.1	Study design and study participants	10
31	2.2.2	Clinical data collection and grouping	10
32	2.2.3	Salivary microbiome sample collection	10
33	2.2.4	16s rRNA gene sequencing	10
34	2.2.5	Bioinformatics analysis	11
35	2.2.6	Data and code availability	11
36	2.3	Results	12
37	2.3.1	Overview of clinical information	12
38	2.3.2	Comparison of salivary microbiomes composition	12
39	2.3.3	Random forest classification to predict PTB risk	12
40	2.4	Discussion	20
41	3	Random forest prediction model for periodontitis statuses based on the salivary microbiomes	22
42	3.1	Introduction	22
43	3.2	Materials and methods	24
44	3.2.1	Study participants enrollment	24
45	3.2.2	Periodontal clinical parameter diagnosis	24
46	3.2.3	Saliva sampling and DNA extraction procedure	26
47	3.2.4	Bioinformatics analysis	26
48	3.2.5	Data and code availability	27
49	3.3	Results	28

50	3.3.1	Summary of clinical information and sequencing data	28
51	3.3.2	Diversity indices reveal differences among the periodontitis severities .	28
52	3.3.3	DAT among multiple periodontitis severities and their correlation . .	28
53	3.3.4	Classification of periodontitis severities by random forest models . .	29
54	3.4	Discussion	49
55	4	Colon microbiome	53
56	4.1	Introduction	53
57	4.2	Materials and methods	54
58	4.2.1	Study participants enrollment	54
59	4.2.2	DNA extraction procedure	54
60	4.2.3	Bioinformatics analysis	54
61	4.2.4	Data and code availability	54
62	4.3	Results	55
63	4.3.1	Summary of clinical characteristics	55
64	4.3.2	Gut microbiome compositions	55
65	4.3.3	Diversity indices	55
66	4.3.4	DAT selection	55
67	4.4	Discussion	57
68	5	Conclusion	58
69	References		59
70	Acknowledgments		70

71

List of Figures

72	1	DAT volcano plot	14
73	2	Salivary microbiome compositions over DAT	15
74	3	Random forest-based PTB prediction model	16
75	4	Diversity indices	17
76	5	PROM-related DAT	18
77	6	Validation of random forest-based PTB prediction model	19
78	7	Diversity indices	35
79	8	Differentially abundant taxa (DAT)	36
80	9	Correlation heatmap	37
81	10	Random forest classification metrics	38
82	11	Random forest classification metrics from external datasets	39
83	12	Rarefaction curves for alpha-diversity indices	40
84	13	Salivary microbiome compositions in the different periodontal statuses	41
85	14	Correlation plots for differentially abundant taxa	42
86	15	Clinical measurements by the periodontitis statuses	43
87	16	Number of read counts by the periodontitis statuses	44
88	17	Proportion of DAT	45

89	18	Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions	46
90			
91	19	Alpha-diversity indices account for evenness	47
92	20	Gradient Boosting classification metrics	48

List of Tables

94	1	Confusion matrix	6
95	2	Standard clinical information of study participants	13
96	3	Clinical characteristics of the study participants	30
97	4	Feature combinations and their evaluations	31
98	5	List of DAT among the periodontally healthy and periodontitis stages	32
99	6	Feature the importance of taxa in the classification of different periodontal statuses.	33
100	7	Beta-diversity pairwise comparisons on the periodontitis statuses	34
101	8	Clinical characteristics of the study participants	56

102

List of Abbreviations

103 **ACC** Accuracy

104 **ASV** Amplicon sequence variant

105 **AUC** Area-under-curve

106 **BA** Balanced accuracy

107 **C-section** Cesarean section

108 **DAT** Differentially abundant taxa

109 **F1** F1 score

110 **Faith PD** Faith's phylogenetic diversity

111 **FTB** Full-term birth

112 **GA** Gestational age

113 **MWU test** Mann-Whitney U-test

114 **PRE** Precision

115 **PROM** Prelabor rupture of membrane

116 **PTB** Preterm birth

117 **ROC curve** Receiver-operating characteristics curve

118 **rRNA** Ribosomal RNA

119 **SD** Standard deviation

120 **SEN** Sensitivity

121 **SPE** Specificity

122 **t-SNE** t-distributed stochastic neighbor embedding

123 **1 Introduction**

124 The microbiome refers to the complex community of microorganisms, including bacteria, viruses, fungi,
125 and other microbes, that inhabit various environments within living organisms (Ursell, Metcalf, Parfrey,
126 & Knight, 2012; Gilbert et al., 2018). In humans, the microbiome plays a crucial role in maintaining
127 health (Lloyd-Price, Abu-Ali, & Huttenhower, 2016), influencing processes such as digestion (Lim, Park,
128 Tong, & Yu, 2020), immune response (Thaiss, Zmora, Levy, & Elinav, 2016; Kogut, Lee, & Santin, 2020;
129 C. H. Kim, 2018), and even mental health (Mayer, Tillisch, Gupta, et al., 2015; X. Zhu et al., 2017;
130 X. Chen, D'Souza, & Hong, 2013). These microbial communities are not static nor constant, but rather
131 dynamic ecosystem that interacts with their host and respond to environmental changes. Recent studies
132 have revealed that imbalances in the microbiome, known as dysbiosis, can contribute to a wide range of
133 diseases, including obesity (John & Mullin, 2016; Tilg, Kaser, et al., 2011; Castaner et al., 2018), diabetes
134 (Barlow, Yu, & Mathur, 2015; Hartstra, Bouter, Bäckhed, & Nieuwdorp, 2015; Sharma & Tripathi, 2019),
135 infections (Whiteside, Razvi, Dave, Reid, & Burton, 2015; Alverdy, Hyoju, Weigerinck, & Gilbert, 2017),
136 inflammatory conditions (Francescone, Hou, & Grivennikov, 2014; Peirce & Alviña, 2019; Honda &
137 Littman, 2012), and cancers (Helmink, Khan, Hermann, Gopalakrishnan, & Wargo, 2019; Cullin, Antunes,
138 Straussman, Stein-Thoeringer, & Elinav, 2021; Sepich-Poore et al., 2021; Schwabe & Jobin, 2013). Thus,
139 understanding the composition of the human microbiomes is essential for developing new therapeutic
140 approaches that target these microbial populations to promote health and prevent diseases.

141 The microbiome participates a crucial role in overall health, influencing not only digestion and immune
142 function but also systemic and neurological processes through the brain-gut axis (Martin, Osadchiy,
143 Kalani, & Mayer, 2018; Aziz & Thompson, 1998; R. Li et al., 2024). The gut microbiota interact with
144 the host through metabolic byproducts, immune signaling, and the production of neurotransmitters, *e.g.*
145 serotonin and dopamine, which are essential for brain function and cognition. Disruptions in microbial
146 composition, known as dysbiosis, have been linked to various diseases, including inflammatory bowel
147 disease (Sultan et al., 2021; Baldelli, Scaldaferrri, Putignani, & Del Chierico, 2021), obesity (Kang et al.,
148 2022; Hamjane, Mechita, Nourouti, & Barakat, 2024; Pezzino et al., 2023), diabetes (Cai et al., 2024;
149 X. Li et al., 2021; Y. Li et al., 2023), and cardiovascular diseases (Manolis, Manolis, Melita, & Manolis,
150 2022; Tian et al., 2021). Furthermore, the brain-gut axis, a bidirectional communication system between
151 the gut microbiome composition and the central nervous system, has been implicated in mental disorders,
152 *e.g.* anxiety disorder, depressive disorder, and neurodegenerative diseases. Emerging evidence suggested
153 that alterations in the host microbiome can influence mood, cognitive function, and even behavior through
154 immune modulation, vagus nerve signaling, and microbial metabolites. These findings highlight the
155 microbiome as a critical factor in maintaining host health and suggest that targeted interventions, namely
156 probiotics, antibiotics, dietary modification, and microbiome-based therapies, may hold promise for
157 improving both physical and mental comfort. Hence, understanding the microbial effects could lead to
158 novel therapeutic strategies for a wide range of health conditions.

159 16S ribosomal RNA (rRNA) gene sequencing is one of the most extensively applied methods for
160 characterizing microbial communities by targeting the conserved 16S rRNA gene, which contains both

161 highly conserved and variable regions in bacteria (Tringe & Hugenholtz, 2008; Janda & Abbott, 2007).
162 The conserved regions enable universal primer binding, while the variable regions provide the specificity
163 needed to differentiate microbial taxa. Among these regions, the V3-V4 region is frequently selected for
164 sequencing due to its balance between phylogenetic resolution and sequencing efficiency (Johnson et al.,
165 2019; López-Aladid et al., 2023). Therefore, the V3-V4 region offers sufficient variability to classify a
166 wide range of bacteria taxa while maintaining compatibility with widely used sequencing platforms.

167 On the other hand, PathSeq is a computational pipeline designed for the identification and analysis
168 of microbial sequences within short-read human sequencing data, such as next-generation sequencing
169 (Kostic et al., 2011; Walker et al., 2018). PathSeq's scalable and effective processing of massive amounts
170 of sequencing data allows large-scale microbial profiling possible. PathSeq workflow consists of two
171 main phases: a subtractive phase and an analytic phase. The subtractive phase is removing human-derived
172 reads by aligning them to a human reference genome; and, the analytic phase is mapping remaining reads
173 to microbial reference databases, not only bacterial reference genome, but also archaeal, fungal, and viral
174 reference genomes. This approach allows for the comprehensive detection of microbiome compositions,
175 without a requirement for targeted amplification. PathSeq presents a more comprehensive and objective
176 evaluation of microbiome compositions than conventional microbiome profiling techniques including 16S
177 rRNA gene sequencing, capturing an assortment of microbial species beyond bacteria. Therefore, PathSeq
178 is an effective instrument for metagenomic research, infectious disease study, and microbiome analysis in
179 environmental and clinical contexts because of its capacity to operate with complex sequencing datasets
180 (Ojesina et al., 2013; Park et al., 2024; Tejeda et al., 2021).

181 Diversity indices are essential techniques for evaluating the complexity and variety of microbial
182 communities, in ecological and microbiological research (Tucker et al., 2017; Hill, 1973). Alpha-diversity
183 index attributes to the heterogeneity within a specific community, obtaining the number of different taxa
184 and the distribution of taxa among the individuals, *i.e.*, richness and evenness. On the other hand, beta-
185 diversity index measures the variations in microbiome compositions between the individuals, highlighting
186 differences among the microbiome compositions of the study participants (B.-R. Kim et al., 2017).
187 Altogether, by providing a thorough understanding of microbiome compositions, diversity indices, *e.g.*
188 alpha-diversity and beta-diversity, allow us to investigate factors that affecting community variability and
189 structure.

190 Differentially abundant taxa (DAT) detection is a key analytical approach in microbiome study to
191 identify microbial taxa that significantly differ in abundance between distinct study participant groups.
192 This DAT detection method is particularly valuable for understanding how microbial communities vary
193 across different conditions, such as disease states, environmental factors, and/or experimental treatments.
194 Various statistical and computational techniques, *e.g.* LEfSe (Segata et al., 2011), DESeq2 (Love, Huber,
195 & Anders, 2014), ANCOM (Lin & Peddada, 2020), and ANCOM-BC (Lin, Eggesbø, & Peddada,
196 2022; Lin & Peddada, 2024), are commonly used to assess differential abundance while accounting for
197 compositional and sparsity-related challenges in microbiome composition data (Swift, Cresswell, Johnson,
198 Stilianoudakis, & Wei, 2023; Cappellato, Baruzzo, & Di Camillo, 2022). Thus, identifying DAT can
199 provide insights into microbial biomarkers associated with specific health conditions or disease statuses,

enabling potential applications in diagnostics and therapeutics. However, due to the nature of microbiome composition data and the influence of sequencing depth, appropriate normalization and statistically adjustments are necessary to ensure reliable and stable detection of differentially abundant microbes (Xia, 2023; Pan, 2021). Integrating DAT detection analysis with functional profiling further enhances our understanding of the biological significance of microbial shifts or dysbiosis. As microbiome research advances, improving methodologies for DAT selection remains essential for uncovering meaningful microbial association and their potential roles in human diseases.

Classification is one of the supervised machine learning techniques used to categorized data into predefined classes based on features within the data (Kotsiantis, Zaharakis, & Pintelas, 2006; Sen, Hajra, & Ghosh, 2020). In other words, the method learns the relationship between input features and their corresponding output classes through the process of training a classification model using labeled data. Classification models are essential for advising choices in a wide range of applications, including medical diagnostics (Omondiagbe, Veeramani, & Sidhu, 2019). Thus, researchers could uncover sophisticated connections in input features and corresponding classes and produce reliable prediction by utilizing machine learning classification.

Random forest classification is one of the ensemble machine learning methods that constructs several decision trees during training and aggregates their results to provide classification predictions (Breiman, 2001). A portion of the features and classes—known as bootstrapping (Jiang & Simon, 2007; Champagne, McNairn, Daneshfar, & Shang, 2014; J.-H. Kim, 2009) and feature bagging (Bryll, Gutierrez-Osuna, & Quek, 2003; Alelyani, 2021; Yaman & Subasi, 2019)—are utilized to construct each tree in the forest. The majority vote from each tree determines the final classification, which lowers the possibility of overfitting in comparison to a single decision tree. Furthermore, random forest classifier offers several advantages, including its robustness to outliers and its ability to calculate the feature importance.

Evaluating the performance of a machine learning classification model is essential to ensure its reliability and effectiveness in real-world solutions and applications (Novaković, Veljović, Ilić, Papić, & Tomović, 2017; Hossin & Sulaiman, 2015; Hand, 2012). A confusion matrix is a tabular representation of predictions of classification, showing the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Table 1). From this matrix, evaluations can be derived: accuracy (ACC; Equation 1), balanced accuracy (BA; Equation 2), F1 score (F1; Equation 3), sensitivity (SEN; Equation 4), specificity (SPE; Equation 5), and precision (PRE; Equation 6). These metrics are in [0, 1] range and high metrics are good metrics. The confusion matrix also helps in identifying specific types of errors, such as a tendency to produce false positive or false negatives, offering valuable insights for improving the classification model. By combining the confusion matrix with other evaluation metrics, researchers can comprehensively assess the classification metrics and refine it for real-world solutions and applications.

The receiver-operating characteristics (ROC) curve is a graphical representation used to evaluate the performance of a classification model by plotting the sensitivity against (1-specificity) at multiple threshold setting (Gonçalves, Subtil, Oliveira, & de Zea Bermudez, 2014; Obuchowski & Bullen, 2018; Centor, 1991). The ROC curve illustrates the trade-off between detecting true positives while minimizing false positives, suggesting determining the optimal decision threshold for classification. A key metric

239 derived from the ROC curve is the area-under-curve (AUC), which quantifies overall ability of the
240 classification model to discriminate between positive and negative predictions. An AUC value of 0.5
241 indicates a model performing no better than random chance, while value closer to 1.0 suggests high
242 predictive accuracy. Thus, by analyzing the AUC value of the ROC curve, researchers can compare
243 different models and select the better classification model that offers the best balance between sensitivity
244 and specificity for a given application.

245 (Limitation & Novelty)

Table 1: Confusion matrix

		Predicted	
		Positive	Negative
Actual	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

246

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

247

$$BA = \frac{1}{2} \times \left(\frac{TP}{TP + FP} + \frac{TN}{TN + FN} \right) \quad (2)$$

248

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (3)$$

249

$$SEN = \frac{TP}{TP + FP} \quad (4)$$

250

$$SPE = \frac{TN}{TN + FN} \quad (5)$$

$$PRE = \frac{TP}{TP + FP} \quad (6)$$

251 **2 Predicting preterm birth using random forest classifier in salivary mi-**
252 **crobiome**

253 **This section includes the published contents:**

254 Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).
255 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1),
256 21105.

257 **2.1 Introduction**

258 Preterm birth (PTB), characterized by the delivery of neonates prior to 37 weeks of gestation, is one
259 of the major cause to neonatal mortality and morbidity (Blencowe et al., 2012). Multiple pregnancies
260 including twins, short cervical length, and infection on genitourinary tract are known risk factor for
261 PTB (Goldenberg, Culhane, Iams, & Romero, 2008). Nevertheless, the extent to which these aspects
262 affect birth outcomes is still up for debate. Henceforth, strategies to boost gestation and enhance delivery
263 outcomes can be more conveniently implemented when pregnant women at high risk of PTB are identified
264 early (Iams & Berghella, 2010).

265 Prediction models that can be utilized as a foundation for intervention methods still have an unac-
266 ceptable amount of classification evaluations, including accuracy, sensitivity, and specificity, despite a
267 great awareness of the risk factors that trigger PTB (Sotiriadis, Papatheodorou, Kavvadias, & Makrydi-
268 mas, 2010). Several attempts have been made to predict PTB through integrating data such as human
269 microbiome composition, inflammatory markers, and prior clinical data with predictive machine learn-
270 ing methods (Berghella, 2012). Because it is affordable and straightforward to use, fetal fibronectin is
271 commonly used in medical applications. However, with a sensitivity of only 56% that merely similar to
272 random prediction, it has a low classification evaluation (Honest et al., 2009). Due to the difficulty and
273 imprecision of the method in general, as well as the requirement for a qualified specialist cervical length
274 measuring is also restricted (Leitich & Kaider, 2003).

275 Preterm prelabor rupture of membranes (PROM) brought on by gestational inflammation and infection
276 contribute to about 70% of PTB cases (Romero, Dey, & Fisher, 2014). Nevertheless, as antibiotics and
277 anti-inflammatory therapeutic strategies were ineffective to decrease PTB occurrence rates, the pathology
278 of PTB has not been entirely elucidated by inflammatory and infectious pathways (Romero, Hassan, et al.,
279 2014). Recent researches on maternal microbiomes were beginning to examine unidentified connections
280 of PTB as a consequence of developmental processes in molecular biological technology (Fettweis et al.,
281 2019).

282 However, as anti-inflammatory and antibiotic therapies were insufficient to lower PTB occurrence
283 rates, infectious and inflammatory processes are insufficient to exhaustively clarify the pathogenesis and
284 pathophysiology of PTB. It has been hypothesized that the microbiota linked to PTB originate from either
285 a hematogenous pathway or the female genitourinary tract increasing through the vagina and/or cervix.
286 (Han & Wang, 2013). Vaginal microbiome compositions have been found in women who eventually

287 acquire PTB, and recent studies have tried to predict PTB risk using cervico-vaginal fluid (Kindinger et
288 al., 2017). Even though previous investigation have confirmed the potential relationships between the
289 vaginal microbiome compositions and PTB, these studies are only able to clarify an upward trajectory.

290 Multiple unfavorable birth outcomes, including PROM and PTB, have been linked to periodontitis
291 as an independence risk factor, according to numerous epidemiological researches (Offenbacher et al.,
292 1996). It is expected that the oral microbiome will be able to explain additional hematogenous pathways
293 in light of these precedents; however, the oral microbiome composition of fetuses is limited understood.

294 Hence, in order to identify the salivary microbiome linked to PTB and to establish a machine learning
295 prediction model of PTB determined by oral microbiome compositions, this study examined the salivary
296 microbiome compositions of PTB study participants with a full-term birth (FTB) study participants.

297 **2.2 Materials and methods**

298 **2.2.1 Study design and study participants**

299 Between 2019 and 2021, singleton pregnant women who received treatment to Jeonbuk National University
300 Hospital for childbirth were the participants of this study. This study was conducted according to the
301 Declaration of Helsinki (Goodyear, Krleza-Jeric, & Lemmens, 2007). The Institutional Review Board
302 authorized this study (IRB file No. 2019-01-024). Participants who were admitted for elective cesarean
303 sections (C-sections) or induction births, as well as those who had written informed consent obtained
304 with premature labor or PROM, were eligible.

305 **2.2.2 Clinical data collection and grouping**

306 Questionnaires and electronic medical records were implemented to gather information on both previous
307 and current pregnancy outcomes. The following clinical data were analyzed:

- 308 • maternal age at delivery
- 309 • diabetes mellitus
- 310 • hypertension
- 311 • overweight and obesity
- 312 • C-section
- 313 • history PROM or PTB
- 314 • gestational week on delivery
- 315 • birth weight
- 316 • sex

317 **2.2.3 Salivary microbiome sample collection**

318 Salivary microbiome samples were collected 24 hours before to delivery using mouthwash. The standard
319 methods of sterilizing were performed. Medical experts oversaw each stage of the sample collecting
320 procedure. Participants received instruction not to eat, drink, or brush their teeth for 30 minutes before
321 sampling salivary microbiome. Saliva samples were gathered by washing the mouth for 30 seconds with
322 12 mL of a mouthwash solution (E-zен Gargle, JN Pharm, Pyeongtaek, Gyeonggi, Korea). The samples
323 were tagged with the anonymous ID for each participant and kept at 4 °C until they underwent further
324 processing. Genomic DNA was extracted using an ExgeneTM Clinic SV kit (GeneAll Biotechnology,
325 Seoul, Korea) following with the manufacturer instructions and store at -20 °C.

326 **2.2.4 16s rRNA gene sequencing**

327 Salivary microbiome samples were transported to the Department of Biomedical Engineering of the
328 Ulsan National Institute of Science and Technology . 16S rRNA sequencing was then carried out using a
329 commissioned Illumina MiSeq Reagent Kit v3 (Illumina, San Diego, CA, USA). Library methods were
330 utilized to amplify the V3-V4 areas. 300 base-pair paired-end reads were produced by sequencing the

331 pooled library using a v3 \times 600 cycle chemistry after the samples had been diluted to a final concentration
332 of 6 pM with a 20% PhiX control.

333 **2.2.5 Bioinformatics analysis**

334 The independent *t*-test was utilized to evaluate the differences of continuous values between from the
335 PTB participants than the FTB participants; χ^2 -square test was applied to decide statistical differences of
336 categorical values. Clinical measurement comparisons were conducted using SPSS (version 20.0) (Spss
337 et al., 2011). At $p < 0.05$, statistical significance was taken into consideration.

338 QIIME2 (version 2022.2) was implemented to import 16S rRNA gene sequences from salivary
339 microbiome samples of study participants for additional bioinformatics processing (Bolyen et al., 2019).
340 DADA2 was used to verify the qualities of raw sequences (Callahan et al., 2016). The remain sequences
341 were clustered into amplicon sequence variants (ASVs). Diversity indices, namely Faith PD for alpha
342 diversity index (Faith, 1992) and Hamming distance for beta diversity index (Hamming, 1950), were
343 calculated. MWU test (Mann & Whitney, 1947), and PERMANOVA multivariate test were evaluated for
344 measuring statistical significance (Anderson, 2014; Kelly et al., 2015).

345 Taxonomic assignment were implemented with HOMD (version 15.22) (T. Chen et al., 2010).
346 Afterward, DESeq2 was implemented to identify differentially abundant taxa (DAT) that could dis-
347 tinguish between salivary microbiome from PTB and FTB participants (Love et al., 2014). Taxa with
348 $|\log_2 \text{FoldChange}| > 1$ and $p < 0.05$ were considered as statistically significant.

349 The taxa for predicting PTB using salivary microbiome data were determined using a random forest
350 classifier (Breiman, 2001). Through stratified *k*-fold cross-validation (*k* = 5) that preserves the existence
351 rate of PTB and FTB participants, consistency and trustworthy classification were ensured (Wong & Yeh,
352 2019).

353 **2.2.6 Data and code availability**

354 All sequences from the 59 study participants have been added to the Sequence Read Archives (project ID
355 PRJNA985119): <https://dataview.ncbi.nlm.nih.gov/object/PRJNA985119>. Docker image that
356 employed throughout this study is available in the DockerHub: https://hub.docker.com/r/fumire/helixco_premature. Every code used in this study can be found on GitHub: https://github.com/CompbioLabUnist/Helixco_Premature.

359 **2.3 Results**

360 **2.3.1 Overview of clinical information**

361 In the beginning, 69 volunteer mothers were recruited for this study. However, due to insufficient clinical
362 information or twin pregnancies, 10 participants were excluded from the study participants. Demographic
363 and clinical information of the study participants are displayed in Table 2. Because PROM is one of the
364 leading factors of PTB, it was prevalent in the PTB group than the FTB group. Other maternal clinical
365 factors did not significantly differ between the FTB and PTB groups. There were no cases in both groups
366 that had a history of simultaneous periodontal disease or cigarette smoking.

367 **2.3.2 Comparison of salivary microbiomes composition**

368 The salivary microbiome composition was composed of 13953804 sequences from 59 study participants,
369 with 102305.95 ± 19095.60 and 64823.41 ± 15841.65 (mean \pm SD) reads/sample before and following
370 the quality-check stage, accordingly. There was not a significant distinction between the PTB and FTB
371 groups with regard to on alpha diversity nor beta diversity metrics (Figure 4).

372 DESeq2 was used to select 32 DAT that distinguish between the PTB and FTB groups out of the 465
373 species that were examined (Love et al., 2014): 26 FTB-enriched DAT and six PTB-enriched DAT. Seven
374 PROM-related DAT were removed from these 32 PTB-related DAT to lessen the confounding effect of
375 PROM (Figure 5). Therefore, there were a total of 25 PTB-related DAT: 22 FTB-enriched DAT and three
376 PTB-enriched DAT (Figure 1).

377 A significant negative correlation was found using Pearson correlation analysis between GW and
378 differences between PTB-enriched DAT and FTB-enriched DAT ($r = -0.542$ and $p = 7.8e-6$; Figure 5).

379 **2.3.3 Random forest classification to predict PTB risk**

380 To classify PTB according to DAT, random forest classifiers were constructed. The nine most significant
381 DAT were used to obtain the best BA (0.765 ± 0.071 ; Figure 3a). Moreover, random forest classification
382 model determined each DAT's importance (Figure 3b). We conducted a validation procedure on nine
383 twin pregnancies that were excluded in the initial study design in order to confirm the reliability and
384 dependability of our random forest-based PTB prediction model (Figure 6). Comparable to the PTB
385 prediction model on the 59 initial singleton study participants, the validation classification on PTB risk of
386 these twin participants have an accuracy of 87.5%.

Table 2: Standard clinical information of study participants.

Continuous variable for independent *t*-test. Categorical variable for Pearson's χ^2 -square test. Continuous variable: mean \pm SD. Categorical variable: count (proportion)

	PTB (n=30)	FTB (n=29)	p-value
Maternal age (years)	31.8 \pm 5.2	33.7 \pm 4.5	0.687
C-section	20 (66.7%)	24 (82.7%)	0.233
Previous PTB history	4 (13.3%)	1 (3.4%)	0.353
PROM	12 (40.0%)	1 (3.4%)	0.001
Pre-pregnant overweight	8 (26.7%)	7 (24.1%)	1.000
Gestational weight gain (kg)	9.0 \pm 5.9	11.5 \pm 4.6	0.262
Diabetes	2 (6.7%)	2 (6.9%)	1.000
Hypertension	11 (36.7%)	4 (13.8%)	0.072
Gestational age (weeks)	32.5 \pm 3.4	38.3 \pm 1.1	\leq 0.001
Birth weight (g)	1973.4 \pm 686.6	3283.4 \pm 402.7	\leq 0.001
Male	14 (46.7%)	13 (44.8%)	1.000

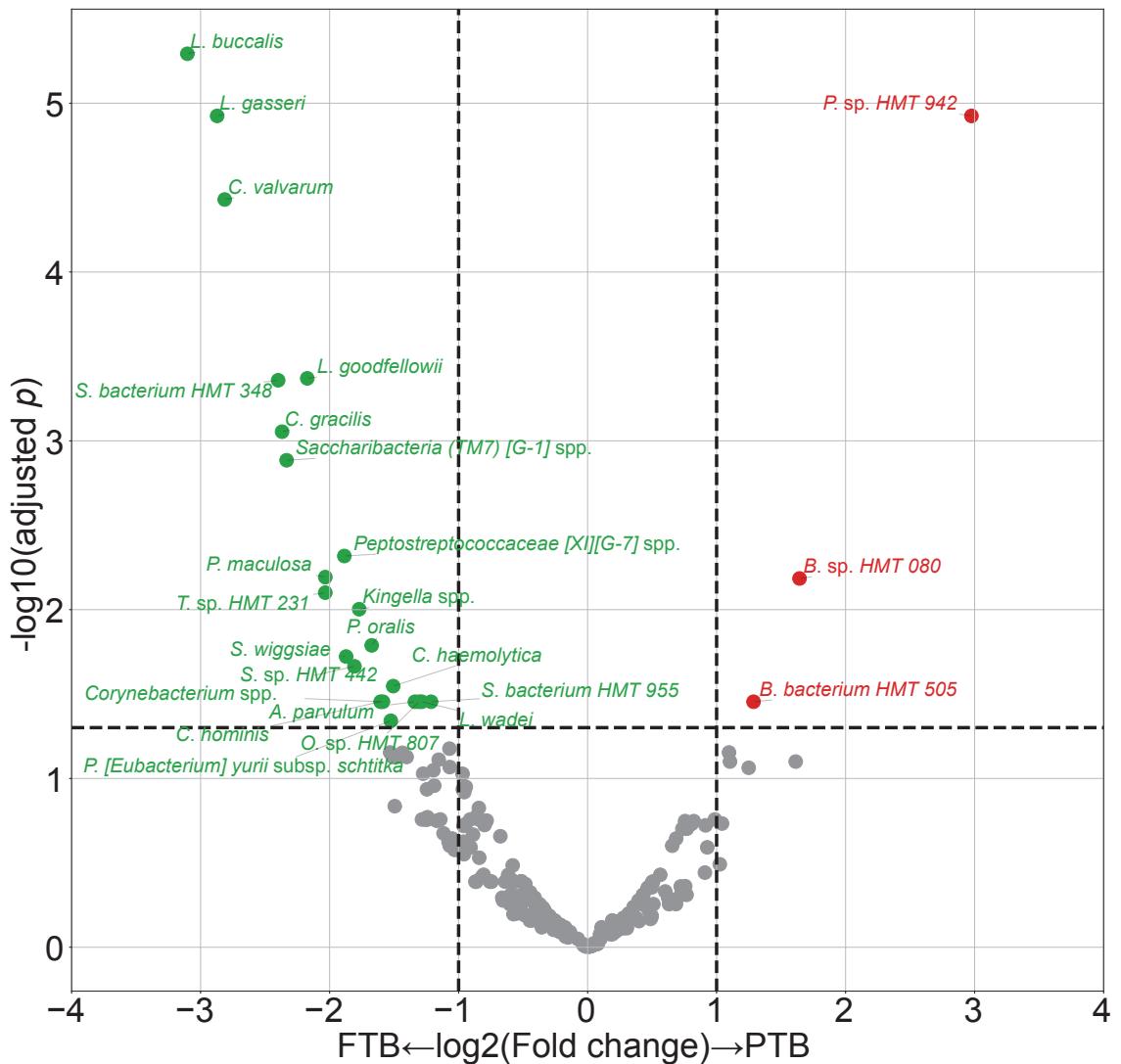


Figure 1: DAT volcano plot.

Red dots represent PTB-enriched DAT, while green dots represent FTB-enriched DAT.

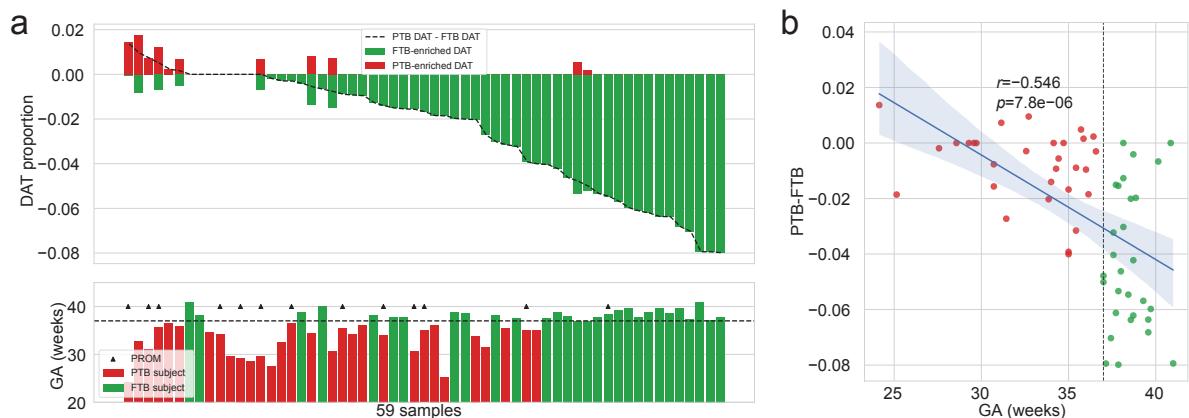


Figure 2: **Salivary microbiome compositions over DAT.**

(a) Frequencies of DAT of study subjects. The study participants are arranged in respect of (PTB-enriched DAT – FTB-enriched DAT). The study participants' GA is displayed in accordance with the upper panel's order (PTB: red bar, FTB: green bar. PROM: arrow head.) **(b)** Correlation plot with GA and (PTB-enriched DAT – FTB-enriched DAT). Strong negative correlation is found with Pearson correlation.

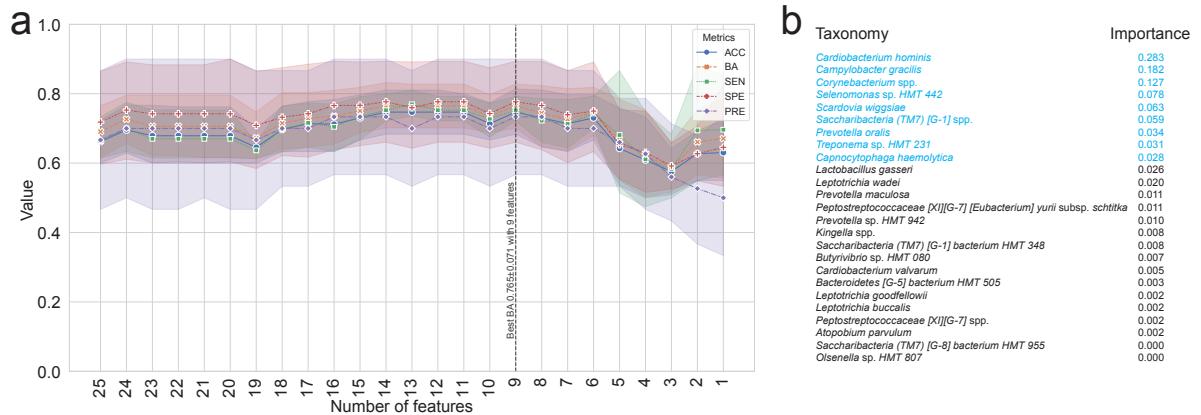


Figure 3: **Random forest-based PTB prediction model.**

(a) Machine learning evaluations upon number of features (DAT). Random Forest classifier has the best BA (0.765 ± 0.071 ; Mean \pm SD) with the nine most important DAT. **(b)** Importance of DAT.

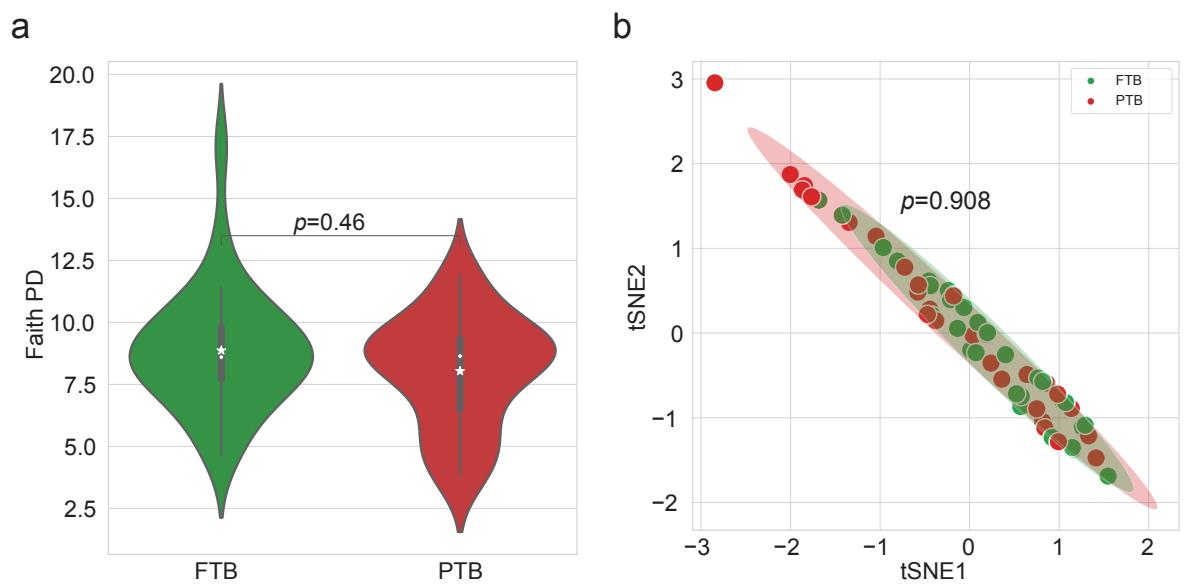


Figure 4: **Diversity indices.**

(a) Alpha diversity index (Faith PD). There is no statistically significant difference between the PTB and FTB group (MWU test $p = 0.46$). **(b)** t-SNE plot with beta diversity index (Hamming distance). There is no statistically significant difference between the PTB and FTB group (PERMANOVA test $p = 0.908$)

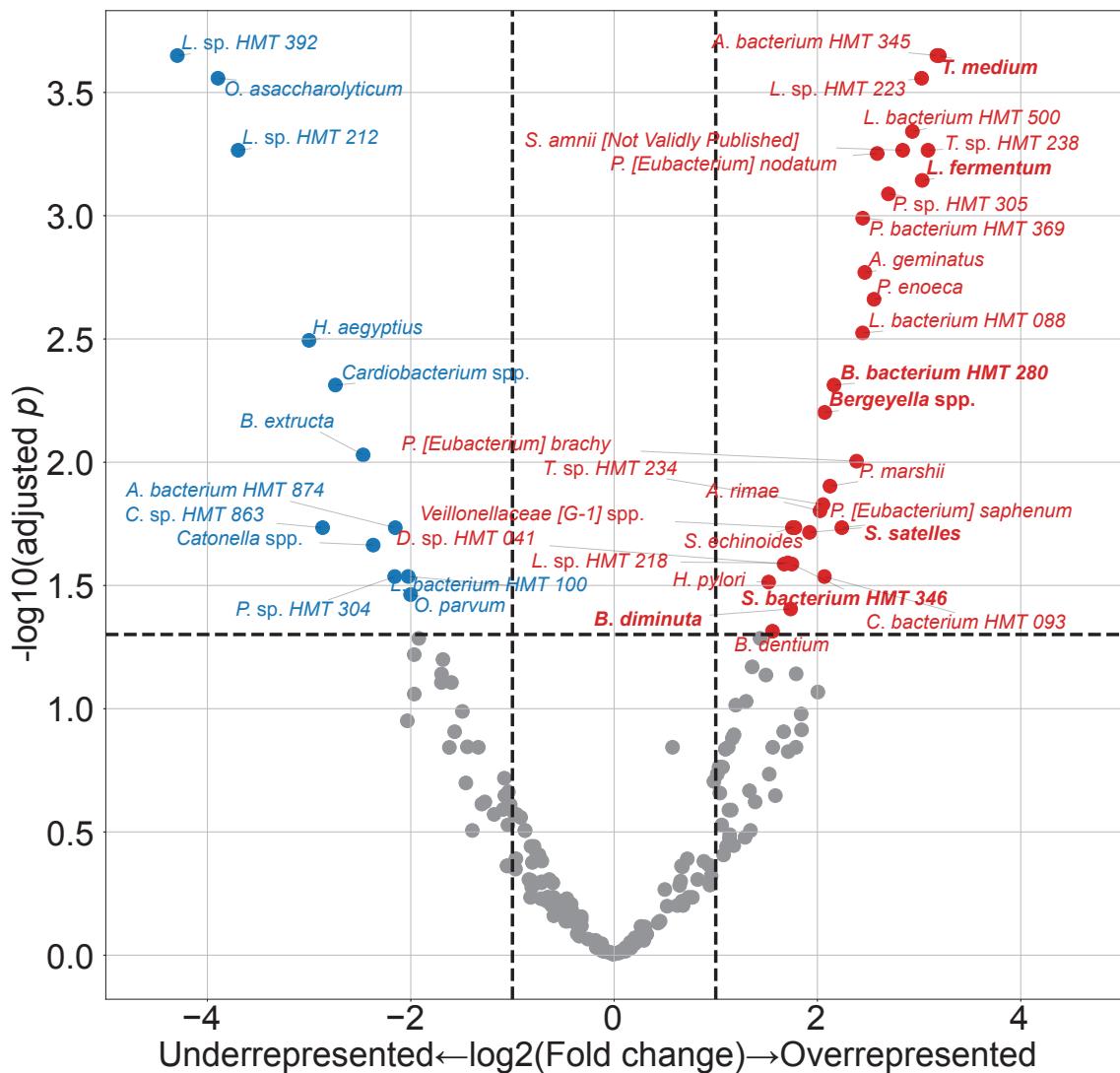


Figure 5: PROM-related DAT.

Only seven of these 42 PROM-related DAT overlapped with PTB-related DAT (bold text). Blue dots represented PROM-underrepresented DAT, while red dots represented PROM-overrepresented DAT.

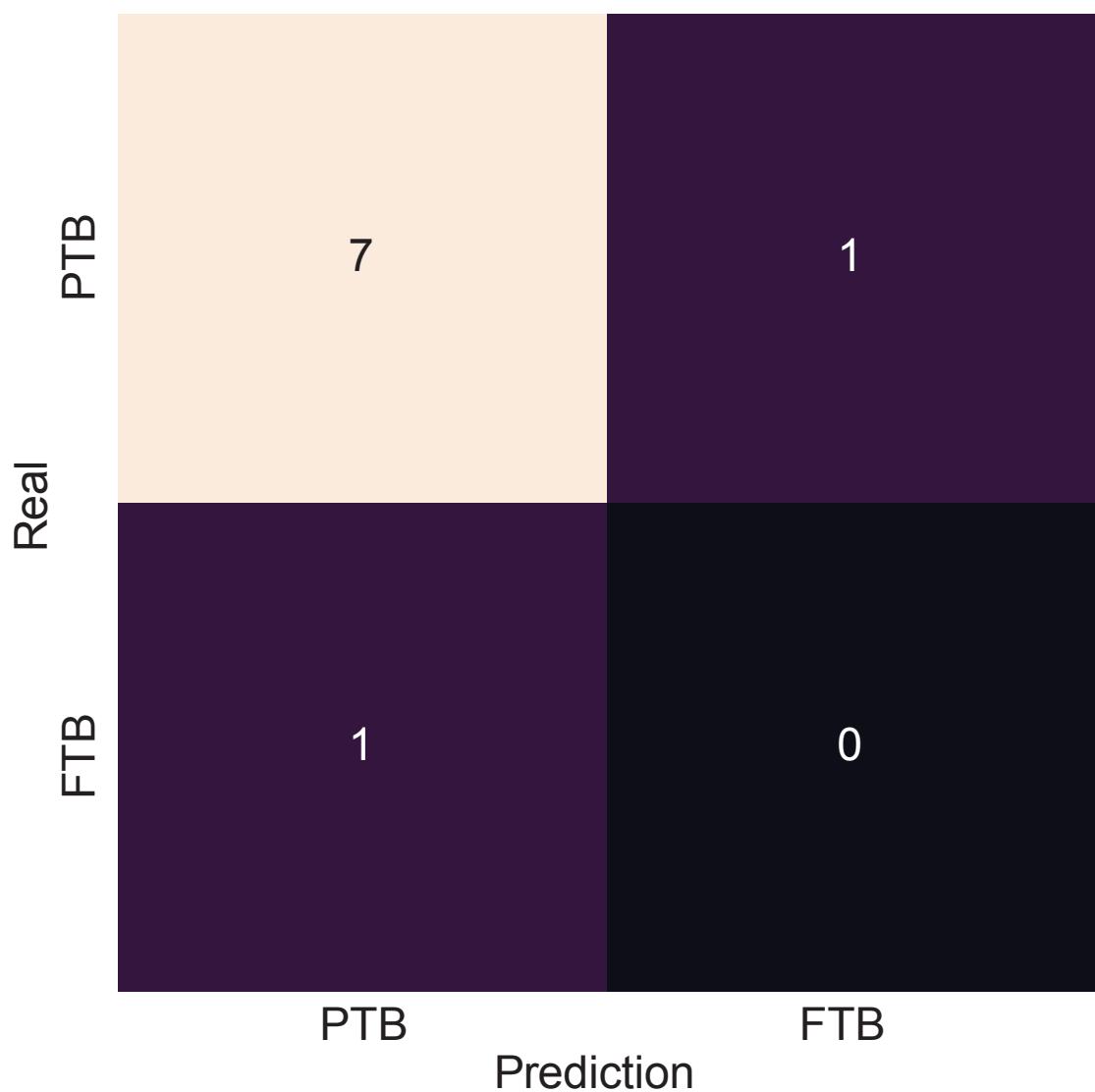


Figure 6: Validation of random forest-based PTB prediction model.

Nine twin pregnancies (eight PTB subjects and a FTB subject) that were excluded in the initial study subjects were subjected to a validation procedure. The random forest-based PTB prediction model shows 87.5% accuracy, comparable to the PTB classification evaluations on the singleton study subjects (0.714 ± 0.061 . Mean \pm SD)

387 **2.4 Discussion**

388 In this study, we employed salivary microbiome compositions to develop the random forest-based PTB
389 prediction models to estimate PTB risks. Previous reports have indicated bidirectional associations
390 between pregnancy outcomes and salivary microbiome compositions (Han & Wang, 2013). Nevertheless,
391 the salivary microbiome composition is not yet elucidated. Salivary microbial dysbiosis, including gingival
392 inflammation and periodontitis, have been connected to unfavorable pregnancy outcomes, such as PTB
393 (Ide & Papapanou, 2013). However, the techniques utilized in recent research that primarily focus on
394 recognized infections have led to inconsistent outcomes.

395 One of the most common salivary taxa that has been examined is *Fusobacterium nucleatum* (Han,
396 2015; Brennan & Garrett, 2019; Bolstad, Jensen, & Bakken, 1996), that is a Gram-negative, anaerobic, and
397 filamentous bacteria. *Fusobacterium nucleatum* can be separated from not only the salivary microbiome
398 but also the vaginal microbiome (Vander Haar, So, Gyamfi-Bannerman, & Han, 2018; Witkin, 2019). In
399 both animal and human investigation, *Fusobacterium nucleatum* infection has been linked to risk of PTB
400 (Doyle et al., 2014). According to recent researches, the placenta women who give birth prematurely may
401 include additional salivary microbiome dysbiosis, such as *Bergeyella* spp. and *Porphyromonas gingivalis*
402 (León et al., 2007; Katz, Chegini, Shiverick, & Lamont, 2009). Although *Bergeyella* spp. were one of the
403 PROM-overrepresented DAT (Figure 5), it was excluded in the final 25 PTB-related DAT. Furthermore,
404 *Porphyromonas gingivalis* and *Campylobacter gracilis* were pathogens of periodontitis in sub-gingival
405 microbiome (Yang et al., 2022). *Lactobacillus gasseri* was also one of the FTB-enriched DAT (Figure
406 1), and it is well established that early PTB risk can be reduced by *Lactobacillus gasseri* in the vaginal
407 microbiome (Basavaprabhu, Sonu, & Prabha, 2020; Payne et al., 2021).

408 With DAT comprising 22 FTB-enriched DAT and three PTB-enriched DAT (Figure 1), we discovered
409 that the FTB study participants had the majority of the essential DAT that distinguished between the PTB
410 and FTB groups. Thus, we hypothesize that the pathogenesis and pathophysiology of PTB may have been
411 triggered by an absence of species with protective characteristics. The association between unfavorable
412 pregnancy outcomes and a dysfunctional microbiome has been explained through two distinct processes.
413 According to the first hypothesis, periodontal pathogens originating in the gingival biofilm might spread
414 from the infected salivary microbiome over the placenta microbiome, invade the intra-amniotic fluid
415 and fetal circulation, and then have a direct impact on the fetoplacental unit, leading to bacteremia
416 (Hajishengallis, 2015). Based on the second hypothesis, inflammatory mediators and endotoxins that
417 generated by the sub-gingival inflammation and derived from dental plaque of periodontitis may spread
418 throughout the body and reach the fetoplacental unit (Stout et al., 2013; Aagaard et al., 2014). Despite
419 belonging to the same species, some subgroups of the salivary microbiome may influence pregnancy
420 outcomes in both favorable and adverse manners. Following this line of argumentation, the salivary
421 microbiome composition or their dysbiosis are more significant than the existence of particular bacteria.

422 Notably, microbial alteration that take place throughout pregnancy may be expected results of a healthy
423 pregnancy. Those pregnancy-related vulnerabilities to dental problem like periodontitis can be explained
424 by three factors. Because of hormone-driven gingival hyper-reactivity to the salivary microbiome in the

425 oral biofilm including sub-gingival biofilm, these conditions are prevalent in pregnant women. For insight
426 at the relationship between the salivary microbiome compositions and PTB, further studies with pathway
427 analysis are warranted.

428 Our study confirmed that salivary microbiome composition could provide potential biomarkers for
429 predicting pregnancy complications including PTB risks using random forest-based classification models,
430 despite a limited number of study participants and a tiny validation sample size. Another limitation of
431 our study was 16S rRNA sequencing. In other words, unlike the shotgun sequencing, 16S rRNA gene
432 sequencing only focused on bacteria, not viruses nor fungi. We did not delve into other variables like
433 nutrition status and socioeconomic statuses of study participants that might affect the salivary microbiome
434 composition.

435 Notwithstanding these limitations, this prospective examination showed the promise of the random
436 forest-based PTB prediction models based on mouthwash-derived salivary microbiome composition.
437 Before applying the methods developed in this study in a clinical context, more multi-center and extensive
438 research is warranted to validate our findings.

439 **3 Random forest prediction model for periodontitis statuses based on the**
440 **salivary microbiomes**

441 **This section includes the published contents:**

442

443 **3.1 Introduction**

444 Saliva microbial dysbiosis brought on by the accumulation of plaque results in periodontitis, a chronic
445 inflammatory disease of the tissue that surrounds the tooth (Kinane, Stathopoulou, & Papapanou, 2017).
446 Loss of periodontal attachment is a consequence of periodontitis, which may lead to irreversible bone loss
447 and, eventually, permanent tooth loss if left untreated. A new classification criterion of periodontal diseases
448 was created in 2018, about 20 years after the 1999 statements of the previous one (Papapanou et al.,
449 2018). Even with this evolution, radiographic and clinical markers of periodontitis progression remain the
450 primary methods for diagnosing periodontitis (Papapanou et al., 2018). Such tools, nevertheless, frequently
451 demonstrate the prior damage from periodontitis rather than its present condition. Certain individuals have
452 a higher risk of periodontitis, a higher chance of developing severe generalized periodontitis, and a worse
453 response to common salivary bacteria control techniques utilized to prevent and treat periodontitis. As a
454 result, the 2017 framework for diagnosing periodontitis additionally allows for the potential development
455 of biomarkers to enhance diagnosis and treatment of periodontitis (Tonetti, Greenwell, & Kornman, 2018).
456 Instead of only depending on the progression of periodontitis, a new etiological indication based on the
457 current state must be introduced in order to enable appropriate intervention through early detection of
458 periodontitis. Thus, the current clinical diagnostic techniques that rely on periodontal probing can be
459 uncomfortable for patients with periodontitis (Canakci & Canakci, 2007).

460 Due to the development of salivaomics, in this manner, the examination of saliva has emerged as
461 a significant alternative to the conventional ways of identifying periodontitis (Altingöz et al., 2021;
462 Melguizo-Rodríguez, Costela-Ruiz, Manzano-Moreno, Ruiz, & Illescas-Montes, 2020). Given that saliva
463 sampling is non-invasive, painless, and accessible to non-specialists, it may be a valuable instrument for
464 diagnosing periodontitis (Zhang et al., 2016). Furthermore, much research has suggested that periodontitis
465 could be a trigger in the development and exacerbation of metabolic syndrome (Morita et al., 2010; Nesbitt
466 et al., 2010). Consequently, alteration in these levels of salivary microbiome markers may serve as high
467 effective diagnostic, prognostic, and therapeutic indicators for periodontitis and other systemic diseases
468 (Miller, Ding, Dawson III, & Ebersole, 2021; Čižmárová et al., 2022). The pathogenesis of periodontitis
469 typically comprises qualitative as well as quantitative alterations in the salivary microbial community,
470 despite that it is a complex disease impacted by a number of contributing factors including age, smoking
471 status, stress, and nourishment (Abusleme, Hoare, Hong, & Diaz, 2021; Lafaurie et al., 2022). Depending
472 on the severity of periodontitis, the salivary microbial community's diversity and characteristics vary
473 (Abusleme et al., 2021), indicating that a new etiological diagnostic standards might be microbial
474 community profiling based on clinical diagnostic criteria. As a consequence, salivary microbiome

475 compositions have been characterized in numerous research in connection with periodontitis. High-
476 throughput sequencing, including 16S rRNA gene sequencing, has recently used in multiple studies to
477 identify variations in the bacterial composition of sub-gingival plaque collections from periodontal healthy
478 individuals and patients with periodontitis (Altabtbaei et al., 2021; Iniesta et al., 2023; Nemoto et al., 2021).
479 This realization has rendered clear that alterations in the salivary microbial community—especially, shifts to
480 dysbiosis—are significant contributors to the pathogenesis and development of periodontitis (Lamont, Koo,
481 & Hajishengallis, 2018). Yet most of these research either focused only on the microbiome alterations in
482 sub-gingival plaque collection, comprised a limited number of periodontitis study participants, or did not
483 account for the impact of multiple severities of periodontitis.

484 For the objective of diagnosing periodontitis, previous research has developed machine learning-based
485 prediction models based on oral microbiome compositions, such as the sub-gingival microbial dysbiosis
486 index (T. Chen, Marsh, & Al-Hebshi, 2022; Chew, Tan, Chen, Al-Hebshi, & Goh, 2024), which have
487 demonstrated good diagnostic evaluation and could be applied to individual saliva collection. Despite
488 offering valuable details, these indicators are frequently restricted by their limited emphasis on classifying
489 the multiple severities of periodontitis. Furthermore, many of these machine learning models currently in
490 practice are trained solely upon the existence of periodontitis rather than on the multiple severities of
491 periodontitis.

492 Recently, we employed multiplex quantitative-PCR and machine learning-based classification model
493 to predict the severity of periodontitis based on the amount of nine pathogens of periodontitis from
494 saliva collections (E.-H. Kim et al., 2020). On the other hand, the fact that we focused merely at nine
495 pathogens for periodontitis and neglected the variety bacterial species associated to the various severities
496 of periodontitis constrained the breadth of our investigation. By developing a machine learning model
497 that could classify multiple severities of periodontitis based on the salivary microbiome composition,
498 this study aims to fill these knowledge gaps and produce more accurate and therapeutically useful
499 guidance to evaluate progression of periodontitis. Hence, in order to examine the salivary microbiome
500 composition of both healthy controls and patients with periodontitis in multiple stages, we applied
501 16S rRNA gene sequencing. Furthermore, employing the 2018 classification criteria, we sought to find
502 biomarkers (species) for the precise prediction of periodontitis severities (Papapanou et al., 2018; Chapple
503 et al., 2018).

504 **3.2 Materials and methods**

505 **3.2.1 Study participants enrollment**

506 Between 2018-08 and 2019-03, 250 study participants—100 healthy controls, 50 patients with stage I
507 periodontitis, 50 patients with stage II periodontitis, and 50 patients with stage III periodontitis—visited
508 visited the Department of Periodontics at Pusan National University Dental Hospital. The Institutional
509 Review Board of the Pusan National University Dental Hospital accepted this study protocol and design
510 (IRB No. PNUDH-2016-019). Every study participants provided their written informed authorization
511 after being fully informed about this study's objectives and methodologies. Exclusion criteria for the
512 study participants are followings:

- 513 1. People who, throughout the previous six months, underwent periodontal therapy, including root
514 planing and scaling.
- 515 2. People who struggle with systemic conditions that may affect periodontitis developments, such as
516 diabetes.
- 517 3. People who, throughout the previous three months, were prescribed anti-inflammatory medications
518 or antibiotics.
- 519 4. Women who were pregnant or breastfeeding.
- 520 5. People who have persistent mucosal lesions, e.g. pemphigus or pemphigoid, or acute infection, e.g.
521 herpetic gingivostomatitis.
- 522 6. Patient with grade C periodontitis or localized periodontitis (< 30% of teeth involved).

523 **3.2.2 Periodontal clinical parameter diagnosis**

524 A skilled periodontist conducted each clinical procedure. Six sites per tooth were used to quantify
525 gingival recession and probing depth: mesiobuccal, midbuccal, distobuccal, mesiolingual, midlingual,
526 and distolingual (Huang et al., 2007). A periodontal probe (Hu-Friedy, IL, USA) was placed parallel to
527 the major axis of the tooth at each tooth location in order to gather measurements. The cementoenamel
528 junction of the tooth was analyzed to determine the clinical attachment level, and the deepest point of
529 probing was taken to determine the periodontal pocket depth from the marginal gingival level of the
530 tooth. Plaque index was measured by probing four surfaces per tooth: mesial, distal, buccal, and palatal
531 or lingual. Plaque index was scored by the following criteria:

- 532 0. No plaque present.
- 533 1. A thin layer of plaque that adheres to the surrounding tissue of the tooth and free gingival margin.
534 Only through the use of a periodontal probe on the tooth surface can the plaque be existed.
- 535 2. Significant development of soft deposits that are visible within the gingival pocket, which is a
536 region between the tooth and gingival margin.

537 3. Considerable amount of soft matter on the tooth, the gingival margin, and the gingival pocket.

538 The arithmetic average of the plaque indices collected from every tooth was determined to calculate
539 plaque index of each study participant. By probing four surfaces per tooth, mesial, distal, buccal, and
540 palatal or lingual, to assess gingival bleeding, the gingival index was scored by the following criteria:

541 0. Normal gingiva: without inflammation nor discoloration.

542 1. Mild inflammation: minimal edema and slight color changes, but no bleeding on probing.

543 2. Moderate inflammation: edema, glazing, redness, and bleeding on probing.

544 3. Severe inflammation: significant edema, ulceration, redness, and spontaneous bleeding.

545 The arithmetic average of the gingival indices collected from every tooth was determined to calculate
546 gingival index of each study participant. The relevant data was not displayed, despite that furcation
547 involvement and bleeding on probing were thoroughly utilized into account during the diagnosis process.

548 Periodontitis was diagnosed in respect to the 2018 classification criteria (Papapanou et al., 2018;
549 Chapple et al., 2018). An experienced periodontist diagnosed the periodontitis severity by considering
550 complexity, depending on clinical examinations including radiographic images and periodontal probing.

551 Periodontitis is categorized into healthy, stage I, stage II, and stage III with the following criteria:

552 • Healthy:

553 1. Bleeding sites < 10%

554 2. Probing depth: \leq 3 mm

555 • Stage I:

556 1. No tooth loss because of periodontitis.

557 2. Inter-dental clinical attachment level at the site of the greatest loss: 1-2 mm

558 3. Radiographic bone loss: < 15%

559 • Stage II:

560 1. No tooth loss because of periodontitis.

561 2. Inter-dental clinical attachment level at the site of the greatest loss: 3-4 mm

562 3. Radiographic bone loss: 15-33%

563 • Stage III:

564 1. Teeth loss because of periodontitis: \leq teeth

565 2. Inter-dental clinical attachment level at the site of the greatest loss: \geq 5 mm

566 3. Radiographic bone loss: > 33%

567 **3.2.3 Saliva sampling and DNA extraction procedure**

568 All study participants received instructions to avoid eating, drinking, brushing, and using mouthwash for
569 at least an hour prior to the saliva sample collection process. These collections were conducted between
570 09:00 and 11:00. Mouth rinse was collected by rinsing the mouth for 30 seconds with 12 mL of a solution
571 (E-zen Gargle, JN Pharm, Korea). All saliva samples were tagged with anonymous ID and stored at -4 °C.

572 Bacteria DNA was extracted from saliva samples using an Exgene™Clinic SV DNA extraction kit
573 (GeneAll, Seoul, Korea), and quality and quantity of bacterial DNA was measured using a NanoDrop
574 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). Hyper-variable regions (V3-V4)
575 of the 16S rRNA gene were amplified using the following primer:

- 576 • Forward: 5' -TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNNGCWGCAG-3'
577 • Reverse: 5' -GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'

578 The standard protocols of the Illumina 16S Metagenomic Sequencing Library Preparation were
579 followed in the preparation of the libraries. The PCR conditions were as follows:

- 580 1. Heat activation for 30 seconds at 95 °C.
581 2. 25 cycles for 30 seconds at 95 °C.
582 3. 30 seconds at 55 °C.
583 4. 30 seconds at 72 °C.

584 NexteraXT Indexed Primer was applied to amplification 10 µL of the purified initial PCR products for
585 the final library creation. The second PCR used the same conditions as the first PCR conditions but with
586 10 cycles. 16S rRNA gene sequencing was performed via 2×300 bp paired-end sequencing at Macrogen
587 Inc. (Macrogen, Seoul, Korea) using Illumina MiSeq platform (Illumina, San Diego, CA, USA).

588 **3.2.4 Bioinformatics analysis**

589 We computed alpha-diversity and beta-diversity indices to quantify the divergence of phylogenetic
590 information. Following alpha-diversity indices were calculated using the scikit-bio Python package
591 (version 0.5.5) (Rideout et al., 2018), and these alpha-diversity indices were compared using the MWU
592 test:

- 593 • Abundance-based Coverage Estimator (ACE) (Chao & Lee, 1992)
594 • Chao1 (Chao, 1984)
595 • Fisher (Fisher, Corbet, & Williams, 1943)
596 • Margalef (Magurran, 2021)
597 • Observed ASVs (DeSantis et al., 2006)
598 • Berger-Parker d (Berger & Parker, 1970)
599 • Gini index (Gini, 1912)

600 • Shannon (Weaver, 1963)
601 • Simpson (Simpson, 1949)
602 Aitchison index for a beta-diversity index was calculated using QIIME2 (version 2020.8) (Aitchison,
603 Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000; Bolyen et al., 2019). We employed the
604 t-SNE algorithm to illustrate multi-dimensional data from the beta-diversity index computation (Van der
605 Maaten & Hinton, 2008). The beta-diversity index was compared using the PERMANOVA test (Anderson,
606 2014; Kelly et al., 2015) and MWU test.

607 DAT between multiple periodontitis stages were identified by ANCOM (Lin & Peddada, 2020). The
608 log-transformed absolute abundances of DAT were analyzed by hierarchical clustering in order to identify
609 sub-groups with similar abundance patterns on periodontitis severities. Additionally, we examined the
610 relative proportions among the 20 DAT in order to reduce the effect of salivary bacteria that differ
611 insignificantly across the multiple severities of periodontitis.

612 Differentially abundant taxa (DAT) among multiple periodontitis severities were selected from the
613 salivary microbiome compositions by ANCOM (Lin & Peddada, 2020). In contrast to conventional
614 techniques that examine raw abundance counts, ANCOM applies log-ratio between taxa to account for
615 the salivary microbiome composition data. The log-transformed abundances of DAT were subjected to
616 hierarchical clustering to discover subgroups of DAT with similar patterns on periodontitis severities.
617 Furthermore, we examined the relative proportion among the DAT in order to reduce the effects of other
618 salivary bacteria that differ non-significantly across the multiple periodontitis severities.

619 As previously stated (E.-H. Kim et al., 2020), we used stratified k -fold cross-validation ($k = 10$)
620 by severity of periodontitis to achieve consistent and trustworthy classification results (Wong & Yeh,
621 2019). Additionally, we utilized various features with confusion matrices and their derivations to evaluate
622 the classification outcomes in order to identify which features optimize classification evaluations and
623 decrease sequencing efforts. Using the DAT discovered by ANCOM, we iteratively removed the least
624 significant taxa from the input features (taxa) of the random forest classification models using the
625 backward elimination method.

626 We investigated external datasets from Spanish individuals (Iniesta et al., 2023) and Portuguese
627 individuals (Relvas et al., 2021) to confirm that our random forest classification was consistent. To
628 ascertain repeatability and dependability, the external datasets were processed using the same pipeline
629 and parameters as those used for our study participants.

630 3.2.5 Data and code availability

631 All sequences from the 250 study participants have been added to the Sequence Read Archives (project
632 ID PRJNA976179): <https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA976179>. Docker
633 image that employed throughout this study is available in the DockerHub: https://hub.docker.com/repository/docker/fumire/periodontitis_16s. Every code used in this study can be found on
634 GitHub: https://github.com/CompbioLabUnist/Periodontitis_16S.

636 **3.3 Results**

637 **3.3.1 Summary of clinical information and sequencing data**

638 Among clinical information of the study participants, clinical attachment level, probing depth, plaque
639 index, and gingival index, were significantly increased with periodontitis severity (Kruskal-Wallis test
640 $p < 0.001$), while sex were observed no significant difference (Table 2). Notably, clinical attachment level
641 and probing depth have significant differences among the periodontitis severities (MWU test $p < 0.01$;
642 Figure 15). Additionally, 71461.00 ± 11792.30 and 45909.78 ± 11404.65 reads per sample were obtained
643 before and after filtering low-quality reads and trimming extra-long tails, respectively (Figure 16).

644 **3.3.2 Diversity indices reveal differences among the periodontitis severities**

645 Rarefaction curves showed that the sequencing depth was sufficient (Figure 12). Alpha-diversity in-
646 dices indicated significant differences between the healthy and the periodontitis stages (MWU test
647 $p < 0.01$; Figure 7a-e); however, there were no significant differences between the periodontitis stages.
648 This emphasizes how essential it is to classify the salivary microbiome compositions and distinguish
649 between the stages of periodontitis using machine learning approaches.

650 The confidence ellipses of the tSNE-transformed beta-diversity index (Aitchison index) indicated
651 distinct distributions among the periodontitis severities (PERMANOVA $p \leq 0.001$; Figure 7f). Aitchison
652 index demonstrated significant differences every pairwise of the periodontitis severities (PERMANOVA
653 test $p \leq 0.001$; Table 7). Significant differences in the distances between periodontitis severities further
654 demonstrated the uniqueness of each severity of periodontitis (MWU test $p \leq 0.05$; Figure 7g-j).

655 **3.3.3 DAT among multiple periodontitis severities and their correlation**

656 Of the 425 total taxa that identified in the salivary microbiome composition (Figure 13), 20 DAT were
657 identified (Table 5). Three separate subgroups were formed from the participants-level abundances of the
658 DAT using a hierarchical clustering methodology (Figure 8a):

- 659 • Group 1
- 660 1. *Treponema* spp.
- 661 2. *Prevotella* sp. HMT 304
- 662 3. *Prevotella* sp. HMT 526
- 663 4. *Peptostreptococcaceae [XI][G-5] saphenum*
- 664 5. *Treponema* sp. HMT 260
- 665 6. *Mycoplasma faecium*
- 666 7. *Peptostreptococcaceae [XI][G-9] brachy*
- 667 8. *Lachnospiraceae [G-8] bacterium* HMT 500
- 668 9. *Peptostreptococcaceae [XI][G-6] nodatum*
- 669 10. *Fretibacterium* spp.

- 670 • Group 2
- 671 1. *Porphyromonas gingivalis*
- 672 2. *Campylobacter showae*
- 673 3. *Filifactor alocis*
- 674 4. *Treponema putidum*
- 675 5. *Tannerella forsythia*
- 676 6. *Prevotella intermedia*
- 677 7. *Porphyromonas* sp. HMT 285

- 678 • Group 3
- 679 1. *Actinomyces* spp.
- 680 2. *Corynebacterium durum*
- 681 3. *Actinomyces graevenitzii*

682 Ten DAT that were significant enriched in stage II and stage III, but deficient in healthy formed Group
683 1. Furthermore, in comparison to the healthy, the seven DAT of Group 2 were significantly enriched in
684 each of the stages of periodontitis. On the other hand, three DAT in Group 3 were deficient in stage II
685 and stage III, but significantly enriched in healthy. The relative proportions of the DAT further supported
686 these findings (Figure 8b), suggesting that the DAT is primarily linked to periodontitis rather than other
687 salivary bacteria.

688 Correlation analysis from the DAT showed that DAT from Group 3 was negatively correlated with
689 Group 1 and Group 2 (Figure 9), and strong correlations were observed the nine pairs of DAT (Figure 14).

690 3.3.4 Classification of periodontitis severities by random forest models

691 Based on the proportion of DAT, random forest classifier were trained to classify the periodontitis
692 severities (Table 6). First of all, we conducted multi-label classification for the multiple periodontitis
693 severities, namely healthy, stage I, stage II, and stage III. In this setting, we classified multiple periodontitis
694 severities with the highest BA of 0.779 ± 0.029 (Table 4). AUC ranged between 0.81 and 0.94 (Figure
695 10b).

696 Second, since timely detection in dentistry is demanding (Tonetti et al., 2018), we implemented a
697 random forest classification for both healthy and stage I. Remarkably, the random forest classifier had
698 the highest BA at 0.793 ± 0.123 (Table 4). In this setting, this model showed high AUC value for the
699 classifying of stage I from healthy (AUC=0.85; Figure 10d).

700 Third, based on the findings that the salivary microbiome composition in stage II is more comparable
701 to those in stage III than to other severities (Figure 7f and Figure 7j), we combined stage II and stage III
702 to perform a multi-label classification.

Table 3: Clinical characteristics of the study participants.

Significant differences were assessed using the Kruskal-Wallis test. NA: Not applicable.

Index	Healthy	Stage I	Stage II	Stage III	p-value
Age (year)	33.83±13.04	43.30±14.28	50.26±11.94	51.08±11.13	6.18E-17
Gender (Male)	44 (44.0%)	22 (44.0%)	25 (50.0%)	25 (50.0%)	NA
Smoking (Never)	83 (83.0%)	36 (72.0%)	34 (68.0%)	29 (58.0%)	NA
Smoking (Ex)	12 (12.0%)	7 (14.0%)	9 (18.0%)	10 (20.0%)	NA
Smoking (Current)	2 (2.0%)	7 (14.0%)	7 (14.0%)	10 (20.0%)	NA
Number of teeth	28.03±2.23	27.36±1.80	26.72±2.89	25.74±4.34	8.07E-05
Attachment level (mm)	2.45±0.29	2.75±0.38	3.64±0.83	4.54±1.14	1.82E-35
Probing depth (mm)	2.42±0.29	2.61±0.40	3.27±0.76	3.95±0.88	6.43E-28
Plaque index	17.66±16.21	35.46±23.75	54.40±23.79	58.30±25.25	3.23E-22
Gingival index	0.09±0.16	0.44±0.46	0.85±0.52	1.06±0.52	2.59E-32

Table 4: Feature combinations and their evaluations

Classification performance with the most important taxon, the two most important taxa, and taxa with the best-balanced accuracy. *P.gingivalis* and *Act.* are *Porphyromonas gingivalis* and *Actinomyces* spp., respectively.

Classification	Features	ACC	AUC	BA	F1	PRE	SEN	SPE
Healthy vs. Stage I vs. Stage II vs. Stage III	<i>P.gingivalis</i>	0.758±0.051	0.716±0.177	0.677±0.068	0.839±0.034	0.839±0.034	0.516±0.102	
	<i>P.gingivalis+Act.</i>	0.792±0.043	0.822±0.105	0.723±0.057	0.861±0.029	0.861±0.029	0.584±0.086	
Top 5 taxa		0.834±0.022	0.870±0.079	0.779±0.029	0.889±0.015	0.889±0.015	0.668±0.033	
Healthy vs. Stage I	<i>Act.</i>	0.687±0.116	0.725±0.145	0.647±0.159	0.762±0.092	0.760±0.128	0.781±0.116	0.513±0.224
	<i>Act.+P.gingivalis</i>	0.733±0.119	0.831±0.081	0.713±0.122	0.797±0.097	0.797±0.126	0.798±0.082	0.627±0.191
Top 9 taxa		0.800±0.103	0.852±0.103	0.793±0.123	0.849±0.080	0.850±0.112	0.857±0.090	0.730±0.193
Healthy vs. Stage I vs. Stages II/III	<i>P.gingivalis</i>	0.776±0.042	0.736±0.196	0.748±0.047	0.832±0.031	0.832±0.031	0.664±0.062	
	<i>P.gingivalis+Act.</i>	0.843±0.035	0.876±0.109	0.823±0.039	0.882±0.026	0.882±0.026	0.764±0.052	
Top 6 taxa		0.885±0.036	0.914±0.027	0.871±0.038	0.914±0.027	0.914±0.025	0.828±0.051	
Healthy vs. Stages I/II/III	<i>P.gingivalis</i>	0.792±0.114	0.856±0.105	0.819±0.088	0.776±0.089	0.840±0.092	0.756±0.175	0.883±0.054
	<i>P.gingivalis+Act.</i>	0.828±0.121	0.926±0.074	0.847±0.116	0.797±0.123	0.800±0.126	0.830±0.191	0.864±0.074
Top 4 taxa		0.860±0.078	0.953±0.049	0.885±0.066	0.832±0.079	0.840±0.128	0.864±0.157	0.905±0.070

Table 5: List of DAT among healthy status and periodontitis stages

No.	Taxonomy	ANCOM W score
1	<i>Porphyromonas gingivalis</i>	424
2	<i>Actinomyces</i> spp.	424
3	<i>Filifactor alocis</i>	421
4	<i>Prevotella intermedia</i>	419
5	<i>Treponema putidum</i>	418
6	<i>Tannerella forsythia</i>	415
7	<i>Porphyromonas</i> sp. HMT 285	412
8	<i>Peptostreptococcaceae [XI][G-6] nodatum</i>	412
9	<i>Fretibacterium</i> spp.	411
10	<i>Mycoplasma faecium</i>	411
11	<i>Prevotella</i> sp. HMT 304	411
12	<i>Lachnospiraceae [G-8] bacterium</i> HMT 500	409
13	<i>Treponema</i> spp.	408
14	<i>Prevotella</i> sp. HMT 526	401
15	<i>Peptostreptococcaceae [XI][G-9] brachy</i>	400
16	<i>Peptostreptococcaceae [XI][G-5] saphenum</i>	398
17	<i>Campylobacter showae</i>	395
18	<i>Treponema</i> sp. HMT 260	393
19	<i>Corynebacterium durum</i>	393
20	<i>Actinomyces graevenitzii</i>	387

Table 6: Feature the importance of taxa in the classification of different periodontal statuses
 Taxa are ranked in descending order of importance; from most important to least important.

Condition	Healthy vs. Stage I vs. Stage II vs. Stage III			Healthy vs. Stage I			Healthy vs. Stage I vs. Stage II/III			Healthy vs. Stage VII/III		
	Rank	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance	
1	<i>Porphyromonas gingivalis</i>	0.297	<i>Actinomyces spp.</i>	0.195	<i>Porphyromonas gingivalis</i>	0.360	<i>Porphyromonas gingivalis</i>	0.426	<i>Porphyromonas gingivalis</i>	0.461	<i>Actinomyces spp.</i>	0.461
2	<i>Actinomyces spp.</i>	0.195	<i>Actinomyces graevenitzii</i>	0.054	<i>Actinomyces graevenitzii</i>	0.125	<i>Actinomyces spp.</i>	0.244	<i>Actinomyces spp.</i>	0.257	<i>Actinomyces graevenitzii</i>	0.257
3	<i>Prevotella intermedia</i>	0.054	<i>Porphyromonas sp. HMT 285</i>	0.052	<i>Porphyromonas sp. HMT 285</i>	0.095	<i>Actinomyces graevenitzii</i>	0.049	<i>Actinomyces spp.</i>	0.059	<i>Corynebacterium durum</i>	0.035
4	<i>Actinomyces graevenitzii</i>	0.052	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.050	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.062	<i>Corynebacterium durum</i>	0.046	<i>Filifactor alocis</i>	0.032	<i>Filifactor alocis</i>	0.032
5	<i>Filifactor alocis</i>	0.050	<i>Campylobacter showae</i>	0.042	<i>Campylobacter showae</i>	0.052	<i>Prevotella intermedia</i>	0.036	<i>Campylobacter showae</i>	0.023	<i>Campylobacter showae</i>	0.023
6	<i>Campylobacter showae</i>	0.042	<i>Porphyromonas sp. HMT 285</i>	0.040	<i>Filifactor alocis</i>	0.050	<i>Prevotella intermedia</i>	0.033	<i>Porphyromonas sp. HMT 285</i>	0.025	<i>Porphyromonas sp. HMT 285</i>	0.022
7	<i>Porphyromonas sp. HMT 285</i>	0.040	<i>Corynebacterium durum</i>	0.032	<i>Corynebacterium durum</i>	0.038	<i>Campylobacter showae</i>	0.023	<i>Prevotella intermedia</i>	0.022	<i>Prevotella intermedia</i>	0.022
8	<i>Corynebacterium durum</i>	0.032	<i>Treponema spp.</i>	0.032	<i>Treponema spp.</i>	0.037	<i>Porphyromonas sp. HMT 285</i>	0.021	<i>Treponema spp.</i>	0.022	<i>Treponema spp.</i>	0.022
9	<i>Treponema spp.</i>	0.032	<i>Tannerella forsythia</i>	0.026	<i>Tannerella forsythia</i>	0.029	<i>Treponema spp.</i>	0.018	<i>Peptostreptococcaceae (XII/G-9) brachy</i>	0.015	<i>Peptostreptococcaceae (XII/G-9) brachy</i>	0.015
10	<i>Tannerella forsythia</i>	0.026	<i>Prevotella intermedia</i>	0.025	<i>Prevotella intermedia</i>	0.026	<i>Peptostreptococcaceae (XII/G-9) brachy</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.010	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.010
11	<i>Prevotella intermedia</i>	0.025	<i>Freibacterium spp.</i>	0.023	<i>Freibacterium spp.</i>	0.018	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.011	<i>Tannerella forsythia</i>	0.009	<i>Tannerella forsythia</i>	0.009
12	<i>Freibacterium spp.</i>	0.023	<i>Peptostreptococcaceae (XII/G-9) brachy</i>	0.021	<i>Peptostreptococcaceae (XII/G-9) brachy</i>	0.018	<i>Peptostreptococcaceae (XII/G-6) nodatum</i>	0.010	<i>Freibacterium spp.</i>	0.009	<i>Freibacterium spp.</i>	0.009
13	<i>Peptostreptococcaceae (XII/G-9) brachy</i>	0.021	<i>Treponema putidum</i>	0.019	<i>Treponema putidum</i>	0.014	<i>Treponema putidum</i>	0.009	<i>Prevotella putidum</i>	0.006	<i>Prevotella putidum</i>	0.006
14	<i>Treponema sp. HMT 260</i>	0.019	<i>Prevotella sp. HMT 526</i>	0.018	<i>Prevotella sp. HMT 526</i>	0.011	<i>Prevotella sp. HMT 526</i>	0.008	<i>Peptostreptococcaceae (XII/G-6) nodatum</i>	0.004	<i>Peptostreptococcaceae (XII/G-6) nodatum</i>	0.004
15	<i>Prevotella sp. HMT 526</i>	0.018	<i>Treponema sp. HMT 260</i>	0.018	<i>Treponema sp. HMT 260</i>	0.008	<i>Freibacterium spp.</i>	0.008	<i>Treponema sp. HMT 260</i>	0.004	<i>Treponema sp. HMT 260</i>	0.004
16	<i>Peptostreptococcaceae (XII/G-6) nodatum</i>	0.018	<i>Prevotella sp. HMT 304</i>	0.017	<i>Peptostreptococcaceae (XII/G-6) nodatum</i>	0.008	<i>Freibacterium spp.</i>	0.008	<i>Mycoplasma faecium</i>	0.004	<i>Mycoplasma faecium</i>	0.004
17	<i>Prevotella sp. HMT 304</i>	0.017	<i>Mycoplasma faecium</i>	0.014	<i>Mycoplasma faecium</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.005	<i>Prevotella sp. HMT 304</i>	0.003	<i>Prevotella sp. HMT 304</i>	0.003
18	<i>Mycoplasma faecium</i>	0.014	<i>Prevotella sp. HMT 304</i>	0.014	<i>Prevotella sp. HMT 304</i>	0.003	<i>Mycoplasma faecium</i>	0.005	<i>Peptostreptococcaceae (XII/G-5) saphenum</i>	0.002	<i>Peptostreptococcaceae (XII/G-5) saphenum</i>	0.002
19	<i>Peptostreptococcaceae (XII/G-5) saphenum</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013	<i>Peptostreptococcaceae (XII/G-5) saphenum</i>	0.003	<i>Peptostreptococcaceae (XII/G-5) saphenum</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.001	<i>Prevotella sp. HMT 304</i>	0.001
20	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013										

Table 7: Beta-diversity pairwise comparisons on the periodontitis statuses

Statistically significant (p-value) was determined by the PERMANOVA test.

Group 1	Group 2	p-value
Healthy	Stage I	0.001
Healthy	Stage II	0.001
Healthy	Stage III	0.001
Stage I	Stage II	0.001
Stage I	Stage III	0.001
Stage II	Stage III	0.737

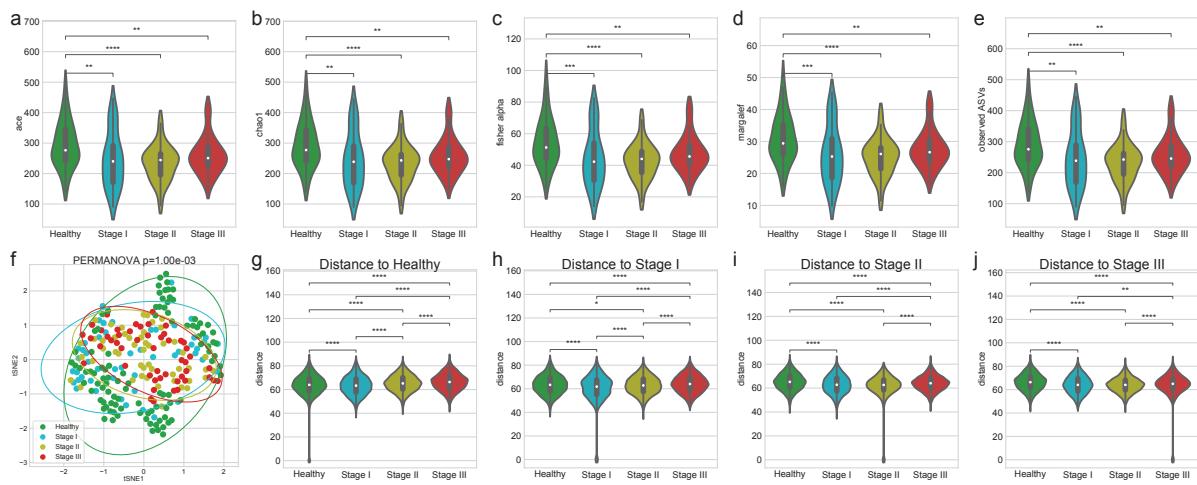


Figure 7: Diversity indices.

Alpha-diversity indices (**a-e**) indicate that healthy controls have increased heterogeneity than periodontitis stages as measured by: (**a**) ace (**b**) chao1 (**c**) Fisher alpha (**d**) Margalef, and (**e**) observed ASVs. (**f**) The beta-diversity index (weighted UniFrac) was visualized using a tSNE-transformed plot. The confidence ellipses are shown to display the distribution of each periodontitis stage. The distance to each stage demonstrated that each periodontitis stage was distinguished from the other periodontitis stages: (**g**) distance to Healthy (**h**) distance to Stage I (**i**) distance to Stage II, and (**j**) distance to Stage III. Statistical significance determined by the MWU test and the PERMANOVA test: $p \leq 0.01$ (**) and $p \leq 0.0001$ (****).

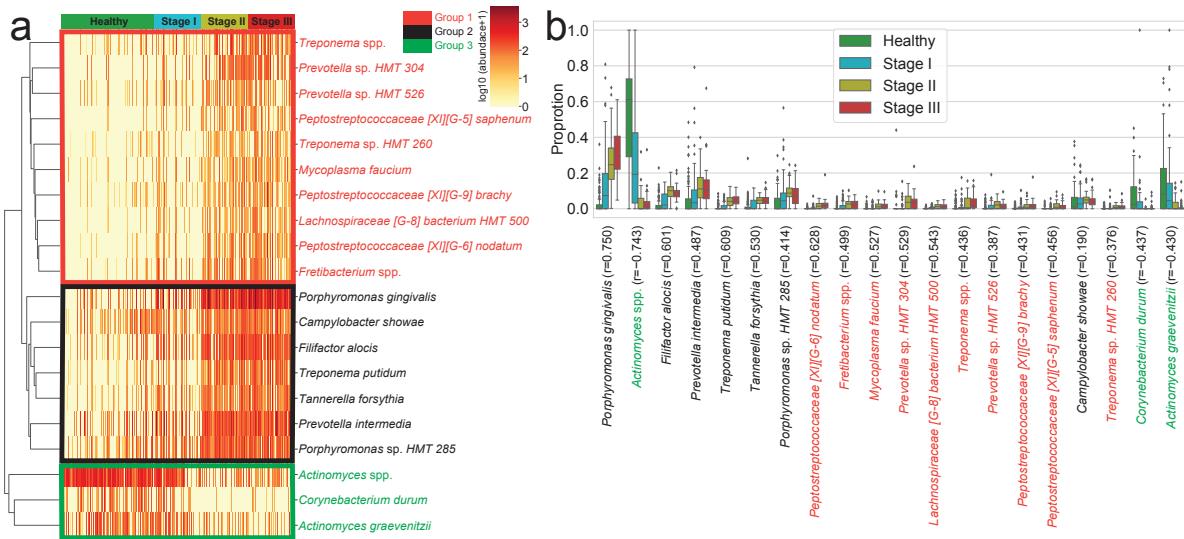


Figure 8: **Differentially abundant taxa (DAT).**

DAT that were identified by ANCOM. **(a)** Heatmap of clustered DAT with similar distribution among subjects. Group 1, Group 2, and Group 3 are marked in red, black, and green, respectively. **(b)** Box plots showing the proportions of DAT. Taxa were sorted by their importance according to ANCOM.

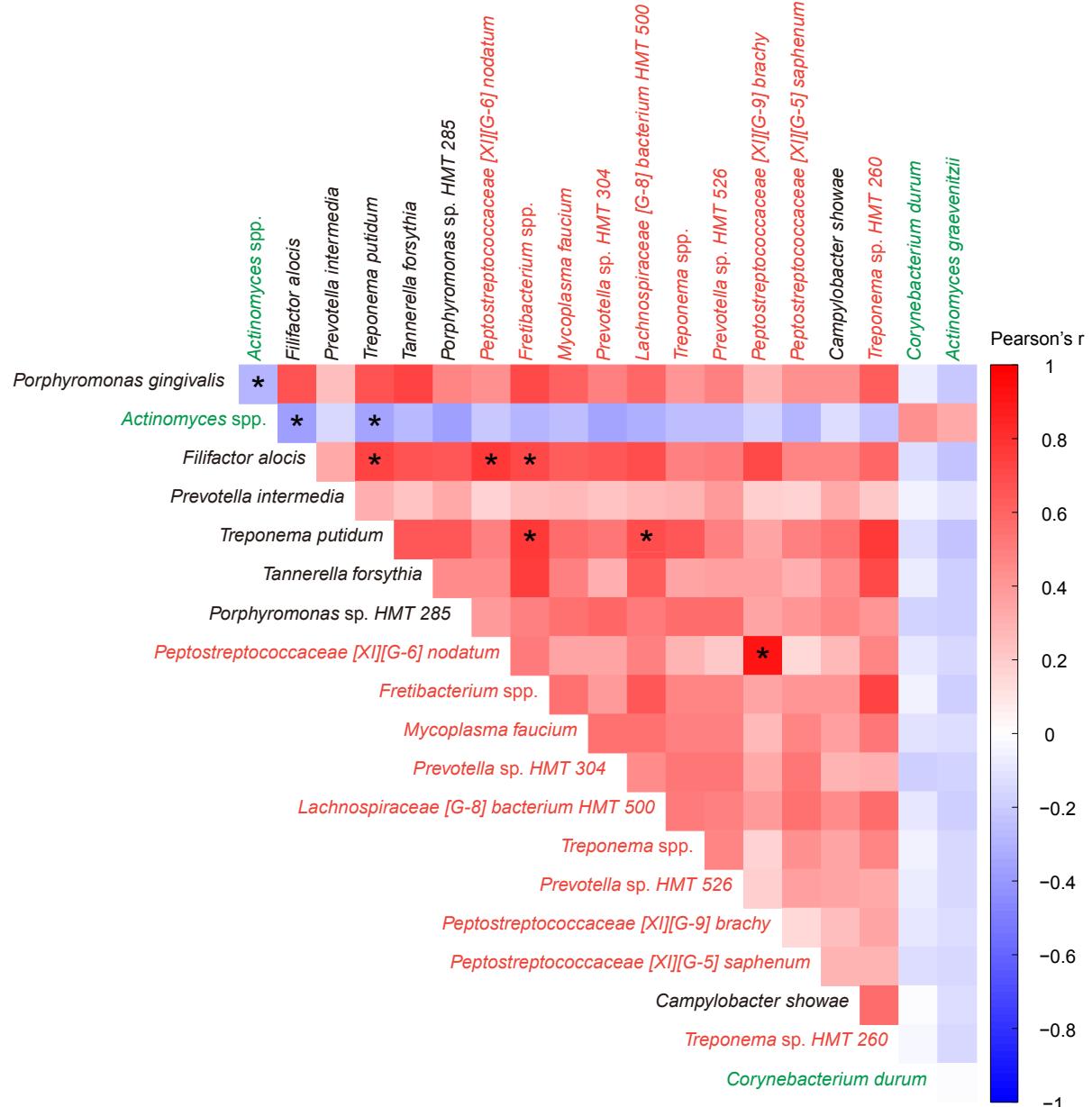


Figure 9: Correlation heatmap.

Pearson's correlations between DAT in healthy status and periodontitis stages. Statistical significance was determined by strong correlation, i.e., $| \text{coefficient} | \geq 0.5$ (*).

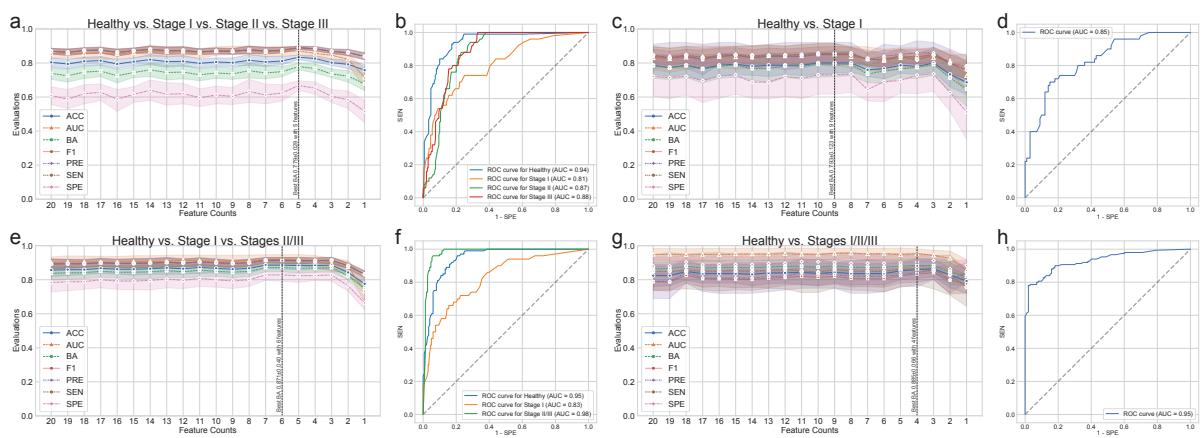


Figure 10: Random forest classification metrics.

The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. **(a)** Classification performance for healthy vs. stage I vs. stage II vs. stage III. **(b)** ROC curve for the highest BA of (a). **(c)** Classification performance for healthy vs. stage I. **(d)** ROC curve on the highest BA of (c). **(e)** Classification performance for healthy vs. stage I vs. stages II/III. **(f)** ROC curve for the highest BA of (e). **(g)** Classification performance for healthy vs. stages I/II/III. **(h)** ROC curve for the highest BA of (h).

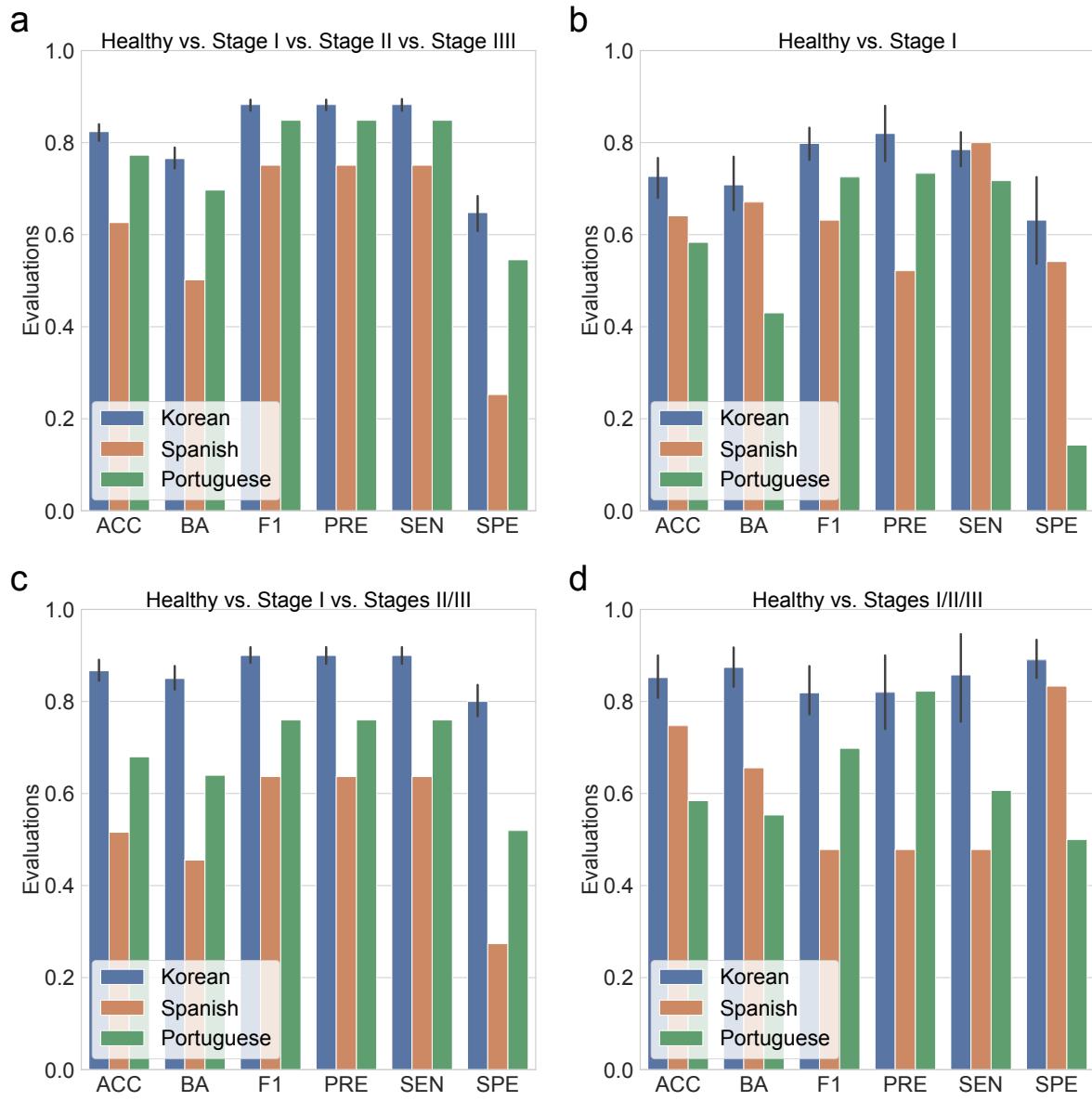


Figure 11: **Random forest classification metrics from external datasets.**

The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. **(a)** Classification performance for healthy vs. stage I vs. stage II vs. stage III. **(b)** Classification performance for healthy vs. stage I. **(c)** Classification performance for healthy vs. stage I vs. stages II/III. **(d)** Classification performance for healthy vs. stages I/II/III.

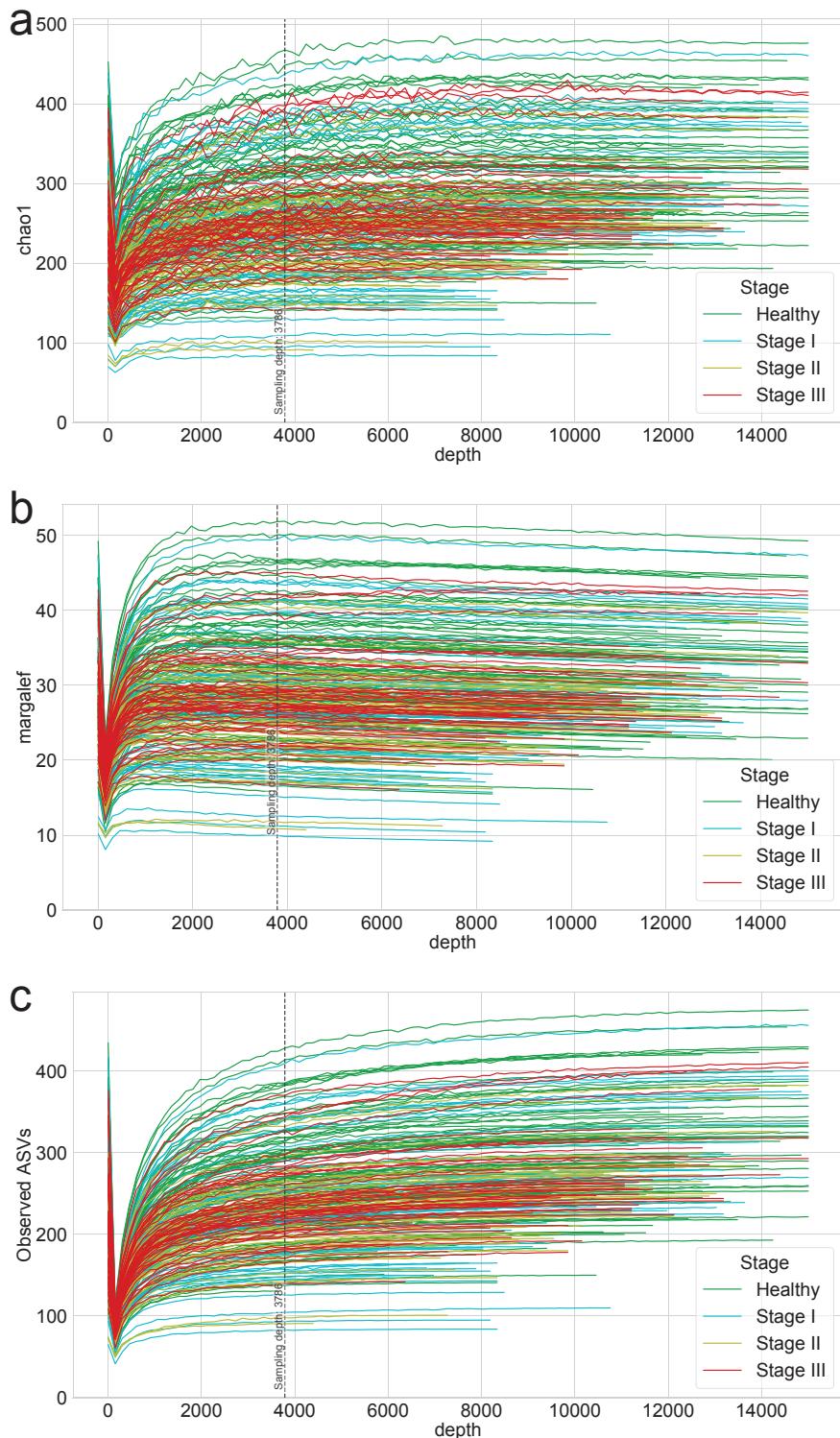


Figure 12: Rarefaction curves for alpha-diversity indices.

Rarefaction of (a) chao1 (b) margalef, and (c) observed ASVs were generated to measure species richness and determine the sampling depth of each sample.

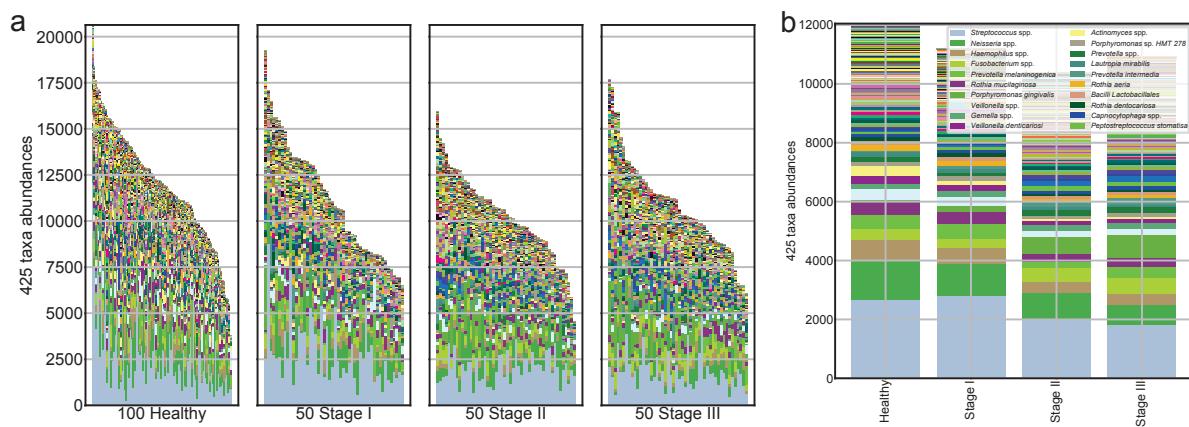


Figure 13: Salivary microbiome compositions in the different periodontal statuses.

Stacked bar plot of the absolute abundance of bacterial species for all samples (a) and the mean absolute abundance of bacterial species in the healthy, stage I, stage II, and stage III groups (b).

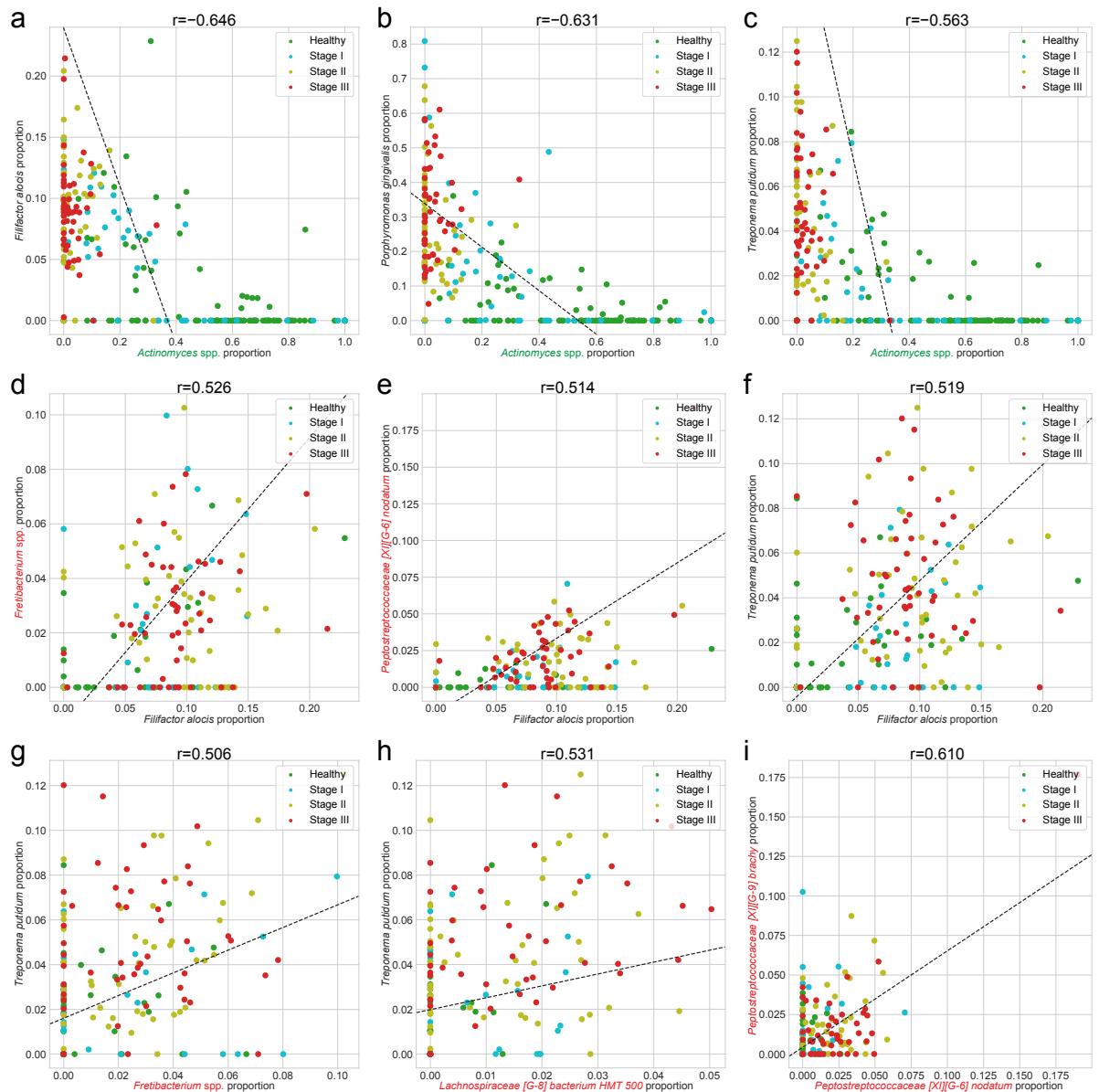


Figure 14: Correlation plots for differentially abundant taxa.

We selected the combinations of DAT with absolute Spearman correlation coefficients greater than 0.5. The color represents periodontal healthy periodontal statuses (green: healthy, cyan: stage I, yellow: stage II, and red: stage III).

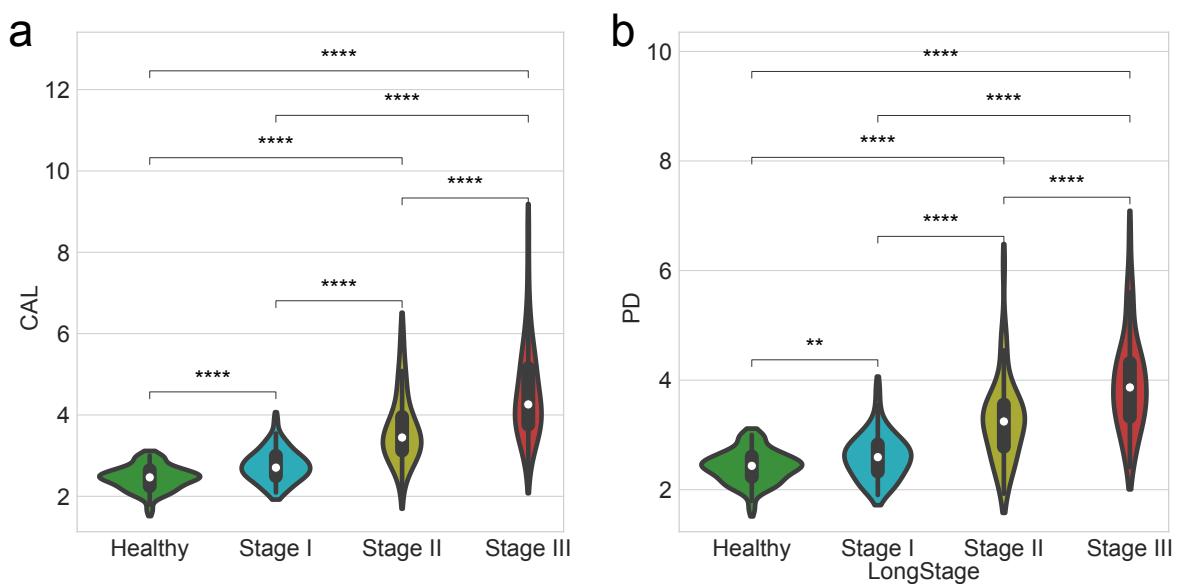


Figure 15: Clinical measurements by the periodontitis statuses.

Comparisons of clinical measurement among healthy controls and patients with various periodontitis stages. **(a)** Clinical attachment level (CAL) **(b)** Probing depth (PD). Statistical significance determined by the MWU test: $p \leq 0.01$ (**) and $p \leq 0.0001$ (****).

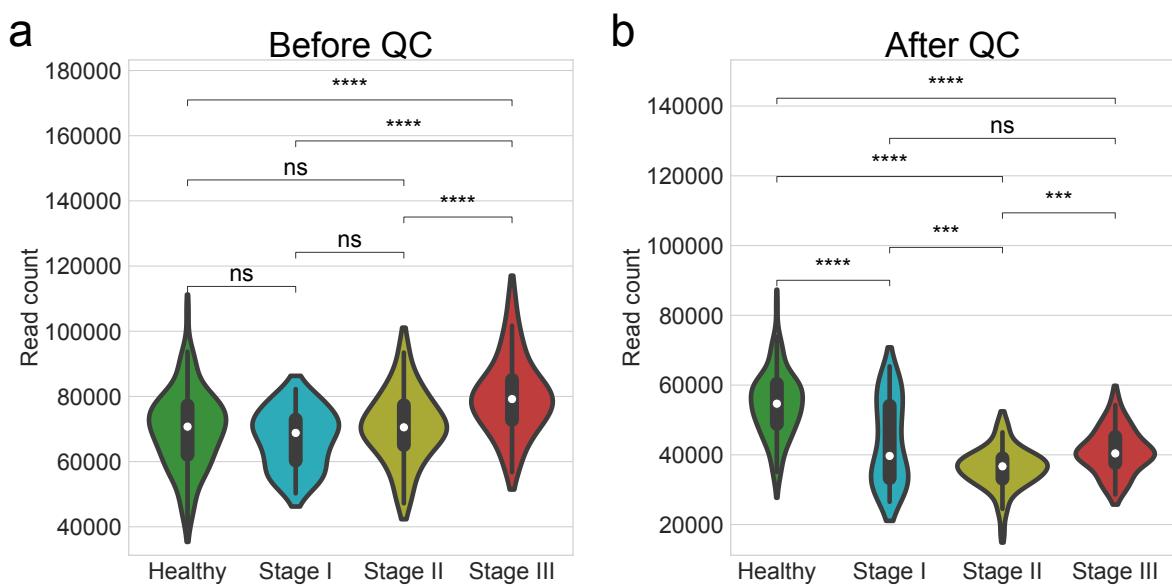


Figure 16: **Number of read counts by the periodontitis statuses.**

Comparisons of the number of read counts among healthy controls and patients with various periodontitis stages. **(a)** Before quality check **(b)** After quality check. Statistical significance determined by the MWU test: $p > 0.05$ (ns), $p \leq 0.001$ (***) , and $p \leq 0.0001$ (****).

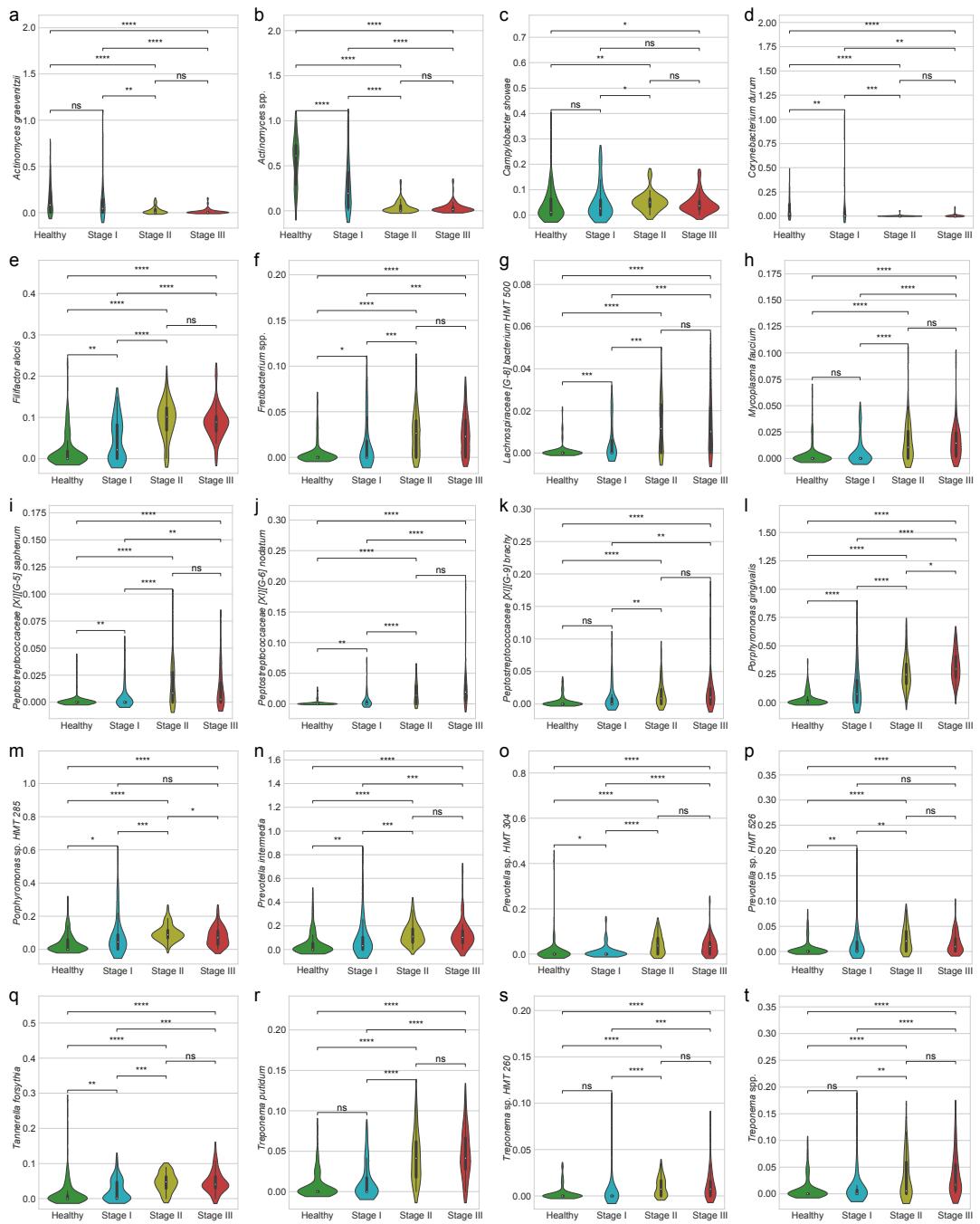


Figure 17: Proportion of DAT.

(a) *Actinomyces graevenitzii* (b) *Actinomyces* spp. (c) *Campylobacter showae* (d) *Corynebacterium durum* (e) *Filifactor alocis* (f) *Fretibacterium* spp. (g) *Lachnospiraceae [G-8] bacterium HMT 500* (h) *Mycoplasma faecium* (i) *Peptostreptococcaceae [XI][G-5] saphenum* (j) *Peptostreptococcaceae [XI][G-6] nodatum* (k) *Peptostreptococcaceae [XI][G-9] brachy* (l) *Porphyromonas gingivalis* (m) *Porphyromonas* sp. HMT 285 (n) *Prevotella intermedia* (o) *Prevotella* sp. HMT 304 (p) *Prevotella* sp. HMT 526 (q) *Tannerella forsythia* (r) *Treponema putidum* (s) *Treponema* sp. HMT 260 (t) *Treponema* spp. Statistical significance determined by the MWU test: $p > 0.05$ (ns), $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), and $p \leq 0.0001$ (****).

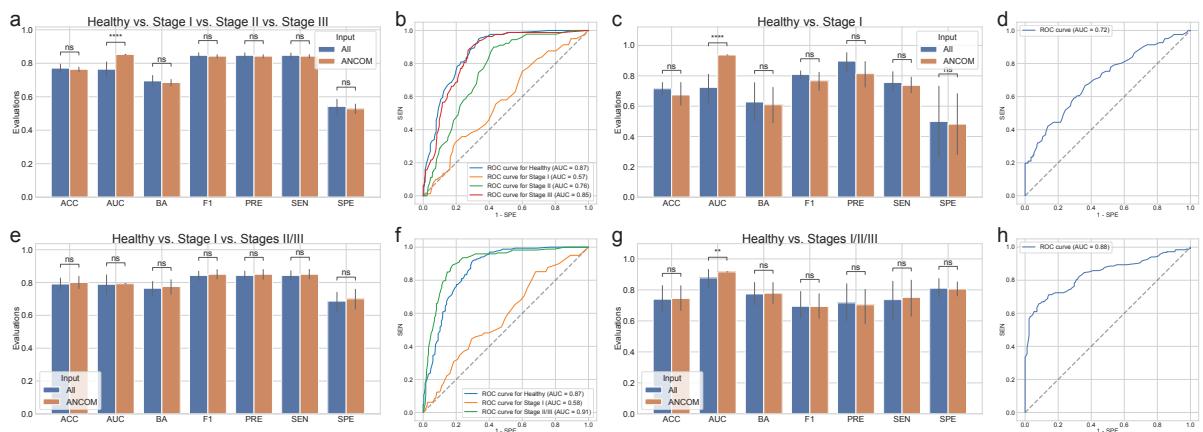


Figure 18: Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions.

The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. **(a)** Classification performance for healthy vs. stage I vs. stage II vs. stage III. **(b)** ROC curve for the highest BA of (a). **(c)** Classification performance for healthy vs. stage I. **(d)** ROC curve on the highest BA of (c). **(e)** Classification performance for healthy vs. stage I vs. stages II/III. **(f)** ROC curve for the highest BA of (e). **(g)** Classification performance for healthy vs. stages I/II/III. **(h)** ROC curve for the highest BA of (g). Statistical significance determined by the MWU test: $p > 0.05$ (ns), $p \leq 0.01$ (**), and $p \leq 0.0001$ (****).

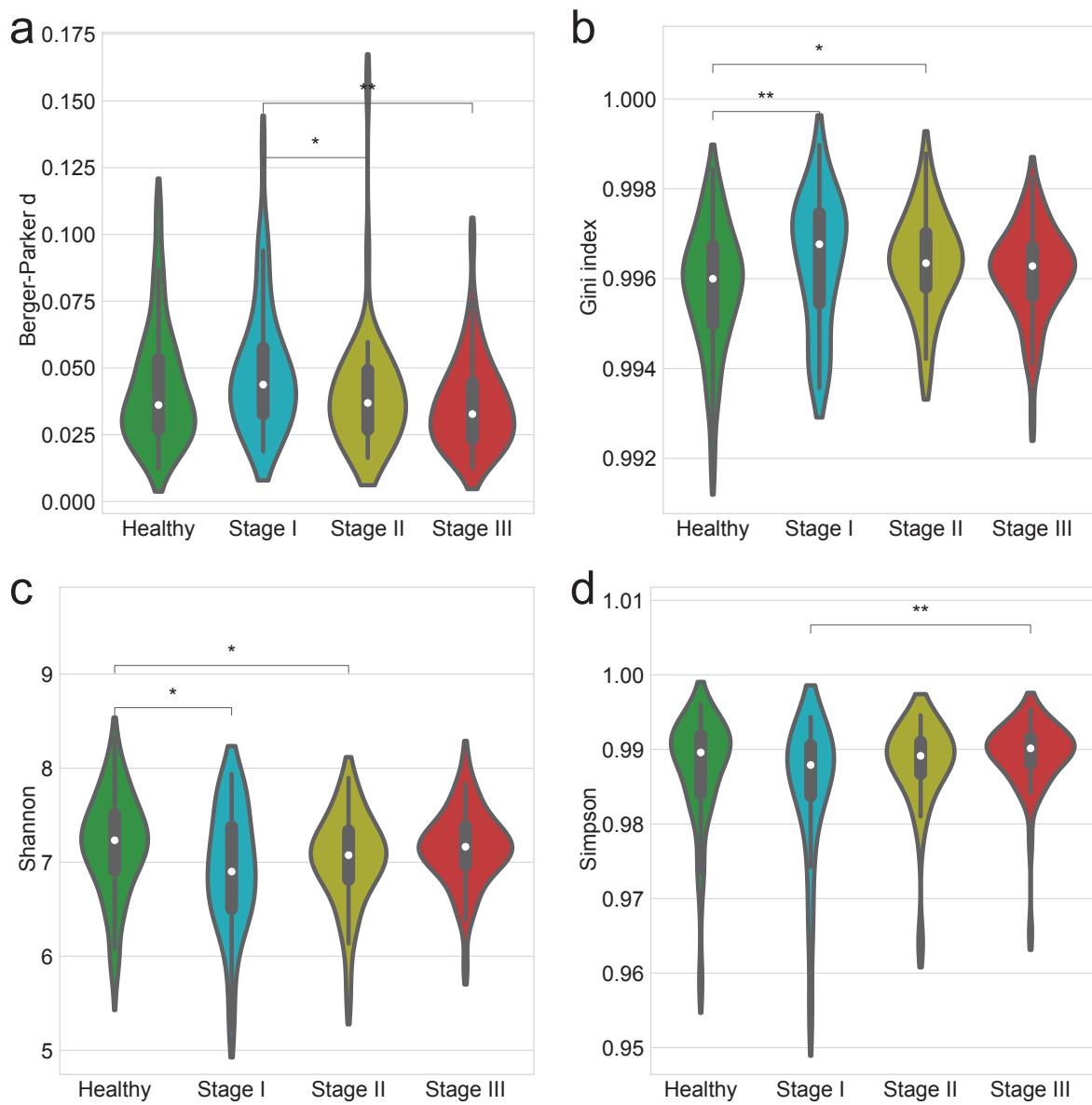


Figure 19: **Alpha-diversity indices account for evenness.**

Alpha-diversity indices (**a-d**) indicate that the heterogeneity between the periodontitis stages as measured by: **(a)** Berger-Parker *d* **(b)** Gini **(c)** Shannon **(d)** Simpson. Statistical significance determined by the MWU test: $p \leq 0.05$ (*) and $p \leq 0.01$ (**)

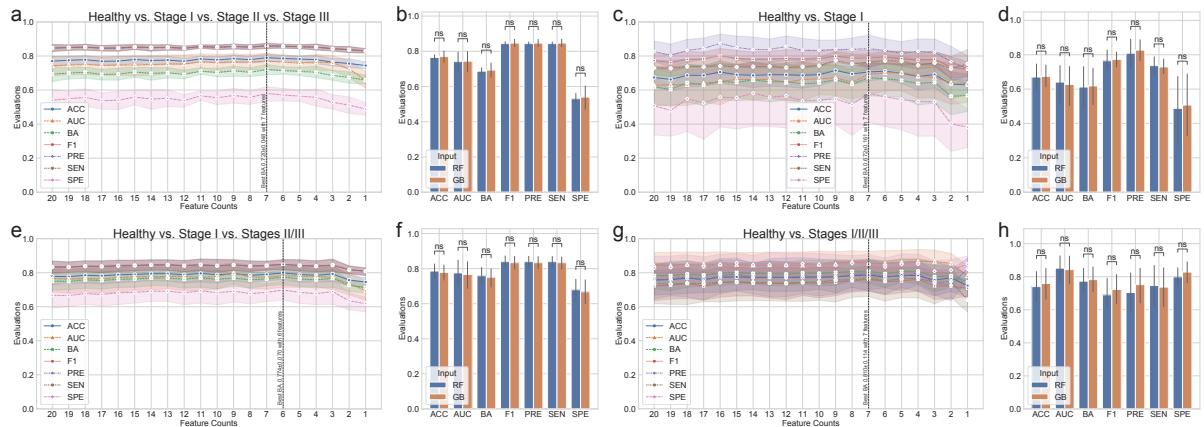


Figure 20: Gradient Boosting classification metrics.

The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. The feature counts mean that the classification model trained on the most important n features as the Table 5. **(a)** Comparison of Random forest (RF) and Gradient boosting (GB) for healthy vs. stage I vs. stage II vs. stage III. **(b)** Comparison of RF and GB for the highest BA of (a). **(c)** Classification performance for healthy vs. stage I. **(d)** Comparison of RF and GB for healthy vs. stage I vs. stages II/III. **(e)** Comparison of RF and GB for the highest BA of (d). **(f)** Comparison of RF and GB for Healthy vs. Stage I vs. Stages II/III. **(g)** Classification performance for healthy vs. stages I/II/III. **(h)** Comparison of RF and GB for Healthy vs. Stages I/II/III.

703 **3.4 Discussion**

704 In order to investigate at potential alterations in the salivary microbiome compositions based on periodontal
705 statuses, including healthy, stage I, stage II, and stage III, we employed 16S rRNA gene sequencing to
706 perform a cross-sectional periodontitis analysis. In this study, the 2018 periodontitis classification served
707 as the basis for the classification of periodontitis severities (Papapanou et al., 2018). There were notable
708 variations in the salivary microbiome composition among the multiple severities of periodontitis (Figure
709 13). Furthermore, our random forest classification model based on the proportions of DAT in the salivary
710 microbiome compositions across study participants to predict multiple periodontitis statuses with high
711 AUC of 0.870 ± 0.079 (Table 4).

712 Previous research identified the red complex as the primary pathogens of periodontitis (Listgarten,
713 1986): *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola*. Other studies, however,
714 have shown that periodontal pathogens communicate with other bacteria in the salivary microbiome
715 networks to generate dental plaque prior to the pathogenesis and development of periodontitis (Lamont &
716 Jenkinson, 2000; Rosan & Lamont, 2000; Yoshimura, Murakami, Nishikawa, Hasegawa, & Kawaminami,
717 2009).

718 Using subgingival plaque collections, recent researches have suggested a connection between the
719 periodontitis severity and the salivary microbiome compositions (Altabtbaei et al., 2021; Iniesta et al.,
720 2023; Nemoto et al., 2021). Therefore, we have examined the salivary microbiome compositions of
721 patients with multiple severities of periodontitis and periodontally healthy controls, extending on earlier
722 studies.

723 According to our findings, the salivary microbiome compositions have 425 taxa (Figure 13). We
724 computed the alpha-diversity indices to determine the variability within each salivary microbiome
725 composition, including ace (Chao & Lee, 1992), chao1 (Chao, 1984), fisher alpha (Fisher et al., 1943),
726 margalef (Magurran, 2021), observed ASVs (DeSantis et al., 2006), Berger-Parker *d* (Berger & Parker,
727 1970), Gini index (Gini, 1912), Shannon (Weaver, 1963), and Simpson (Simpson, 1949) (Figure 7 and
728 Figure 19). Alpha-diversity indices suggested that the microbial richness of periodontally healthy controls
729 was higher than that of patients with periodontitis (Figure 7a-e and Figure 19). These results are in line with
730 findings with that patients with advanced periodontitis, namely stage II and stage III, have less diversified
731 communities than periodontally healthy controls (Jorth et al., 2014). Recognizing that the periodontitis
732 severity increases the amount of *Porphyromonas gingivalis*, the salivary microbiome compositions from
733 periodontally healthy controls conserved microbial networks dominated by *Streptococcus* spp. (Figure
734 13). *Porphyromonas gingivalis* is one of the known periodontal pathogen that could cause dysbiosys
735 in the salivary microbiomes, suggesting in the pathophysiology of periodontitis. Despite this finding,
736 earlier research found that subgingival microbiome of patients with periodontitis had a greater alpha-
737 diversity index (observed ASVs) than that of healthy controls (Iniesta et al., 2023), might due to the
738 different sampling sites between saliva and subgingival plaque. On the other hand, another research
739 has addressed significant discrepancies in alpha-diversity indices from subgingival plaque, saliva, and
740 tongue biofilms from healthy controls and periodontitis patients, resulting the highest alpha-diversity

741 index in saliva collections (Belstrøm et al., 2021). Moreover, early-stage periodontitis, namely stage I,
742 did not determine statisticall ysiginificant differences in alpha-diversity indices compared to advanced
743 periodontitis, including stage II and stage III (Figure 7a-e). Accordingly, saliva collection of stage I
744 periodontitis may exhibit heterogeneity, indicating a midpoint condition between a healthy state and
745 advanced periodontitis (stage II and stage III). Likewise, gingivitis is often associated with low abundances
746 of the majority of periodontal pathogens, including *Porphyromonas gingivalis*, *Tannerella forsythia*, and
747 *Treponema denticola* (Abusleme et al., 2021). Compared to healthy controls, patients with stage I
748 periodontitis have higher detection rates of *Porphyromonas gingivalis* and *Tannerella forsythia* (Tanner et
749 al., 2006, 2007).

750 Therefore, we calculated beta-diversity indices to analyze the differences between the study partici-
751 pants. The distances for the multiple stages of periodontitis, including stage I, stage II, and stage III, as
752 well as healthy controls (Figure 4g-j and Table 7), suggesting notable differences among the multiple
753 periodontitis severities. In other words, the composition of the salivary microbiome compositions varies
754 depending on the periodontitis stages, so that supporting the findings from a previous study (Iniesta et al.,
755 2023). Taken together that it is nearly impossible to fully restore the attachment level after it has been lost
756 due to the progression and development of periodontitis, the ability to rapidly screen for periodontitis in
757 its early phases using saliva collections would be highly beneficial for effective disease management and
758 treatment.

759 Of the total of 425 taxa in the salivary microbiome composition that have been identified (Figure 13),
760 ANCOM was applied to select 20 taxa as the DAT that indicated notable abundance variation among
761 the periodontitis severities (Figure 8 and Table 5). Three sub-groups were formed from the DAT using
762 hierarchical clustering (Figure 8a). Surprisingly, two of the red complex pathogens (Rôças, Siqueira Jr,
763 Santos, Coelho, & de Janeiro, 2001), *Porphyromonas gingivalis* and *Tannerella forsythia*, were classified
764 in Group 2 and were more prevalent in stage II and stage II periodontitis compared to healthy controls.
765 *Campylobacter showae* was additionally placed in Group 2 of the orange complex pathogens (Gambin et
766 al., 2021). Furthermoe, some of the DAT in Group 2 have reported their crucial roles in pathogenesis
767 and development of periodontitis: *Filifactor alocis* (Aruni et al., 2015), *Treponema putidum* (Wyss et
768 al., 2004), *Tannerella forsythia* (Stafford, Roy, Honma, & Sharma, 2012; W. Zhu & Lee, 2016), and
769 *Prevotella intermedia* (Karched, Bhardwaj, Qudeimat, Al-Khabbaz, & Ellepol, 2022). Taken together,
770 this indicates that DAT in Group 2 is essential to periodontitis. The portion of some Group 1 DAT,
771 including *Peptostreptococcaceae[XI][G-5] saphenum*, *Peptostreptococcaceae[XI][G-6] nodatum*, and
772 *Peptostreptococcaceae[XI][G-9] brachy*, in healthy controls and patients with periodontitis significantly
773 differed, according to earlier research (Lafaurie et al., 2022). These outcomes support our research,
774 implying that Group 1 DAT are also essential to the etiology and progression of periodontitis. However,
775 in contrast to patients with periodontitis, Group 3 DAT, namely *Corynebacterium durum* and *Actinomyces*
776 *graevenitzii*, were enriched in healthy controls, which is consistent with earlier research (Redanz et al.,
777 2021; Nibali et al., 2020).

778 In our correlation analysis (Figure 9), we have discovered strongly negative correlations (coefficient \leq
779 -0.5) between DAT of Group 3 and these of Group 1 and Group 2; we have also identified nine DAT

pairs with strong correlations (coefficient $\leq -0.5 \vee$ coefficient ≥ 0.5) (Figure 14). Interestingly, there were strongly negative correlations (coefficient ≤ -0.5) between Group 2 DAT and *Actinomyces* spp., taxa which belong to Group 3: *Filifactor alocis* (Figure 14a), *Porphyromonas gingivalis* (Figure 14b), and *Treponema putidum* (Figure 14c). Taken together that pathogens, including *Filifactor alocis* (Aja, Mangar, Fletcher, & Mishra, 2021; Hiranmayi, Sirisha, Rao, & Sudhakar, 2017), *Porphyromonas gingivalis* (Rôças et al., 2001), and *Treponema putidum* (Wyss et al., 2004), become dominant taxa in patients with stage III periodontitis. On the other hand, commensal salivary bacteria, such as *Actinomyces* spp., gradually declined. Additionally, several DAT from Group 1 and Group 2 exhibited strong positive correlations (coefficient ≥ 0.5) (Figure 14d-i). It has been established that all of these DAT from Group 1 and Group 2 are periodontal pathogens: *Filifactor alocis* (Aja et al., 2021; Hiranmayi et al., 2017), *Fretibacterium* spp. (Teles, Wang, Hajishengallis, Hasturk, & Marchesan, 2021), *Lachnospiraceae[G-8] bacterium HMT 500* (Lafaurie et al., 2022), *Peptostreptococcaceae[XI][G-6] nodatum* (Lafaurie et al., 2022; Haffajee, Teles, & Socransky, 2006), *Peptostreptococcaceae[XI][G-9] brachy* (Lafaurie et al., 2022), and *Treponema putidum* (Wyss et al., 2004). Thus, these fundamental roles of identified periodontal pathogens in the pathophysiology and progression of periodontitis are further supported by these strong positive correlations (coefficient ≥ 0.5), suggesting that advanced periodontitis, i.e., stage III, might arise from the additional DAT from Group 1 and Group 2.

Moreover, to predict periodontitis statuses from salivary microbiome composition, we have constructed machine-learning classification models based on random forest for four classification settings:

1. healthy vs. stage I vs. stage II vs. stage III
2. healthy vs. stage I
3. healthy vs. stage I vs. stages II/III
4. healthy vs. stages I/II/III

Porphyromonas gingivalis and *Actinomyces* spp. were the two most important taxa (feature) in all classification settings. This finding aligns with a recent study that identifies *Actinomyces* spp. as the most prevalent bacteria in both the healthy gingivitis controls, while *Porphyromonas gingivalis* is recognized as the most predominant taxon within the periodontitis subjects, based on analyses of subgingival plaque samples (Nemoto et al., 2021). We have previously developed machine learning models for the classification of periodontitis, with the objective of predicting the severities of chronic periodontitis by analyzing the copy numbers of nine known salivary bacteria species. We classified healthy controls and patients with periodontitis utilizing bacterial combinations in conjunction with a random forest model (E.-H. Kim et al., 2020):

- AUC: 94%
- BA: 84%
- SEN: 95%
- SPE: 72%

Another study established a machine-learning model for the classification of periodontitis, employing 266 species derived from the buccal microbiome (Na et al., 2020):

- AUC: 92%

- 819 • BA: 84%
820 • SEN: 94%
821 • SPE: 74%

822 By separating patients with periodontitis from healthy controls using only four DAT, *e.g.* *Actinomyces*
823 *graevenitzii*, *Actinomyces* spp., *Corynebacterium durum*, and *Porphyromonas gingivalis*, our machine
824 learning model performed better than previously published models (Figure 10, Table 4, and Table 6):

- 825 • AUC: $95.3\% \pm 4.9\%$
826 • BA: $88.5\% \pm 6.6\%$
827 • SEN: $86.4\% \pm 15.7\%$
828 • SPE: $90.5\% \pm 7.0\%$

829 This result showed that by detecting Group 3 bacteria that were substantially abundant in health
830 controls than patients with periodontitis, our study increased BA by at least 5% and SPE by at least 17%.

831 Furthermore, we have validated our machine-learning prediction model using openly accessible 16S
832 gene rRNA sequencing data from Portuguese (Iniesta et al., 2023) and Spanish participants (Relvas et
833 al., 2021) in order to ensure the consistency of our random forest classification model (Figure 11). Our
834 classification models employed in this study were primarily developed and assessed on Korean study par-
835 ticipants, which may limit their generalizability to other ethnic groups with different salivary microbiome
836 compositions (Premaraj et al., 2020; Renson et al., 2019). Therefore, the evaluations of this periodonti-
837 tis classification models can be affected by ethnic-specific variances and differences, highlighting the
838 necessity for additional validation and adjustment across a spectrum of ethnic backgrounds.

839 Regarding the clinical characteristics and potential confounders influencing the analysis of salivary
840 microbiome compositions connected with periodontitis severity, this study had a number of limitations
841 that were pointed out. We did not offer clinical information, such as the percentage of teeth, the percentage
842 of bleeding on probing, nor dental furcation involvement, even though we did gather information on
843 attachment level, probing depth, plaque index, and gingival index; this might have it challenging to present
844 thorough and in-depth data about periodontal health. Moreover, the broad age range may make it tougher
845 to evaluate the relationship between age and periodontitis statuses, providing the necessity for future
846 studies to consider into account more comprehensive clinical characteristics associated with periodontitis.
847 Additionally, potential confounders—*e.g.* body mass index and e-cigarette use—which might have affected
848 dental health and salivary microbiome composition were disregarding consideration in addition to smoking
849 status and systemic diseases. Thus, future research incorporating these components would offer a more
850 thorough knowledge of how lifestyle factors interact and affect the salivary microbiome composition and
851 periodontal health. Throughout, resolving these limitations will advance our understanding in pathogenesis
852 and development of periodontitis, offering significant novel insights on the causal connection between
853 systemic diseases and the salivary microbiome compositions.

854 **4 Colon microbiome**

855 **4.1 Introduction**

856 Colorectal cancer (CRC) is one of the most prevalent and life-threatening malignancies worldwide, with
857 its incidence influenced by a combination of genetic, environmental, and lifestyle factors. Established risk
858 factors include a often diet in red and processed meats, low fiber intake, obesity, cigarette smoking, alcohol
859 consumption, and a sedentary lifestyle, all of which contribute to chronic inflammation, mutagenesis, and
860 metabolic regulation. Additionally, underlying conditions, *e.g.* inflammatory bowel disease and familial
861 adenomatous polyposis, significantly increase risk of CRC due to persistent mucosal inflammation and
862 somatic mutations that promote tumorigenesis.

863 The gut microbiome plays a fundamental role in maintaining host health by helping digestion,
864 regulating metabolism, adjusting immune function, and even coordinating neurological processes by
865 the brain-gut axis. Comprising these gut microbiota, including, archaea, bacteria, fungi, and viruses, the
866 gut microbiome contributes to the synthesis of essential vitamins, and production of fatty acids, which
867 influence intestinal integrity and immune responses. Thus, well-balanced gut microbiome composition
868 modulates systemic immune function by interacting with gut-associated lymphoid tissue, shaping immune
869 tolerance and response to infections. Hence, emerging evidence suggests that dysbiosis in the gut
870 microbiome composition are associated not only a narrow range of diseases, *e.g.* diarrhea and enteritis,
871 but also a wide range of diseases, *e.g.* obesity, diabetes, and cancers.

872 Recent studies have highlighted the crucial role of the gut microbiome in tumorigenesis and progression
873 of CRC, with dysbiosis emerging as a potential risk factor. Dysbiosis in gut microbiome compositions
874 can promote tumorigenesis of many cancers, including CRC, through several signaling cascades, including
875 inflammation, mutagenesis, and altered metabolism in host. Certain bacteria species, such as
876 *Fusobacterium* genus, *Bacteroides* genus, and *Escherichia coli*, have been associated with development
877 and progression of CRC by producing pro-inflammatory signals, generating toxins including mutagens,
878 and disrupting the intestinal barriers including mucous surface. In contrast, beneficial bacteria, such as
879 *Lactobacillus* genus and *Bifidobacterium* genus, are regarded to apply protective roles by maintaining
880 homeostasis of gut microbiome compositions and regulating immune responses including inflammation.

881 (Novelty)

882 **4.2 Materials and methods**

883 **4.2.1 Study participants enrollment**

884 **4.2.2 DNA extraction procedure**

885 **4.2.3 Bioinformatics analysis**

886 **4.2.4 Data and code availability**

887 **4.3 Results**

888 **4.3.1 Summary of clinical characteristics**

889 **4.3.2 Gut microbiome compositions**

890 **4.3.3 Diversity indices**

891 **4.3.4 DAT selection**

Table 8: Clinical characteristics of the study participants

892 **4.4 Discussion**

893 **5 Conclusion**

894 In conclusion, the research described in this doctoral dissertation was conducted to identify significant ...

895 In the section 2, I show that

896 References

- 897 Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors
898 a unique microbiome. *Science translational medicine*, 6(237), 237ra65–237ra65.
- 899 Abusleme, L., Hoare, A., Hong, B.-Y., & Diaz, P. I. (2021). Microbial signatures of health, gingivitis,
900 and periodontitis. *Periodontology 2000*, 86(1), 57–78.
- 901 Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio
902 analysis and compositional distance. *Mathematical geology*, 32, 271–275.
- 903 Aja, E., Mangar, M., Fletcher, H., & Mishra, A. (2021). Filifactor alocis: recent insights and advances.
904 *Journal of dental research*, 100(8), 790–797.
- 905 Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, 8(1), 11.
- 906 Altabtbaei, K., Maney, P., Ganesan, S. M., Dabdoub, S. M., Nagaraja, H. N., & Kumar, P. S. (2021). Anna
907 karenina and the subgingival microbiome associated with periodontitis. *Microbiome*, 9, 1–15.
- 908 Altingöz, S. M., Kurgan, Ş., Önder, C., Serdar, M. A., Ünlütürk, U., Uyanık, M., ... Günhan, M.
909 (2021). Salivary and serum oxidative stress biomarkers and advanced glycation end products in
910 periodontitis patients with or without diabetes: A cross-sectional study. *Journal of periodontology*,
911 92(9), 1274–1285.
- 912 Alverdy, J., Hyoju, S., Weigerinck, M., & Gilbert, J. (2017). The gut microbiome and the mechanism of
913 surgical infection. *Journal of British Surgery*, 104(2), e14–e23.
- 914 Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref:
915 statistics reference online*, 1–15.
- 916 Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., & Fletcher, H. M. (2015). Filifactor
917 alocis—a new emerging periodontal pathogen. *Microbes and infection*, 17(7), 517–530.
- 918 Aziz, Q., & Thompson, D. G. (1998). Brain-gut axis in health and disease. *Gastroenterology*, 114(3),
919 559–578.
- 920 Baldelli, V., Scaldaferrri, F., Putignani, L., & Del Chierico, F. (2021). The role of enterobacteriaceae in
921 gut microbiota dysbiosis in inflammatory bowel diseases. *Microorganisms*, 9(4), 697.
- 922 Barlow, G. M., Yu, A., & Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus.
923 *Nutrition in clinical practice*, 30(6), 787–797.
- 924 Basavaprabhu, H., Sonu, K., & Prabha, R. (2020). Mechanistic insights into the action of probiotics
925 against bacterial vaginosis and its mediated preterm birth: An overview. *Microbial pathogenesis*,
926 141, 104029.

- 927 Belstrøm, D., Constancias, F., Drautz-Moses, D. I., Schuster, S. C., Veleba, M., Mahé, F., & Givskov, M.
928 (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. *npj*
929 *Biofilms and Microbiomes*, 7(1), 76.
- 930 Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments.
931 *Science*, 168(3937), 1345–1347.
- 932 Berghella, V. (2012). Universal cervical length screening for prediction and prevention of preterm birth.
933 *Obstetrical & gynecological survey*, 67(10), 653–657.
- 934 Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., ... others (2012).
935 National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends
936 since 1990 for selected countries: a systematic analysis and implications. *The lancet*, 379(9832),
937 2162–2172.
- 938 Bolstad, A., Jensen, H. B., & Bakken, V. (1996). Taxonomy, biology, and periodontal aspects of
939 *fusobacterium nucleatum*. *Clinical microbiology reviews*, 9(1), 55–71.
- 940 Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... others
941 (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2.
942 *Nature biotechnology*, 37(8), 852–857.
- 943 Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.
- 944 Brennan, C. A., & Garrett, W. S. (2019). *Fusobacterium nucleatum*—symbiont, opportunist and
945 oncobacterium. *Nature Reviews Microbiology*, 17(3), 156–166.
- 946 Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier
947 ensembles by using random feature subsets. *Pattern recognition*, 36(6), 1291–1302.
- 948 Cai, Y., Li, Y., Xiong, Y., Geng, X., Kang, Y., & Yang, Y. (2024). Diabetic foot exacerbates gut
949 mycobiome dysbiosis in adult patients with type 2 diabetes mellitus: revealing diagnostic markers.
950 *Nutrition & Diabetes*, 14(1), 71.
- 951 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016).
952 Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7),
953 581–583.
- 954 Canakci, V., & Canakci, C. F. (2007). Pain levels in patients during periodontal probing and mechanical
955 non-surgical therapy. *Clinical oral investigations*, 11, 377–383.
- 956 Cappellato, M., Baruzzo, G., & Di Camillo, B. (2022). Investigating differential abundance methods in
957 microbiome data: A benchmark study. *PLoS computational biology*, 18(9), e1010467.
- 958 Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., & Schröder, H. (2018).
959 The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*,
960 2018(1), 4095789.
- 961 Centor, R. M. (1991). Signal detectability: the use of roc curves and their analyses. *Medical decision
making*, 11(2), 102–106.
- 962 Champagne, C., McNairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing
963 classification accuracy and confidence for agricultural land use mapping in canada. *International
Journal of Applied Earth Observation and Geoinformation*, 29, 44–52.

- 966 Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian
967 Journal of statistics*, 265–270.
- 968 Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the
969 American statistical Association*, 87(417), 210–217.
- 970 Chapple, I. L., Mealey, B. L., Van Dyke, T. E., Bartold, P. M., Dommisch, H., Eickholz, P., ... others
971 (2018). Periodontal health and gingival diseases and conditions on an intact and a reduced
972 periodontium: Consensus report of workgroup 1 of the 2017 world workshop on the classification
973 of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S74–S84.
- 974 Chen, T., Marsh, P., & Al-Hebshi, N. (2022). Smdi: an index for measuring subgingival microbial
975 dysbiosis. *Journal of dental research*, 101(3), 331–338.
- 976 Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The human
977 oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and
978 genomic information. *Database*, 2010.
- 979 Chen, X., D’Souza, R., & Hong, S.-T. (2013). The role of gut microbiota in the gut-brain axis: current
980 challenges and perspectives. *Protein & cell*, 4, 403–414.
- 981 Chew, R. J. J., Tan, K. S., Chen, T., Al-Hebshi, N. N., & Goh, C. E. (2024). Quantifying periodontitis-
982 associated oral dysbiosis in tongue and saliva microbiomes—an integrated data analysis. *Journal
983 of Periodontology*.
- 984 Čižmárová, B., Tomečková, V., Hubková, B., Hurajtová, A., Ohlasová, J., & Birková, A. (2022). Salivary
985 redox homeostasis in human health and disease. *International Journal of Molecular Sciences*,
986 23(17), 10076.
- 987 Cullin, N., Antunes, C. A., Straussman, R., Stein-Thoeringer, C. K., & Elinav, E. (2021). Microbiome
988 and cancer. *Cancer Cell*, 39(10), 1317–1341.
- 989 DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L.
990 (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with
991 arb. *Applied and environmental microbiology*, 72(7), 5069–5072.
- 992 Doyle, R., Alber, D., Jones, H., Harris, K., Fitzgerald, F., Peebles, D., & Klein, N. (2014). Term and
993 preterm labour are associated with distinct microbial community structures in placental membranes
994 which are independent of mode of delivery. *Placenta*, 35(12), 1099–1101.
- 995 Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1),
996 1–10.
- 997 Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., ... others
998 (2019). The vaginal microbiome and preterm birth. *Nature medicine*, 25(6), 1012–1021.
- 999 Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and
1000 the number of individuals in a random sample of an animal population. *The Journal of Animal
1001 Ecology*, 42–58.
- 1002 Francescone, R., Hou, V., & Grivennikov, S. I. (2014). Microbiome, inflammation, and cancer. *The
1003 Cancer Journal*, 20(3), 181–189.
- 1004 Gambin, D. J., Vitali, F. C., De Carli, J. P., Mazzon, R. R., Gomes, B. P., Duque, T. M., & Trentin, M. S.

- 1005 (2021). Prevalence of red and orange microbial complexes in endodontic-periodontal lesions: a
1006 systematic review and meta-analysis. *Clinical Oral Investigations*, 1–14.
- 1007 Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current
1008 understanding of the human microbiome. *Nature medicine*, 24(4), 392–400.
- 1009 Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Tipografia di Paolo Cuppini,*
1010 *Bologna, Italy*, 156.
- 1011 Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm
1012 birth. *The lancet*, 371(9606), 75–84.
- 1013 Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). Roc curve estimation: An
1014 overview. *REVSTAT-Statistical journal*, 12(1), 1–20.
- 1015 Goodey, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). *The declaration of helsinki* (Vol. 335) (No.
1016 7621). British Medical Journal Publishing Group.
- 1017 Haffajee, A., Teles, R., & Socransky, S. (2006). Association of eubacterium nodatum and treponema
1018 denticola with human periodontitis lesions. *Oral microbiology and immunology*, 21(5), 269–282.
- 1019 Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation.
1020 *Nature reviews immunology*, 15(1), 30–44.
- 1021 Hamjane, N., Mechita, M. B., Nourouti, N. G., & Barakat, A. (2024). Gut microbiota dysbiosis-associated
1022 obesity and its involvement in cardiovascular diseases and type 2 diabetes. a systematic review.
1023 *Microvascular Research*, 151, 104601.
- 1024 Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*,
1025 29(2), 147–160.
- 1026 Han, Y. W. (2015). Fusobacterium nucleatum: a commensal-turned pathogen. *Current opinion in
1027 microbiology*, 23, 141–147.
- 1028 Han, Y. W., & Wang, X. (2013). Mobile microbiome: oral bacteria in extra-oral infections and
1029 inflammation. *Journal of dental research*, 92(6), 485–491.
- 1030 Hand, D. J. (2012). Assessing the performance of classification methods. *International Statistical Review*,
1031 80(3), 400–414.
- 1032 Hartstra, A. V., Bouter, K. E., Bäckhed, F., & Nieuwdorp, M. (2015). Insights into the role of the
1033 microbiome in obesity and type 2 diabetes. *Diabetes care*, 38(1), 159–165.
- 1034 Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome,
1035 cancer, and cancer therapy. *Nature medicine*, 25(3), 377–388.
- 1036 Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2),
1037 427–432.
- 1038 Hiranmayi, K. V., Sirisha, K., Rao, M. R., & Sudhakar, P. (2017). Novel pathogens in periodontal
1039 microbiology. *Journal of Pharmacy and Bioallied Sciences*, 9(3), 155–163.
- 1040 Honda, K., & Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annual
1041 review of immunology*, 30(1), 759–795.
- 1042 Honest, H., Forbes, C., Durée, K., Norman, G., Duffy, S., Tsourapas, A., ... others (2009). Screening to
1043 prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with

- 1044 economic modelling. *Health Technol Assess*, 13(43), 1–627.
- 1045 Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M.-G., ... J. K. (2023). Predicting preterm
1046 birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.
- 1047 Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations.
1048 *International journal of data mining & knowledge management process*, 5(2), 1.
- 1049 Huang, R.-Y., Lin, C.-D., Lee, M.-S., Yeh, C.-L., Shen, E.-C., Chiang, C.-Y., ... Fu, E. (2007). Mandibular
1050 disto-lingual root: a consideration in periodontal therapy. *Journal of periodontology*, 78(8), 1485–
1051 1490.
- 1052 Iams, J. D., & Berghella, V. (2010). Care for women with prior preterm birth. *American journal of
1053 obstetrics and gynecology*, 203(2), 89–100.
- 1054 Ide, M., & Papapanou, P. N. (2013). Epidemiology of association between maternal periodontal
1055 disease and adverse pregnancy outcomes—systematic review. *Journal of clinical periodontology*,
1056 40, S181–S194.
- 1057 Iniesta, M., Chamorro, C., Ambrosio, N., Marín, M. J., Sanz, M., & Herrera, D. (2023). Subgingival
1058 microbiome in periodontal health, gingivitis and different stages of periodontitis. *Journal of
1059 Clinical Periodontology*, 50(7), 905–920.
- 1060 Janda, J. M., & Abbott, S. L. (2007). 16s rRNA gene sequencing for bacterial identification in the diagnostic
1061 laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.
- 1062 Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach
1063 for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29),
1064 5320–5334.
- 1065 John, G. K., & Mullin, G. E. (2016). The gut microbiome and obesity. *Current oncology reports*, 18,
1066 1–7.
- 1067 Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., ... others (2019).
1068 Evaluation of 16s rRNA gene sequencing for species and strain-level microbiome analysis. *Nature
1069 communications*, 10(1), 5029.
- 1070 Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics
1071 of the human oral microbiome during health and disease. *MBio*, 5(2), 10–1128.
- 1072 Kang, Y., Kang, X., Yang, H., Liu, H., Yang, X., Liu, Q., ... others (2022). Lactobacillus acidophilus ame-
1073 liorates obesity in mice through modulation of gut microbiota dysbiosis and intestinal permeability.
1074 *Pharmacological research*, 175, 106020.
- 1075 Karched, M., Bhardwaj, R. G., Qudeimat, M., Al-Khabbaz, A., & Ellepola, A. (2022). Proteomic analysis
1076 of the periodontal pathogen Prevotella intermedia secretomes in biofilm and planktonic lifestyles.
1077 *Scientific Reports*, 12(1), 5636.
- 1078 Katz, J., Chegini, N., Shiverick, K., & Lamont, R. (2009). Localization of *P. gingivalis* in preterm delivery
1079 placenta. *Journal of dental research*, 88(6), 575–578.
- 1080 Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., ... Li, H. (2015).
1081 Power and sample-size estimation for microbiome studies using pairwise distances and permanova.
1082 *Bioinformatics*, 31(15), 2461–2468.

- 1083 Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., ... Isaacson, R. E. (2017).
1084 Deciphering diversity indices for a better understanding of microbial communities. *Journal of*
1085 *Microbiology and Biotechnology*, 27(12), 2089–2093.
- 1086 Kim, C. H. (2018). Immune regulation by microbiome metabolites. *Immunology*, 154(2), 220–229.
- 1087 Kim, E.-H., Kim, S., Kim, H.-J., Jeong, H.-o., Lee, J., Jang, J., ... others (2020). Prediction of chronic
1088 periodontitis severity using machine learning models based on salivary bacterial copy number.
1089 *Frontiers in Cellular and Infection Microbiology*, 10, 571515.
- 1090 Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and
1091 bootstrap. *Computational statistics & data analysis*, 53(11), 3735–3745.
- 1092 Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. *Nature reviews*
1093 *Disease primers*, 3(1), 1–14.
- 1094 Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Caciato, S., ... MacIntyre,
1095 D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone
1096 treatment for preterm birth risk. *Microbiome*, 5, 1–14.
- 1097 Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune
1098 system. *Poultry science*, 99(4), 1906–1913.
- 1099 Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G., Getz, G., & Meyerson, M. (2011).
1100 Pathseq: software to identify or discover microbes by deep sequencing of human tissue. *Nature*
1101 *biotechnology*, 29(5), 393–396.
- 1102 Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification
1103 and combining techniques. *Artificial Intelligence Review*, 26, 159–190.
- 1104 Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., ... oth-
1105 ers (2022). Differences in the subgingival microbiome according to stage of periodontitis: A
1106 comparison of two geographic regions. *PLoS one*, 17(8), e0273523.
- 1107 Lamont, R. J., & Jenkinson, H. F. (2000). Subgingival colonization by porphyromonas gingivalis. *Oral*
1108 *Microbiology and Immunology: Mini-review*, 15(6), 341–349.
- 1109 Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and
1110 host interactions. *Nature reviews microbiology*, 16(12), 745–759.
- 1111 Leitich, H., & Kaider, A. (2003). Fetal fibronectin—how useful is it in the prediction of preterm birth?
1112 *BJOG: An International Journal of Obstetrics & Gynaecology*, 110, 66–70.
- 1113 León, R., Silva, N., Ovalle, A., Chaparro, A., Ahumada, A., Gajardo, M., ... Gamonal, J. (2007).
1114 Detection of porphyromonas gingivalis in the amniotic fluid in pregnant women with a diagnosis
1115 of threatened premature labor. *Journal of periodontology*, 78(7), 1249–1255.
- 1116 Li, R., Miao, Z., Liu, Y., Chen, X., Wang, H., Su, J., & Chen, J. (2024). The brain–gut–bone axis in
1117 neurodegenerative diseases: insights, challenges, and future prospects. *Advanced Science*, 11(38),
1118 2307971.
- 1119 Li, X., Yu, D., Wang, Y., Yuan, H., Ning, X., Rui, B., ... Li, M. (2021). The intestinal dysbiosis of
1120 mothers with gestational diabetes mellitus (gdm) and its impact on the gut microbiota of their
1121 newborns. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2021(1), 3044534.

- 1122 Li, Y., Qian, F., Cheng, X., Wang, D., Wang, Y., Pan, Y., ... Tian, Y. (2023). Dysbiosis of oral microbiota
1123 and metabolite profiles associated with type 2 diabetes mellitus. *Microbiology spectrum*, 11(1),
1124 e03796–22.
- 1125 Lim, J. W., Park, T., Tong, Y. W., & Yu, Z. (2020). The microbiome driving anaerobic digestion and
1126 microbial analysis. In *Advances in bioenergy* (Vol. 5, pp. 1–61). Elsevier.
- 1127 Lin, H., Eggesbø, M., & Peddada, S. D. (2022). Linear and nonlinear correlation estimators unveil
1128 undescribed taxa interactions in microbiome data. *Nature communications*, 13(1), 4946.
- 1129 Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature
1130 communications*, 11(1), 3514.
- 1131 Lin, H., & Peddada, S. D. (2024). Multigroup analysis of compositions of microbiomes with covariate
1132 adjustments and repeated measures. *Nature Methods*, 21(1), 83–91.
- 1133 Listgarten, M. A. (1986). Pathogenesis of periodontitis. *Journal of clinical periodontology*, 13(5),
1134 418–425.
- 1135 Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome
1136 medicine*, 8, 1–11.
- 1137 López-Aladid, R., Fernández-Barat, L., Alcaraz-Serrano, V., Bueno-Freire, L., Vázquez, N., Pastor-
1138 Ibáñez, R., ... Torres, A. (2023). Determining the most accurate 16s rrna hypervariable region for
1139 taxonomic identification from respiratory samples. *Scientific reports*, 13(1), 3974.
- 1140 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for
1141 rna-seq data with deseq2. *Genome biology*, 15, 1–21.
- 1142 Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, 31(19), R1174–R1177.
- 1143 Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically
1144 larger than the other. *The annals of mathematical statistics*, 50–60.
- 1145 Manolis, A. A., Manolis, T. A., Melita, H., & Manolis, A. S. (2022). Gut microbiota and cardiovascular
1146 disease: symbiosis versus dysbiosis. *Current Medicinal Chemistry*, 29(23), 4050–4077.
- 1147 Martin, C. R., Osadchiy, V., Kalani, A., & Mayer, E. A. (2018). The brain-gut-microbiome axis. *Cellular
1148 and molecular gastroenterology and hepatology*, 6(2), 133–148.
- 1149 Mayer, E. A., Tillisch, K., Gupta, A., et al. (2015). Gut/brain axis and the microbiota. *The Journal of
1150 clinical investigation*, 125(3), 926–938.
- 1151 Melguizo-Rodríguez, L., Costela-Ruiz, V. J., Manzano-Moreno, F. J., Ruiz, C., & Illescas-Montes, R.
1152 (2020). Salivary biomarkers and their application in the diagnosis and monitoring of the most
1153 common oral pathologies. *International journal of molecular sciences*, 21(14), 5173.
- 1154 Miller, C. S., Ding, X., Dawson III, D. R., & Ebersole, J. L. (2021). Salivary biomarkers for discriminating
1155 periodontitis in the presence of diabetes. *Journal of clinical periodontology*, 48(2), 216–225.
- 1156 Morita, T., Yamazaki, Y., Mita, A., Takada, K., Seto, M., Nishinoue, N., ... Maeno, M. (2010). A cohort
1157 study on the association between periodontal disease and the development of metabolic syndrome.
1158 *Journal of periodontology*, 81(4), 512–519.
- 1159 Na, H. S., Kim, S. Y., Han, H., Kim, H.-J., Lee, J.-Y., Lee, J.-H., & Chung, J. (2020). Identification of
1160 potential oral microbial biomarkers for the diagnosis of periodontitis. *Journal of clinical medicine*,

- 1161 9(5), 1549.
- 1162 Nemoto, T., Shiba, T., Komatsu, K., Watanabe, T., Shimogishi, M., Shibasaki, M., ... others (2021).
1163 Discrimination of bacterial community structures among healthy, gingivitis, and periodontitis
1164 statuses through integrated metatranscriptomic and network analyses. *Msystems*, 6(6), e00886–21.
- 1165 Nesbitt, M. J., Reynolds, M. A., Shiau, H., Choe, K., Simonsick, E. M., & Ferrucci, L. (2010). Association
1166 of periodontitis and metabolic syndrome in the baltimore longitudinal study of aging. *Aging clinical
1167 and experimental research*, 22, 238–242.
- 1168 Nibali, L., Sousa, V., Davrandi, M., Spratt, D., Alyahya, Q., Dopico, J., & Donos, N. (2020). Differences
1169 in the periodontal microbiome of successfully treated and persistent aggressive periodontitis.
1170 *Journal of Clinical Periodontology*, 47(8), 980–990.
- 1171 Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Tomović, M. (2017). Evaluation of classification
1172 models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7(1),
1173 39.
- 1174 Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: review of
1175 methods with applications in diagnostic medicine. *Physics in Medicine & Biology*, 63(7), 07TR01.
- 1176 Offenbacher, S., Katz, V., Fertik, G., Collins, J., Boyd, D., Maynor, G., ... Beck, J. (1996). Periodontal
1177 infection as a possible risk factor for preterm low birth weight. *Journal of periodontology*, 67,
1178 1103–1113.
- 1179 Ojesina, A. I., Pedamallu, C. S., Kostic, A., Jung, J., Auclair, D., Lohr, J., ... Meyerson, M. (2013). High
1180 throughput sequencing-based pathogen discovery in multiple myeloma. *Blood*, 122(21), 5322.
- 1181 Omundiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine learning classification techniques
1182 for breast cancer diagnosis. In *Iop conference series: materials science and engineering* (Vol. 495,
1183 p. 012033).
- 1184 Pan, A. Y. (2021). Statistical analysis of microbiome data: the challenge of sparsity. *Current Opinion in
1185 Endocrine and Metabolic Research*, 19, 35–40.
- 1186 Papapanou, P. N., Sanz, M., Buduneli, N., Dietrich, T., Feres, M., Fine, D. H., ... others (2018).
1187 Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of
1188 periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S173–S182.
- 1189 Park, J., Park, S. H., Lee, D., Lee, J. E., Lee, D., Na, K. J., ... Im, H.-J. (2024). Detecting cancer microbiota
1190 using unmapped rna reads on spatial transcriptomics. *Cancer Research*, 84(6_Supplement), 4881–
1191 4881.
- 1192 Payne, M. S., Newnham, J. P., Doherty, D. A., Furfarro, L. L., Pendal, N. L., Loh, D. E., & Keelan, J. A.
1193 (2021). A specific bacterial dna signature in the vagina of australian women in midpregnancy
1194 predicts high risk of spontaneous preterm birth (the predict1000 study). *American journal of
1195 obstetrics and gynecology*, 224(2), 206–e1.
- 1196 Peirce, J. M., & Alviña, K. (2019). The role of inflammation and the gut microbiome in depression and
1197 anxiety. *Journal of neuroscience research*, 97(10), 1223–1241.
- 1198 Pezzino, S., Sofia, M., Greco, L. P., Litrico, G., Filippello, G., Sarvà, I., ... Latteri, S. (2023). Microbiome
1199 dysbiosis: a pathological mechanism at the intersection of obesity and glaucoma. *International*

- 1200 *Journal of Molecular Sciences*, 24(2), 1166.
- 1201 Premaraj, T. S., Vella, R., Chung, J., Lin, Q., Hunter, P., Underwood, K., ... Zhou, Y. (2020). Ethnic
1202 variation of oral microbiota in children. *Scientific reports*, 10(1), 14788.
- 1203 Redanz, U., Redanz, S., Treerat, P., Prakasam, S., Lin, L.-J., Merritt, J., & Kreth, J. (2021). Differential
1204 response of oral mucosal and gingival cells to corynebacterium durum, streptococcus sanguinis, and
1205 porphyromonas gingivalis multispecies biofilms. *Frontiers in cellular and infection microbiology*,
1206 11, 686479.
- 1207 Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J., Cabral, C., ... Tomás, I.
1208 (2021). Relationship between dental and periodontal health status and the salivary microbiome:
1209 bacterial diversity, co-occurrence networks and predictive models. *Scientific reports*, 11(1), 929.
- 1210 Renson, A., Jones, H. E., Beghini, F., Segata, N., Zolnik, C. P., Usyk, M., ... others (2019). Sociodemo-
1211 graphic variation in the oral microbiome. *Annals of epidemiology*, 35, 73–80.
- 1212 Rideout, J. R., Caporaso, G., Bolyen, E., McDonald, D., Baeza, Y. V., Alastuey, J. C., ... Sharma, K.
1213 (2018, December). *biocore/scikit-bio: scikit-bio 0.5.5: More compositional methods added*. Zenodo.
1214 Retrieved from <https://doi.org/10.5281/zenodo.2254379> doi: 10.5281/zenodo.2254379
- 1215 Rôças, I. N., Siqueira Jr, J. F., Santos, K. R., Coelho, A. M., & de Janeiro, R. (2001). “red com-
1216 plex”(bacteroides forsythus, porphyromonas gingivalis, and treponema denticola) in endodontic
1217 infections: a molecular approach. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology,*
1218 and *Endodontontology*, 91(4), 468–471.
- 1219 Romero, R., Dey, S. K., & Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science*,
1220 345(6198), 760–765.
- 1221 Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., ... others (2014). The
1222 composition and stability of the vaginal microbiota of normal pregnant women is different from
1223 that of non-pregnant women. *Microbiome*, 2, 1–19.
- 1224 Rosan, B., & Lamont, R. J. (2000). Dental plaque formation. *Microbes and infection*, 2(13), 1599–1607.
- 1225 Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, 13(11),
1226 800–812.
- 1227 Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011).
1228 Metagenomic biomarker discovery and explanation. *Genome biology*, 12, 1–18.
- 1229 Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A
1230 survey and review. In *Emerging technology in modelling and graphics: Proceedings of iem graph*
1231 2018 (pp. 99–111).
- 1232 Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., & Knight, R. (2021). The
1233 microbiome and human cancer. *Science*, 371(6536), eabc4552.
- 1234 Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go?
1235 *The Journal of nutritional biochemistry*, 63, 101–108.
- 1236 Simpson, E. (1949). Measurement of diversity. *Nature*, 163.
- 1237 Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical
1238 length measurement for prediction of preterm birth in women with threatened preterm labor: a

- meta-analysis. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 35(1), 54–64.
- Spss, I., et al. (2011). Ibm spss statistics for windows, version 20.0. *New York: IBM Corp*, 440, 394.
- Stafford, G., Roy, S., Honma, K., & Sharma, A. (2012). Sialic acid, periodontal pathogens and tannerella forsythia: stick around and enjoy the feast! *Molecular Oral Microbiology*, 27(1), 11–22.
- Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., ... Mysorekar, I. U. (2013). Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. *American journal of obstetrics and gynecology*, 208(3), 226–e1.
- Sultan, S., El-Mowafy, M., Elgaml, A., Ahmed, T. A., Hassan, H., & Mottawea, W. (2021). Metabolic influences of gut microbiota dysbiosis on inflammatory bowel disease. *Frontiers in physiology*, 12, 715506.
- Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., & Wei, X. (2023). A review of normalization and differential abundance methods for microbiome counts data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1), e1586.
- Tanner, A. C., Kent Jr, R., Kanasi, E., Lu, S. C., Paster, B. J., Sonis, S. T., ... Van Dyke, T. E. (2007). Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults. *Journal of clinical periodontology*, 34(11), 917–930.
- Tanner, A. C., Paster, B. J., Lu, S. C., Kanasi, E., Kent Jr, R., Van Dyke, T., & Sonis, S. T. (2006). Subgingival and tongue microbiota during early periodontitis. *Journal of dental research*, 85(4), 318–323.
- Tejeda, M., Farrell, J., Zhu, C., Haines, J. L., Wang, L.-S., Schellenberg, G. D., ... others (2021). Multiple viruses detected in human dna are associated with alzheimer disease risk. *Alzheimer's & Dementia*, 17, e054585.
- Teles, F., Wang, Y., Hajishengallis, G., Hasturk, H., & Marchesan, J. T. (2021). Impact of systemic factors in shaping the periodontal microbiome. *Periodontology 2000*, 85(1), 126–160.
- Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*, 535(7610), 65–74.
- Tian, R., Liu, H., Feng, S., Wang, H., Wang, Y., Wang, Y., ... Zhang, S. (2021). Gut microbiota dysbiosis in stable coronary artery disease combined with type 2 diabetes mellitus influences cardiovascular prognosis. *Nutrition, Metabolism and Cardiovascular Diseases*, 31(5), 1454–1466.
- Tilg, H., Kaser, A., et al. (2011). Gut microbiome, obesity, and metabolic dysfunction. *The Journal of clinical investigation*, 121(6), 2126–2132.
- Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *Journal of periodontology*, 89, S159–S172.
- Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rrna gene. *Current opinion in microbiology*, 11(5), 442–446.
- Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, 92(2), 698–715.

- 1278 Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome.
1279 *Nutrition reviews*, 70(suppl_1), S38–S44.
- 1280 Vander Haar, E. L., So, J., Gyamfi-Bannerman, C., & Han, Y. W. (2018). *Fusobacterium nucleatum* and
1281 adverse pregnancy outcomes: epidemiological and mechanistic evidence. *Anaerobe*, 50, 55–59.
- 1282 Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning
1283 research*, 9(11).
- 1284 Walker, M. A., Pedamallu, C. S., Ojesina, A. I., Bullman, S., Sharpe, T., Whelan, C. W., & Meyerson, M.
1285 (2018). Gatk pathseq: a customizable computational tool for the discovery and identification of
1286 microbial sequences in libraries from eukaryotic hosts. *Bioinformatics*, 34(24), 4287–4289.
- 1287 Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.
- 1288 Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary
1289 tract—a role beyond infection. *Nature Reviews Urology*, 12(2), 81–90.
- 1290 Witkin, S. (2019). Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An
1291 International Journal of Obstetrics & Gynaecology*, 126(3), 359–359.
- 1292 Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE
1293 Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594.
- 1294 Wyss, C., Moter, A., Choi, B.-K., Dewhirst, F., Xue, Y., Schüpbach, P., ... Guggenheim, B. (2004).
1295 Treponema putidum sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of
1296 human periodontitis and acute necrotizing ulcerative gingivitis. *International journal of systematic
1297 and evolutionary microbiology*, 54(4), 1117–1122.
- 1298 Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome
1299 cancer research. *Gut Microbes*, 15(2), 2244139.
- 1300 Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods
1301 for automated emg signal classification. *BioMed research international*, 2019(1), 9152506.
- 1302 Yang, I., Claussen, H., Arthur, R. A., Hertzberg, V. S., Geurs, N., Corwin, E. J., & Dunlop, A. L. (2022).
1303 Subgingival microbiome in pregnancy and a potential relationship to early term birth. *Frontiers in
1304 cellular and infection microbiology*, 12, 873683.
- 1305 Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface
1306 components of *porphyromonas gingivalis*. *Journal of periodontal research*, 44(1), 1–12.
- 1307 Zhang, C.-Z., Cheng, X.-Q., Li, J.-Y., Zhang, P., Yi, P., Xu, X., & Zhou, X.-D. (2016). Saliva in the
1308 diagnosis of diseases. *International journal of oral science*, 8(3), 133–137.
- 1309 Zhu, W., & Lee, S.-W. (2016). Surface interactions between two of the main periodontal pathogens:
1310 *Porphyromonas gingivalis* and *tannerella forsythia*. *Journal of periodontal & implant science*,
1311 46(1), 2–9.
- 1312 Zhu, X., Han, Y., Du, J., Liu, R., Jin, K., & Yi, W. (2017). Microbiota-gut-brain axis and the central
1313 nervous system. *Oncotarget*, 8(32), 53829.

1314

Acknowledgments

1315 I would like to disclose my earnest appreciation for my advisor, Professor Semin Lee, who provided
 1316 solicitous supervision and cherished opportunities throughout the course of my research. His advice and
 1317 consultation encouraged me to become as a researcher and to receive all humility and gentleness. I am
 1318 also grateful to all of my committee members, Professor AAA, Professor BBB, Professor CCC, and
 1319 Professor DDD, for their critical and meaningful mentions and suggestions.

1320 I extend my deepest gratitude to my Lord, *the Flying Spaghetti Monster*, His Noodly Appendage
 1321 has guided me through the twist and turns of this academic journey. His presence, ever comforting and
 1322 mysterious, has been a source of strength and humor during both highs and lows. In moments of doubt, I
 1323 found solace in the belief that you were there, gently reminding me to keep faith in the process. His Holy
 1324 Noodle has nourished my mind, and for that, I am truly overwhelmed. May His Holy Noodle continue to
 1325 guide me in all my future endeavors. *R’Amen.*

1326 (Professors)

1327 I would like to extend my heartfelt gratitude to my colleagues of the Computational Biology Lab @
 1328 UNIST, whose collaboration, friendship, brotherhood, and support have been an invaluable part of my
 1329 journey. Your willingness to share insights, engage in thoughtful discussions, and offer encouragement
 1330 during the challenging moments of research has significantly shaped my academic experience. The
 1331 camaraderie in Computational Biology Lab made even the most demanding days more enjoyable, and I
 1332 am deeply grateful for the collaborative environment we created together. I appreciate you for standing
 1333 by my side throughout this Ph.D. journey.

1334 I would like to express my heartfelt gratitude to my family, whose unwavering support has been the
 1335 foundation of everything I have achieved. Your love, encouragement, and belief in me have sustained me
 1336 through every challenge, and I could not have come this far without you. From your words of wisdom to
 1337 your patience and understanding, each of you has played a vital role in helping me navigate this journey.
 1338 The strength and comfort I have drawn from our family bond have been my greatest source of resilience.
 1339 Your presence, both near and far, has filled my life with warmth and motivation. I am deeply grateful for
 1340 your unconditional love and for always being there when I needed you the most. Thank you for being my
 1341 constant source of strength and inspiration.

1342 I am incredibly pleased to my friends, especially my GSHS alumni (○망특), for their unwavering
 1343 support and encouragement throughout this journey. The bonds we formed back in our school days have
 1344 only grown stronger over the years, and I am fortunate to have had such loyal and understanding friends
 1345 by my side. Your constant words of motivation, and even moments of levity during stressful times have
 1346 helped keep me grounded. Whether it was a late-night conversations, a shared laugh, or a simple message
 1347 of reassurance, you all have played a vital role in keeping me focused and motivated. I am relieved for the
 1348 ways you celebrated each small achievement with me and how you patiently listened to my worries. The
 1349 memories of our shared past provided me with comfort and a sense of stability when the road ahead felt
 1350 uncertain. I could not have reached this point without the love and friendship that you all have generously
 1351 given. Each of your, in your unique way, has contributed to this dissertation, even if indirectly, and for

1352 that, I am forever beholden. I look forward to continuing our friendship as we all grow in our individual
1353 paths, knowing that the support we share is something truly special.

1354 (Girlfriend)

1355 I would like to express my sincere gratitude to the amazing members of my animal protection groups,
1356 DRDR (두루두루) and UNIMALS (유니멀스), whose dedication and compassion have been a constant
1357 source of motivation. Your unwavering commitment to improving the lives of animals has inspired me
1358 throughout this journey. I am also thankful for the beautiful cats we have cared for, whose presence
1359 brought both joy and purpose to our allegiance. Their playful spirits and gentle companionship served as
1360 daily reminders of why we continue to fight for animal rights. The bond we share, both with each other
1361 and with the animals we protect, has enriched my life in countless ways. I appreciate you all again for
1362 your support, dedication, and for being part of this meaningful cause.

1363 I would like to express my deepest gratitude to everyone I have had the honor of meeting throughout
1364 this journey. Your kindness, encouragement, and support have carried me through both the challenging
1365 and rewarding moments of my life. Whether through a kind word, thoughtful advice, or simply being
1366 there when I needed it most, your presence has made all the difference. I am incredibly fortunate to have
1367 received such generosity and warmth from those around me, and I do not take it for granted. Every act
1368 of kindness, no matter how big or small, has been a source of strength and motivation for me. To all
1369 my friends, colleagues, mentors, and beloved ones, thank you for your unwavering support. I am truly
1370 grateful for each of you, and your kindness has left an indelible mark on my journey.

1371 My Lord, *the Flying Spaghetti Monster*,
1372 give us grace to accept with serenity the things that cannot be changed,
1373 courage to change the things that should be changed,
1374 and the wisdom to distinguish the one from the other.

1375
1376 Glory be to *the Meatball*, to *the Sauce*, and to *the Holy Noodle*.
1377 As it was in the beginning, is now, and ever shall be.

1378 *R'Amen.*



May your progress be evident to all

