Methods

Data collection

We searched for periodontitis-related oral microbiome datasets in the NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject) database using the following keyword combinations: "('human metagenome'[Organism] OR human metagenome[All Fields]) AND oral[All Fields]" and "('human oral metagenome'[Organism] OR human oral metagenome[All Fields])"; please consult Figure S1 for the selection process and results.

Our collection criteria includes:

1) Case-control studies with clear disease information: This means that we only select studies where cases have been diagnosed with periodontitis disease, and healthy controls without periodontitis;
2) At least 10 valid samples in each case group and control group: This ensures that our results have sufficient stability and reproducibility;
3) No recent use of antibiotics: Antibiotic use may affect the composition of oral microbiota, thereby affecting the accuracy of our research.

After filtering out duplicates and data without detailed metadata or meeting minimum samples requirements, we selected a total of eight datasets. Among these, six datasets including both case and control were used for machine learning (ML) modeling and validation, and two datasets with only patient samples were used for independent testing of the ML models. The project accession numbers used for modeling include PRJDB11203, PRJNA230363, PRJNA396840, PRJNA678453, PRJNA717815, and PRJNA932553, and the project accession numbers used for external validation include PRJDB6966 and PRJNA552294 (Table S1-S2).

Raw metagenomic data processing

The raw sequencing data were downloaded from the NCBI Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra) [1]. We used Trimmomatic [2] (v.0.39) with TruSeq3 adapter files to remove adapter sequences and low-quality reads from the raw data. The TruSeq3-PE.fa file was used for paired-end sequences, while the TruSeq3-SE.fa file was used for single-end sequences. Reads shorter than 50bp were removed. Since our data was generated through metagenomic next-generation sequencing (mNGS), we utilized Bowtie2 [3] (v.2.5.1) to align the filtered reads to the human genome (hg19) to remove human DNA contaminations. The remaining reads were referred to as "clean data" and were used for subsequent analyses. Please consult Supplementary Table S2 for a complete list of projects, sample and run IDs used in this study.

Taxonomic and functional profiling

We used the MetaPhlAn4 [4] (v.4.0.3) software with default parameters for taxonomic analysis and retained the relative abundances at the species level for subsequent analysis. For the functional profiles including metabolic pathways, we used the HUMAnN3 [5] (v.3.6) software with default parameters.

To avoid the impact of lowly abundant taxonomic and functional entities, we filtered the species abundance spectrum and metabolic pathway abundance spectrum within each project, removing species and pathways with maximum relative abundance lower than 0.001 in all samples of the project, according to Li *et al.* [6] and the reference manual of “SIAMCAT” R package [7] (v.1.9.0, https://bioconductor.org/packages/SIAMCAT). The filtered microbial abundance data and metabolic pathway abundance data were were z-score standardization for subsequent modeling and statistical analysis.

Identification and removal of confounding factors within cohorts

Due to the influence of confounding factors, subsequent analyses may introduce bias. These issues can potentially compromise the reliability and effectiveness of subsequent biomarker identification and disease prediction modeling. To avoid these problems, we identified confounding factors for each project and subsequently eliminated their impact on the taxonomic and functional relative abundance profile to ensure the quality of results. More specifically, we examined all available factors in the metadata, such as age, gender, body mass index (BMI), disease stage, and geographic location. We tested for significant differences between the case group and the control group. For qualitative variables (including age and BMI), we used the Fisher‘s exact test, while for quantitative variables (including gender, disease stage, and geographic location), we used the non-parametric Wilcoxon rank sum test. We then used the “removeBatchEffect” function from the “limma” R Package [8] (v.3.56.0) to adjust the factors with $p<0.05$. Singnificant qualitative and quantitative variables were considered as covariates and batch factors, respectively, with other variables set as default.

Batch effect removal across cohorts

We evaluated the effectiveness of the state-of-art batch removal methods, including the “ComBat” function in the “sva” R package [9] (v.3.48.0), “removeBatchEffect” in the “limma” R package [8] (v.3.56.2), “adjust_batch” in the “MMUPHin” R package [10] (v.1.14.0), and “ConQuR” in the “ConQuR” R package [11] (v.2.0) that are widely used in current microbiome research. The batch effect was quantified based on the R-squared value from PERMANOVA testing and the distribution in PcoA plots (Figure S7). The smaller the R-square value, the smaller the proportion of batch effects in the model and the less impact it has on our model construction. Based on this criterion, we selected and employed the “adjust_batch” function in the “MMUPHin” R package (v.1.4.2) to reduce batch effects, using project ID and shared confounding factors as control factors. After removing confounding or reducing batch effects, subsequent modeling and biostatistical analysis were conducted on relative abundance data.

Biomarker identification using Linear discriminant analysis Effect Size (LEfSe)

We used the “run.lefse” function in the “microbiomeMarker” R package [12] (v.1.0.2) to perform LEfSe analysis to identify disease-specific taxa. The output of the LefSe analysis provides effect size scores, which represent the Linear Discriminant Analysis (LDA) scores, reflecting the degree of difference between case and control groups. The higher the score, the more significant the difference. Taxa with p<=0.05 and LDA score of 2 in at least one cohort or greater were considered biomarkers. In this study, we added a plus (+) or minus (-) sign to the score to indicate their enrichment in the case or control group, respectively. We extracted biomarkers and metabolic pathways that were enriched in three or more projects for further validation analysis and mapping.

Microbial network analysis

To characterize the relationships among the marker species and the resulting interaction networks, we used the spearman method in the “corr.test” function of the “psych” (v.2.3.9, https://CRAN.R-project.org/package=psych) package to calculate the correlation. Correlations with a p-value < 0.05 and an absolute value of the correlation coefficient > 0.5 were retained for further analysis. We used the “ggraph” package (v.4.3.2, https://CRAN.R-project.org/package=ggraph) to visualize the network, with the size of the nodes corresponding to the number of the projects, and positive and negative correlation edges painted green and red, respectively.

The normalized number of health-enriched species connectivity and disease-enriched species represent the ratio of each health-enriched species divided by the total number of health-enriched species (12) and each disease-enriched species divided by the total number of disease-enriched species (42). Two-sided Wilcoxon rank-sum test was used for group-wise comparisons.

Machine Learning Modeling, Validation and Testing

To check whether oral microbiome data can be used to distinguish periodontitis from healthy controls, we used the “SIAMCAT” R package [7] (v.1.9.0, https://bioconductor.org/packages/SIAMCAT) to construct disease grading classifiers (or models). Related abundances were normalized using the “normalize.features” function and then used as input for model training, validation and testing.

To select the best machine learning algorithm, we compared four methods including Elastic Net (Enet) [13], Lasso [14], Random Forest (RF) [15], and Ridge Regression (Ridge) [16] that corresponded to the parameters “lasso”, “enet”, “ridge”, “randomForest” respectively in the “train.model” function of the SIAMCAT package. By plotting a boxplot, we observed that there was no significant difference in modeling between the four methods. However, the results of the random forest modeling were the best (Figure S8). Therefore, we chose random forest as our machine learning algorithm.

The num.folds and num.resample parameters in the "create.data.split" function were used to adjust for different datasets, including intra-cohort (num.folds = 5,num.resample = 3) and combined cohort (num.folds = 10,num.resample = 3) modeling and validation. The model was then established using the "train_model" function. The "make.predictions" and "evaluate.predictions" functions were used for prediction. The "pROC" R package [17] (v.1.18.5) was then used to calculate the area under the receiver operating characteristic curve (AUC) score as a measure of prediction performance. By default, all features were used for model training and validation, as recommended by refs [7, 18].

In addition to training models on individual datasets, leave-one-dataset-out (LODO) analyses [19] were also performed, which involved training a classifier model on n-1 datasets and validating it on the remaining one dataset at a time when there were at least three datasets available [20]. The LODO analysis examines whether incorporating multiple cohorts for model training can improve the predictive performance of classifiers, in order to test whether the cross-validation model is biased towards a specific dataset.

In this study, we also employed the Sample-Cumulation Modeling (SCM) approach to determine the relationship between sample sizes and AUC values as a combined-cohort modeling and validation strategy [6]. For the above combined-cohort modeling(LODO and SCM), we used ten folds three times repeated cross-validation (num.folds = 10, num.resample = 3 in the create.data.split function) and only noted external (cross-cohort) validation AUCs. The SCM analysis examines whether the model performance increases with the increasing number of samples in modeling training.

Statistical analysis and bioinformatics methods

All processed data, if not specifically stated, were loaded into R (version 4.1.2, https://www.r-pro ject.org/) for analysis and visualization. Wilcoxon rank sum test was used for two-group comparison , while the Kruskal-Wallis test was used for multi-group comparisons, using the "ggpubr" R package (v.0.4.0, https://github.com/kassambara/ggpubr) in "stat_compare_means" function with default parameters. When performing multiple hypothesis tests, the corrected Wilcoxon rank test was performed by the "ggpubr" package "compare_means" function with default parameters. The correlation analysis used the spearman correlation test. All tests were two-sided with P-value < 0.05 (when two groups are compared) or a corrected P-value of < 0.05 (when multiple groups are compared) considered statistically significant. We used the diversity function in the "vegan" R package [21] (v2.6.4) to calculate the alpha diversity of each dataset using the Shannon diversity index and abundance coverage estimator (ACE), which queries species diversity and abundance. We used the Bray-Curtis index to calculate the differences between samples and estimate beta diversity. And we perform PcoA analysis and visualization using the "pcoa" function in the ape R package [22] (v.5.7-1).

1. Katz, K., O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, C. O'Sullivan. 2022. "The Sequence Read Archive: a decade more of explosive growth." *Nucleic Acids Res* 50: D387-D390. https://doi.org/10.1093/nar/gkab1053

2. Bolger, A. M., M. Lohse, B. Usadel. 2014. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30: 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

3. Langmead, B., S. L. Salzberg. 2012. "Fast gapped-read alignment with Bowtie 2." *Nat Methods* 9: 357-359. https://doi.org/10.1038/nmeth.1923

4. Blanco-Miguez, A., F. Beghini, F. Cumbo, L. J. McIver, K. N. Thompson, M. Zolfo, P. Manghi, et al. 2023. "Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4." *Nat Biotechnol* 41: 1633-1644. https://doi.org/10.1038/s41587-023-01688-w

5. Beghini, F., L. J. McIver, A. Blanco-Miguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, et al. 2021. "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3." *Elife* 10: https://doi.org/10.7554/eLife.65088

6. Li, M., J. Liu, J. Zhu, H. Wang, C. Sun, N. L. Gao, X. M. Zhao, W. H. Chen. 2023. "Performance of Gut Microbiome as an Independent Diagnostic Tool for 20 Diseases: Cross-Cohort Validation of Machine-Learning Classifiers." *Gut Microbes* 15: 2205386. https://doi.org/10.1080/19490976.2023.2205386

7. Wirbel, J. Auid-Orcid, K. Auid-Orcid Zych, M. Auid-Orcid Essex, N. Auid-Orcid Karcher, E. Auid-Orcid X. Kartal, G. Auid-Orcid Salazar, P. Auid-Orcid X. Bork, S. Auid-Orcid Sunagawa, G. Auid-Orcid Zeller. "Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox."

8. Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth. 2015. "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Res* 43: e47. https://doi.org/10.1093/nar/gkv007

9. Leek, J. T., W. E. Johnson, H. S. Parker, A. E. Jaffe, J. D. Storey. 2012. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." *Bioinformatics* 28: 882-883. https://doi.org/10.1093/bioinformatics/bts034

10. Ma, S., D. Shungin, H. Mallick, M. Schirmer, L. H. Nguyen, R. Kolde, E. Franzosa, H. Vlamakis, R. Xavier, C. Huttenhower. 2022. "Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin." *Genome Biol* 23: 208. https://doi.org/10.1186/s13059-022-02753-4

11. Ling, W., J. Lu, N. Zhao, A. Lulla, A. M. Plantinga, W. Fu, A. Zhang, et al. 2022. "Batch effects removal for microbiome data via conditional quantile regression." *Nat Commun* 13: 5418. https://doi.org/10.1038/s41467-022-33071-9

12. Cao, Y., Q. Dong, D. Wang, P. Zhang, Y. Liu, C. Niu. 2022. "microbiomeMarker: an R/Bioconductor package for microbiome marker identification and visualization." *Bioinformatics* 38: 4027-4029. https://doi.org/10.1093/bioinformatics/btac438

13. Zou, Hui, Trevor Hastie. 2005. "Regularization and Variable Selection Via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67: 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

14. Tibshirani, Robert. 2018. "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58: 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

15. Chen, H., N. Tang, Q. Ye, X. Yu, R. Yang, H. Cheng, G. Zhang, X. Zhou. 2022. "Alternation of the gut microbiota in metabolically healthy obesity: An integrated multiomics analysis." *Front Cell Infect Microbiol* 12: 1012028. https://doi.org/10.3389/fcimb.2022.1012028

16. Goldstein, M., A. F. M. Smith. 2018. "Ridge-Type Estimators for Regression Analysis." *Journal of the Royal Statistical Society: Series B (Methodological)* 36: 284-291. https://doi.org/10.1111/j.2517-6161.1974.tb01006.x

17. Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, Markus Müller. 2011. "pROC: an open-source package for R and S+ to analyze and compare ROC curves." *BMC Bioinformatics* 12: https://doi.org/10.1186/1471-2105-12-77

18. Zhu, Jiaying, Chuqing Sun, Min Li, Guoru Hu, Xing-Ming Zhao, Wei-Hua Chen. 2023. "Compared to histamine-2 receptor antagonist, proton pump inhibitor induces stronger oral-to-gut microbial transmission and gut microbiome alterations: a randomised controlled trial." *Gut* https://doi.org/10.1136/gutjnl-2023-330168

19. Wirbel, J., P. T. Pyl, E. Kartal, K. Zych, A. Kashani, A. Milanese, J. S. Fleck, et al. 2019. "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer." *Nat Med* 25: 679-689. https://doi.org/10.1038/s41591-019-0406-6

20. Markus Riester, Wei Wei, Levi Waldron, Aedin C. Culhane, Lorenzo Trippa, Esther Oliva, Sung-hoon Kim, Franziska Michor, Curtis Huttenhower, Giovanni Parmigiani, Michael J. Birrer. Risk Prediction for late-Stage Ovarian cancer by Meta-analysis of 1525 Patient Samples. https://doi.org/10.1093/jnci/dju048

21. Oksanen, Jari, F. Guillaume Blanchet, Roeland Kindt, P. Legendre, R. G. O'Hara, Gavin Simpson, Peter Solymos, Hank Stevens, Helene Wagner. 2013. "Multivariate analysis of ecological communities in R: vegan tutorial. R package version 1.7."

22. Paradis, E., K. Schliep. 2019. "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R." *Bioinformatics* 35: 526-528. https://doi.org/10.1093/bioinformatics/bty633