

Interpretable Machine Learning Models for Early Childhood Caries Prediction in Primary Dentition: A Cohort Study of Children Aged 3–6

Li Zhang

Shenzhen Children's Hospital

Lixin Wu

Shenzhen Children's Hospital

Ming Li

Kashi People's Hospital

Qin Wang

Shenzhen Children's Hospital

Xuezhen Peng

Shenzhen Children's Hospital

Lulu Chen

Shenzhen Children's Hospital

Guicong Ding

Shenzhen Children's Hospital

Guobing Wang

rogasan@163.com

Shenzhen Children's Hospital

Research Article

Keywords: dental caries, validation study, predictive modelling, interpretable machine learning, data visualization

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10772592/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Interpretable Machine Learning Models for Early Childhood Caries Prediction in Primary Dentition: A Cohort Study of Children Aged 3–6

Li Zhang 12, Lixin Wu 3 * Ming Li 2, Qin Wang 1, Xuezhen Peng 1, Lulu Chen2†, Guicong Ding 2†, Guobing Wang 4†

1. Department of Stomatology, Shenzhen Children's Hospital, Shenzhen, 518038, China
2. Department of Stomatology, Kashi People's Hospital, Kashi, Xinjiang 844000, China.
3. Department of Anesthesiology, Shenzhen Children's Hospital, Shenzhen, 518038, China
4. Shenzhen Institute of Pediatrics, Shenzhen Children's Hospital, Shenzhen, 518038, China

Lixin Wu is **co-first author**.

Lulu Chen, Guicong Ding and Guobing Wang are **Corresponding authors**.

Lulu Chen, Tel: (0086)18938691392, E-mail: 18938691392@163.com

Guicong Ding, Tel: (0086)18938691361, E-mail: dgc_63@163.com

Guo-bing Wang, Tel: (0086)18938691579, E-mail: rogasan@163.com

Author Contributions

Li Zhang, contributed to design and data interpretation, drafted and critically revised the manuscript; Ming Li, contributed to data analysis and interpretation, drafted and critically revised the manuscript; Xuezhen Peng and Qin Wang contributed to data analysis, critically revised the manuscript; Lixin Wu contributed to design, data analysis, and interpretation, critically revised the manuscript; Guo-bing Wang, contributed to data analysis and interpretation, critically revised the manuscript; Guicong Ding contributed to design, data interpretation, critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

Interpretable Machine Learning Models for Early Childhood Caries Prediction

in Primary Dentition: A Cohort Study of Children Aged 3–6

Background: Despite its status as a prevalent chronic disease affecting children's well-being, the prediction of severe early childhood caries (SECC) is not yet fully exploited. This study was aimed to construct a clinical predictive model for caries to facilitate the early identification of high-risk individuals and guide tailored clinical interventions.

Methods: A total of 93 children aged 3-6 were enrolled in this study and stratified into three cohorts: healthy controls (HC, n = 32), patients with severe early childhood caries (SECC, n = 31), and patients with both severe early childhood caries and extrinsic black stain (SECCBS, n = 30). Supragingival plaque samples were obtained from all participants for parallel 16S rRNA gene sequencing to characterize the microbial composition. R4.5.3 was used to build fifteen machine learning algorithms. Model performance was validated using discrimination and calibration metrics including accuracy, balanced accuracy, sensitivity, specificity, Negative predictive value (NPV), Positive Predictive Value (PPV), Youden's J Index, F1 score, and area under the receiver operating characteristic curve (AUC-ROC), while the clinical relevance was determined through Decision Curve Analysis (DCA). To elucidate the underlying decision-making logic, SHAP (SHapley Additive exPlanations) were utilized to assess feature importance and the clinical relevance of prediction models. UpSet plot was used to visualize machine learning models among microbial indicators.

Results: In the HC-SECC cohort, the LightGBM model demonstrated superior predictive performance; simultaneously, the Dtree exhibited robust generalizability across external validation cohorts. Regarding the SECC-SECCBS group, the MLP model maintained notable performance consistency across both internal and external validations. Specifically, the SVM-poly and MLP models yielded higher net benefits at elevated threshold probabilities, thereby underscoring their clinical reliability for identifying high-risk individuals within the HC-SECC and SECC-SECCBS cohorts, respectively. Furthermore, consensus analysis across 15 models identified *S. mutans* as the keystone biomarker with the highest predictive stability, a finding corroborated by its dominant SHAP values in boosting algorithms. Furthermore, our ensemble ML framework consistently prioritized *Actinomyces oris*, *Prevotella denticola*, and *Veillonella parvula* as core consensus

biomarkers, ranging from linear models to neural networks and reflecting the successional shift from eubiosis to SECC-associated dysbiosis. In contrast, the depletion of *Corynebacterium matruchotii* and *Leptotrichia buccalis* in the SECC cohort relative to the HC and SECCBS groups highlights their roles as potential health-associated anchors. *Rothia dentocariosa*, and *Leptotrichia A_993758 wadei* were identified as core consensus biomarkers significantly associated with SECCBS. Ultimately, these findings, visualized through high-intersectionality UpSet plots, demonstrate that the transition to SECC is driven by high-magnitude taxonomic changes that are consistently captured across diverse algorithmic frameworks.

Conclusions: The LightGBM algorithm has a good prediction effect of dental caries. The utilization of SHAP for model interpretation provided transparent insights into the decision-making process. The nomogram facilitates the clinical identification of individuals at high risk for dental caries, providing a robust framework for targeted preventive interventions.

Keywords: dental caries; validation study; predictive modelling; interpretable machine learning; data visualization

Introduction

Severe early childhood caries (SECC) is a complex, multi-factorial disease that negatively impact the life quality of young children, including functional, psychological, and social well-being, as well as their families [1]. The global prevalence of ECC ranges from 20.6% to 98.0% [2]. Data from the Fifth National Oral Health Epidemiological Survey in China further underscores this public health burden, revealing caries prevalence rates of 50.8%, 63.6%, and 70.9% in children aged 3, 4, and 5 years, respectively [3]. These figures exhibit a marked incremental trend, reflecting a significant escalation in caries susceptibility during early childhood and emphasizing the urgency of prioritizing ECC within the broader public health agenda. Meanwhile, black tooth stain (BTS) is also commonly found in primary dentition [4], exhibit high tenacity, adhering firmly to the dental pellicle. It not only compromises the self-esteem and social confidence of affected children but also imposes a long-term psychological and management burden on their families [5].

In recent years, machine learning (ML) techniques have emerged as powerful tools in

clinical prediction, excelling at capturing intricate patterns within high-dimensional datasets [6]. ML can accurately apply in multiply facets, such as predicting the treatment outcome of ECC [7], detecting dental caries [8], exploring complex associations between systematic biomarkers and periodontal disease [9], and etc. However, most current models remain opaque 'black boxes' that lack transparency and fail to account for the synergistic interactions between risk factors [10]. Furthermore, systematic frameworks for individual-level interpretability are notably absent [11], which undermines clinician confidence and stymies the integration of ML-driven insights into evidence-based dental practice. Addressing these interpretability gaps is essential for fostering trust and facilitating the clinical adoption of advanced predictive algorithms. To address this, we integrated the SHapley Additive exPlanations (SHAP) method to elucidate the decision-making logic of our models. By calculating SHAP values, we assigned a directional importance to each feature—representing its contribution toward or against a specific prediction [12, 13]. Through the implementation of SHAP bar charts and beeswarm plots, we visualized the global importance and local impact of each predictor, thereby enhancing the clinical interpretability of the treatment outcome models. Nomograms and Decision Curve Analysis (DCA) represent innovative modeling approaches that achieve high predictive accuracy by synthesizing multidimensional factors into a visual system. This technological advancement offers a streamlined, efficient, and quantifiable framework for assessing disease risk and clinical prognosis [14].

In this study, 93 children were enrolled and stratified into three cohorts: healthy controls (HC, $n = 32$), severe early childhood caries (SECC, $n = 31$), and SECC with extrinsic black stain (SECCBS, $n = 30$). Leveraging microbial sequencing data, we developed 15 ML algorithms and identified the optimal predictive model. The SHAP framework was subsequently employed to quantify feature contributions, elucidate inter-variable interactions, and analyze non-linear relationships, thereby exploring the feasibility of individualized risk interpretation. As traditional Venn diagrams are typically constrained to fewer than five sets due to increasing visual complexity [15], the matrix-based UpSet plots were generated in R demonstrated intricate interactions between the models and the oral microbiota. DCA was used to evaluate

the clinical guiding practicality of the nomogram. This research aims to provide a robust scientific foundation for the early prediction and precision management of SECC.

Materials and Methods

Study cohorts

The study design, protocol, and informed consent were approved by the Institutional Review Board of Shenzhen Children's Hospital (No. 202319202). Written informed consent was obtained from all participants involved in the study. The sample size was determined based on established methodologies from previous studies of similar design [16] and practical feasibility. Initially, approximately 700 preschoolers (aged 3–6 years) were screened during oral examinations at the Department of Stomatology, Shenzhen Children's Hospital. Following rigorous selection, a total of 93 participants were finally enrolled (Fig.1). To minimize confounding variables, participants were recruited from the same kindergarten within a single geographic region, ensuring that socioeconomic and environmental backgrounds remained highly homogeneous across all study groups. The exclusion criteria were as follows: (1) presence of systemic diseases; (2) developmental dental anomalies,

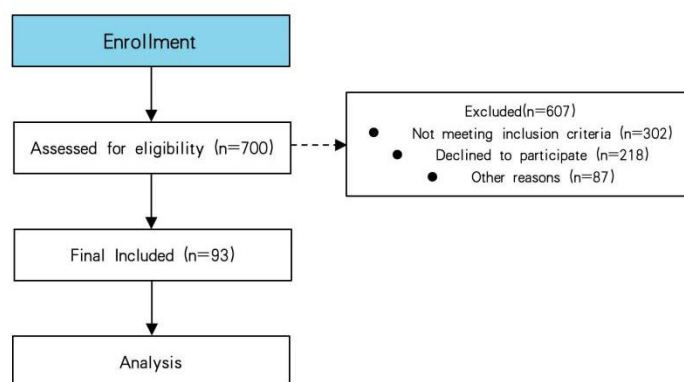


Figure 1

including enamel or dentin hypoplasia; (3) use of antibiotics or anti-inflammatory agents within the two weeks prior to sample collection; and (4) receipt of professional fluoride therapy within the preceding month.

Clinical evaluations were carried out by one trained pediatric dentist; participants were examined in the supine position using a dental unit to ensure standardized posture. The

presence of dental caries was recorded using the World Health Organization (WHO) diagnostic criteria [1] and dental caries status was recorded as sound, decayed, missing, filled surfaces. In the interest of diagnostic reliability, white spot lesions were specifically excluded. Participants were stratified into the SECC group according to age-specific dmft thresholds: ≥ 4 , ≥ 5 , and ≥ 6 for ages 3, 4, and 5, respectively. The BTS relied on the observation of dark pigmented lines or dots along the gingival margin, generally confined to the cervical third of the tooth [17]. Furthermore, tactile exploration with a dental probe was performed to distinguish these stains from arrested carious lesions.

Participants were stratified into three cohorts: healthy controls (HC), Severe Early Childhood Caries (SECC), and SECC with BTS (SECCBS). As detailed in Table 1, the study enrolled a total of 93 children: 32 in the HC group, 31 in the SECC group, and 30 in the SECCBS group. No statistically significant differences were observed between the groups regarding age or sex distribution ($P > 0.05$), confirming the demographic comparability of the cohorts.

TABLE 1. Demographic characteristics of the subjects and sample size

Variables	HC (n=32)	SECC (n=31)	SECCBS (n=30)	P-value	t/ χ^2
Age(Month) ^a	47.66±6.75	49.61±6.31	49.47±7.02	0.240 ¹ ,0.305 ² ,0.932 ³	1.19 ¹ ,1.04 ² ,0.09 ³
Gender(Girl / Boy) ^b	15(46.9)/17(53.1)	17(54.8)/14(45.2)	17(56.3)/ 13(43.3)	0.441 ¹ ,0.441 ² ,0.886 ³	0.59 ¹ ,0.59 ² ,0.02 ³

^a is represented in mean $\pm$ SD. ^b is represented in number(percentage). ¹Represents an difference between HC and SECC.

²Represents an difference between HC and SECCBS. ³Represents an difference between SECC and SECCBS.

Sampling, 16S rRNA sequencing and Raw Data Processing, Taxonomic Classification and Feature Consensus

The experimental procedures for supragingival plaque sampling, 16S rRNA gene amplification, and sequencing were performed as described in our previous study [18] with minor modifications. Briefly, plaque samples were collected from the smooth labial surfaces using a sterile dental excavator. The V3–V4 hypervariable regions of the 16S rRNA gene were amplified, and the resulting libraries were sequenced on the Illumina platform. Raw data processing, including quality filtering, denoising,

and chimera removal, was performed using the UNOISE3 algorithm within the USEARCH v11.0.667 environment. Taxonomic classification of ASVs was achieved using a Naive Bayes classifier trained on the Greengenes2 database (2024.09).

Machine Learning Modeling

Based the tidymodels R package framework, fifteen widely accepted ML algorithms across three architectural categories were constructed to ensure a comprehensive and robust evaluation of predictive performance, which comprised of: 1) Linear Models, including Logistic Regression (LR), Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Linear Support Vector Machine (Linear SVM), Linear Discriminant Analysis (LDA), and Naive Bayes Classifier (NBC); 2) Neural/Non-linear Models, consists of Multi-layer Perceptron (MLP), Support Vector Machine with Radial Basis Function kernel (SVM-RBF), Polynomial SVM (SVM-Poly), and K-Nearest Neighbors (KNN); 3) Tree-based and Ensemble Models, encompasses Decision Tree (Dtree), Random Forest (RF), Extremely Randomized Trees (Extra-Trees), and high-performance boosting algorithms such as Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM).

To improve model interpretability and mitigating overfitting, a rigorous feature selection strategy was developed to filter the irrelevant species according the following criteria: 1) relative abundance was less than 0.0002; 2) occurrence ratio in samples was less than 0.1; 3) the ratio of standard deviation div mean was less than 0.1; 4) the lowest 10% of features based on mutual information(MI) score; 5) the absolute value of the correlation coefficient of multicollinearity was greater than 0.8. Accounting for the varying requirements of different machine learning algorithms regarding data structures and feature characteristics, feature selection for linear models incorporated strategies 1-6 detailed above, whereas non-linear and Tree-based and Ensemble Models applied strategies 1-4. After feature selection, the relative abundances of species were normalized and scaled to the range of 0 to 1 for linear and non-linear models.

Nested Cross-Validation and Hyperparameter Optimization

The schematic overview of the framework of this study was illustrated in Fig. 2. All datasets were

stratified and randomly partitioned into training and validation sets at an 8:2 ratio. To ensure rigorous model validation and generalization, a nested cross-validation (Nested CV) framework was implemented: the inner loop utilized stratified 5-fold cross-validation for hyperparameter tuning, while the outer loop executed 20 independent repeats of 5-fold cross-validation (100 total resamples) to mitigate random bias and estimate performance stability. In the inner loop, a racing method with ANOVA was employed to efficiently explore the hyperparameter space, and identify the most robust predictive model by the AUC-ROC values across 15 algorithms. Inversely, a comprehensive suite of evaluation metrics was defined for the cross-validation in outer loop, including accuracy, balanced accuracy, sensitivity, negative predictive value (NPV), positive Predictive Value (PPV), Youden's J Index, F1 score, and area under the ROC Curve (ROC-AUC) to capture classification performance across multiple dimensions.

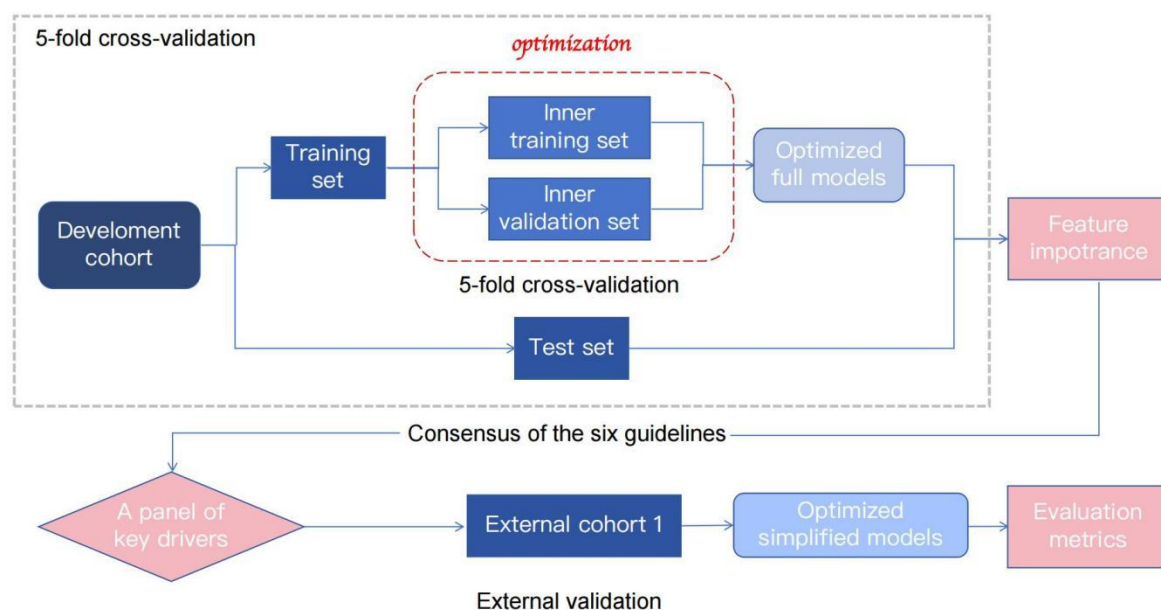


Figure 2

Feature Importance Extraction and Statistical Robustness Validation

To enhance the reliability of feature scoring, a permutation-based validation mechanism was integrated into the importance extraction process. For non-linear models such as SVM and Neural Networks, the `vi_permute` function included in VIP R Package v0.4.5, was employed to perform 100 independent random permutations ($n_{sim} = 100$) for each predictor. The relative importance was then quantified as the average decline in model ROC-AUC following feature shuffling. This

permutation-based strategy effectively minimizes potential artifacts stemming from specific data distributions. By re-fitting the top 15 optimized ML models on the entire dataset, we extract the top 15 features of every algorithm and calculated the selection frequency across various algorithm. An UpSet plot was utilized to identify consensus biomarkers validated by multiple models, ensuring that the identified variables possess both statistical robustness and cross-model predictive consistency. On the other way, an integrated biomarker-performance heatmap was constructed to visualize the correlation between the selection frequency of the top 20 core consensus ASVs across the models and their abundance patterns after centered log-ratio (CLR) transformation, with hierarchical clustering within the Complex Heatmap framework.

Clinical Utility Evaluation and Integrated Biomarker Analysis

To evaluate the potential clinical value of the machine learning models, Decision Curve Analysis (DCA) was implemented to assess their practical utility beyond traditional performance metrics. Considering the baseline SECC prevalence of 55.2% in specialized domestic dental clinics [19], this study eschewed clinically impractical full-threshold evaluations. Instead, the analysis was stratified into three distinct clinical decision intervals: a low-risk zone (10%–30%), a moderate-risk zone (30%–60%), and a high-risk zone (60%–85%). Y calculating the Net Benefit across the three ranges of threshold probabilities, we systematically compared the clinical efficacy of 15 distinct algorithms. Using the dcurves framework, the finalized models were compared against two baseline scenarios, "Treat All" and "Treat None", to determine the optimal model or strategy in different risk regions where the models provide clinical added value.

Model Explainability and Global SHAP Contribution Analysis

To demystify the "black-box" nature of the machine learning models and quantify the individual contribution of each microbial taxon to the predictive output, the SHAP framework, rooted in cooperative game theory, was implemented. Differentiated computational pathways were utilized based on the model architectures: the native TreeSHAP algorithm was applied for gradient-boosting tree models (XGBoost and LightGBM), while the Kernel SHAP method was employed for non-tree-based models (e.g., neural networks and SVMs). To ensure computational stability, SHAP values were estimated using a Monte Carlo simulation involving an explanatory set of 100 representative samples and a 30-sample background reference set. Through the construction of

global feature importance bars, beeswarm plots, and dependence plots, this study prioritized the predictive weights of microbial features and elucidated the non-linear dynamics between taxonomical abundance and diagnostic probability, as well as potential inter-feature interactions. Beyond global interpretations, force plots were utilized for local interpretation of individual clinical samples, demonstrating how specific features shifted the model's base value toward the final predicted probability.

To distill the most robust biomarkers from the ensemble of 15 models, mean absolute SHAP values were calculated across all algorithms, and the intersections of prioritized features were systematically evaluated using UpSet matrix visualization. Features identified as core predictors in at least 5 distinct models (frequency $\geq 33\%$) were designated as "consensus biomarkers." Furthermore, Spearman correlation coefficients between SHAP values and raw abundances were calculated to determine the directionality of each biomarker's influence (Enriched vs. Depleted in the target group), thereby harmonizing ML-based predictions with biological functional interpretations.

Statistical Analysis

Continuous variables with a normal distribution were expressed as mean $\pm$ standard deviation ($\bar{X} \pm S$), and inter-group comparisons were performed using the independent-samples *t*-test. For non-normally distributed continuous variables, data were presented as median and interquartile range [M (P25, P75)], and compared using the Mann-Whitney U test. Categorical variables were expressed as frequencies square test; Fisher's exact test or the Wilcoxon rank-sum test was employed for ordinal data or where appropriate. $P < 0.05$ was considered to indicate a statistically significant difference.

Results

Model Performance Selection and Evaluation

During internal validation, the LightGBM model implemented on the HC-SECC cohort demonstrated superior predictive performance, yielding a peak AUC-ROC of 0.855 (95% CI: 0.835–0.874) (Table 2). The model also achieved high precision and stability across other metrics, including an accuracy of 0.775 (95% CI: 0.753–0.797), sensitivity of 0.763 (95% CI: 0.732–0.793), specificity of 0.787 (95% CI: 0.749–0.826), and an F1-score of 0.768 (95% CI: 0.746–0.790). In

TABLE 2. Performance of machine learning predictive models

Group	Model_Name	Accuracy	Bal_Acuracy	F1 Score	Youden's J_Index	NPV	PPV	AUC_ROC	Sensitivity	Specificity
HC- SECC	LightGBM	0.775 ± 0.113	0.775 ± 0.113	0.768 ± 0.114	0.550 ± 0.226	0.786 ± 0.128	0.804 ± 0.149	0.855 ± 0.098	0.763 ± 0.157	0.787 ± 0.197
		[0.753-0.797]	[0.753-0.797]	[0.746-0.790]	[0.506-0.594]	[0.761-0.811]	[0.775-0.833]	[0.835-0.874]	[0.732-0.793]	[0.749-0.826]
	SVM_RBF	0.758 ± 0.104	0.757 ± 0.104	0.741 ± 0.122	0.515 ± 0.208	0.766 ± 0.127	0.785 ± 0.134	0.843 ± 0.104	0.729 ± 0.173	0.785 ± 0.156
		[0.737-0.778]	[0.737-0.778]	[0.718-0.765]	[0.474-0.555]	[0.741-0.791]	[0.759-0.812]	[0.822-0.863]	[0.695-0.763]	[0.755-0.816]
	SVM_Poly	0.768 ± 0.105	0.767 ± 0.105	0.747 ± 0.129	0.535 ± 0.210	0.770 ± 0.126	0.803 ± 0.136	0.842 ± 0.101	0.726 ± 0.181	0.809 ± 0.148
		[0.747-0.788]	[0.747-0.788]	[0.722-0.772]	[0.493-0.576]	[0.746-0.795]	[0.777-0.830]	[0.822-0.862]	[0.691-0.761]	[0.780-0.838]
HC- SECC	XGBoost	0.733 ± 0.106	0.733 ± 0.106	0.727 ± 0.107	0.466 ± 0.213	0.752 ± 0.129	0.752 ± 0.144	0.825 ± 0.112	0.732 ± 0.156	0.734 ± 0.186
		[0.712-0.754]	[0.712-0.754]	[0.706-0.748]	[0.425-0.508]	[0.726-0.777]	[0.724-0.781]	[0.803-0.847]	[0.702-0.763]	[0.698-0.771]
	KNN	0.708 ± 0.119	0.707 ± 0.119	0.661 ± 0.168	0.414 ± 0.238	0.703 ± 0.130	0.775 ± 0.174	0.823 ± 0.109	0.625 ± 0.218	0.789 ± 0.192
		[0.684-0.731]	[0.684-0.730]	[0.628-0.694]	[0.367-0.460]	[0.678-0.728]	[0.741-0.809]	[0.801-0.844]	[0.582-0.667]	[0.751-0.827]
HC- SECC	Extra_Tree	0.697 ± 0.119	0.698 ± 0.119	0.675 ± 0.139	0.396 ± 0.237	0.702 ± 0.135	0.727 ± 0.157	0.805 ± 0.123	0.660 ± 0.183	0.736 ± 0.181
		[0.674-0.721]	[0.675-0.721]	[0.648-0.702]	[0.349-0.442]	[0.675-0.728]	[0.697-0.758]	[0.781-0.829]	[0.624-0.695]	[0.701-0.772]
HC- SECC	MLP	0.714 ± 0.107	0.714 ± 0.107	0.696 ± 0.120	0.427 ± 0.213	0.725 ± 0.125	0.748 ± 0.150	0.787 ± 0.102	0.685 ± 0.175	0.742 ± 0.191
		[0.693-0.735]	[0.693-0.734]	[0.673-0.720]	[0.385-0.469]	[0.701-0.750]	[0.719-0.778]	[0.767-0.807]	[0.651-0.720]	[0.704-0.779]

SECC -	Random Forest	0.683 ± 0.112 [0.661-0.705]	0.684 ± 0.113 [0.662-0.706]	0.671 ± 0.131 [0.645-0.696]	0.368 ± 0.225 [0.324-0.412]	0.708 ± 0.146 [0.679-0.736]	0.698 ± 0.144 [0.670-0.727]	0.776 ± 0.130 [0.750-0.802]	0.679 ± 0.190 [0.642-0.717]	0.689 ± 0.190 [0.652-0.726]
	DTree	0.711 ± 0.117 [0.688-0.734]	0.712 ± 0.117 [0.689-0.735]	0.696 ± 0.140 [0.669-0.724]	0.424 ± 0.234 [0.378-0.470]	0.746 ± 0.161 [0.715-0.778]	0.722 ± 0.136 [0.695-0.748]	0.715 ± 0.146 [0.686-0.743]	0.710 ± 0.211 [0.669-0.752]	0.713 ± 0.172 [0.680-0.747]
	Ridge	0.593 ± 0.128 [0.568-0.618]	0.594 ± 0.128 [0.569-0.619]	0.590 ± 0.131 [0.565-0.616]	0.188 ± 0.256 [0.137-0.238]	0.605 ± 0.147 [0.576-0.633]	0.607 ± 0.156 [0.576-0.637]	0.608 ± 0.142 [0.580-0.635]	0.600 ± 0.187 [0.564-0.637]	0.587 ± 0.223 [0.544-0.631]
	SVM_Linear	0.583 ± 0.129 [0.558-0.608]	0.584 ± 0.129 [0.559-0.610]	0.585 ± 0.136 [0.558-0.611]	0.169 ± 0.257 [0.118-0.219]	0.587 ± 0.179 [0.552-0.622]	0.599 ± 0.166 [0.567-0.632]	0.579 ± 0.170 [0.546-0.613]	0.615 ± 0.190 [0.578-0.653]	0.553 ± 0.247 [0.505-0.602]
	Lasso	0.561 ± 0.130 [0.535-0.586]	0.560 ± 0.130 [0.534-0.585]	0.542 ± 0.153 [0.513-0.572]	0.119 ± 0.260 [0.068-0.170]	0.573 ± 0.155 [0.542-0.603]	0.561 ± 0.148 [0.532-0.590]	0.560 ± 0.142 [0.532-0.588]	0.551 ± 0.196 [0.513-0.590]	0.568 ± 0.205 [0.528-0.608]
	LDA	0.499 ± 0.135 [0.473-0.526]	0.499 ± 0.134 [0.473-0.525]	0.488 ± 0.148 [0.458-0.517]	-0.002 ± 0.268 [-0.054-0.051]	0.499 ± 0.149 [0.470-0.529]	0.499 ± 0.154 [0.469-0.530]	0.496 ± 0.148 [0.467-0.526]	0.499 ± 0.180 [0.463-0.534]	0.499 ± 0.207 [0.459-0.540]
	Logistic	0.482 ± 0.145 [0.453-0.510]	0.482 ± 0.145 [0.453-0.510]	0.459 ± 0.174 [0.425-0.493]	-0.037 ± 0.291 [-0.094-0.020]	0.486 ± 0.152 [0.457-0.516]	0.475 ± 0.185 [0.438-0.511]	0.484 ± 0.152 [0.454-0.513]	0.470 ± 0.207 [0.429-0.510]	0.494 ± 0.219 [0.451-0.537]
	NBC	0.451 ± 0.129 [0.425-0.476]	0.450 ± 0.129 [0.424-0.475]	0.397 ± 0.159 [0.366-0.429]	-0.101 ± 0.257 [-0.151--0.050]	0.457 ± 0.124 [0.433-0.482]	0.438 ± 0.179 [0.403-0.473]	0.437 ± 0.155 [0.407-0.467]	0.384 ± 0.178 [0.349-0.419]	0.515 ± 0.196 [0.477-0.554]
SECC BS	MLP	0.600 ± 0.123 [0.576-0.624]	0.600 ± 0.122 [0.576-0.624]	0.588 ± 0.127 [0.563-0.613]	0.200 ± 0.244 [0.153-0.248]	0.601 ± 0.129 [0.575-0.626]	0.614 ± 0.151 [0.585-0.644]	0.637 ± 0.148 [0.608-0.666]	0.588 ± 0.160 [0.557-0.620]	0.612 ± 0.206 [0.572-0.653]
	SVM_RBF	0.627 ± 0.145 [0.599-0.656]	0.627 ± 0.145 [0.599-0.656]	0.593 ± 0.175 [0.559-0.627]	0.255 ± 0.290 [0.198-0.311]	0.631 ± 0.146 [0.602-0.659]	0.652 ± 0.195 [0.614-0.690]	0.619 ± 0.202 [0.580-0.659]	0.577 ± 0.207 [0.536-0.617]	0.678 ± 0.204 [0.638-0.718]

Extra_Tree	0.567 ± 0.130	0.566 ± 0.130	0.514 ± 0.162	0.132 ± 0.260	0.569 ± 0.137	0.589 ± 0.190	0.592 ± 0.158	0.492 ± 0.202	0.641 ± 0.207
	[0.541-0.592]	[0.541-0.592]	[0.482-0.546]	[0.082-0.183]	[0.542-0.596]	[0.551-0.626]	[0.561-0.623]	[0.452-0.531]	[0.600-0.681]
LightGBM	0.584 ± 0.125	0.584 ± 0.125	0.565 ± 0.148	0.167 ± 0.251	0.595 ± 0.149	0.582 ± 0.139	0.586 ± 0.155	0.572 ± 0.193	0.595 ± 0.186
	[0.559-0.608]	[0.559-0.608]	[0.536-0.594]	[0.118-0.216]	[0.566-0.624]	[0.555-0.609]	[0.556-0.616]	[0.534-0.609]	[0.559-0.632]
Random Forest	0.543 ± 0.131	0.542 ± 0.131	0.504 ± 0.155	0.085 ± 0.263	0.547 ± 0.142	0.554 ± 0.183	0.583 ± 0.155	0.490 ± 0.191	0.595 ± 0.210
	[0.517-0.569]	[0.517-0.568]	[0.474-0.534]	[0.033-0.136]	[0.519-0.575]	[0.518-0.590]	[0.553-0.614]	[0.453-0.527]	[0.553-0.636]
KNN	0.567 ± 0.129	0.564 ± 0.129	0.487 ± 0.190	0.129 ± 0.258	0.565 ± 0.118	0.588 ± 0.238	0.583 ± 0.150	0.447 ± 0.214	0.682 ± 0.205
	[0.541-0.592]	[0.539-0.590]	[0.450-0.525]	[0.078-0.179]	[0.542-0.588]	[0.541-0.635]	[0.553-0.612]	[0.405-0.489]	[0.642-0.722]
XGBoost	0.568 ± 0.142	0.568 ± 0.142	0.540 ± 0.170	0.135 ± 0.283	0.577 ± 0.153	0.574 ± 0.176	0.579 ± 0.155	0.542 ± 0.208	0.594 ± 0.208
	[0.540-0.595]	[0.540-0.595]	[0.507-0.573]	[0.080-0.191]	[0.546-0.607]	[0.540-0.609]	[0.549-0.609]	[0.501-0.583]	[0.553-0.634]
SVM_Poly	0.605 ± 0.135	0.604 ± 0.135	0.532 ± 0.182	0.208 ± 0.270	0.597 ± 0.123	0.640 ± 0.215	0.577 ± 0.196	0.482 ± 0.201	0.727 ± 0.177
	[0.579-0.632]	[0.578-0.631]	[0.496-0.568]	[0.155-0.261]	[0.573-0.621]	[0.598-0.683]	[0.539-0.615]	[0.442-0.521]	[0.692-0.761]
NBC	0.569 ± 0.126	0.569 ± 0.126	0.540 ± 0.167	0.137 ± 0.253	0.583 ± 0.132	0.563 ± 0.165	0.574 ± 0.135	0.542 ± 0.204	0.595 ± 0.178
	[0.544-0.594]	[0.544-0.593]	[0.507-0.572]	[0.088-0.187]	[0.557-0.609]	[0.530-0.595]	[0.547-0.601]	[0.502-0.582]	[0.561-0.630]
SVM_Linear	0.485 ± 0.057	0.487 ± 0.054	0.479 ± 0.120	-0.026 ± 0.109	0.484 ± 0.111	0.411 ± 0.088	0.543 ± 0.159	0.120 ± 0.259	0.854 ± 0.307
	[0.473-0.496]	[0.476-0.498]	[0.456-0.503]	[-0.047--0.004]	[0.463-0.506]	[0.394-0.428]	[0.512-0.575]	[0.069-0.171]	[0.794-0.915]
DTree	0.527 ± 0.128	0.527 ± 0.128	0.497 ± 0.156	0.054 ± 0.255	0.533 ± 0.152	0.529 ± 0.161	0.528 ± 0.130	0.500 ± 0.200	0.554 ± 0.216
	[0.502-0.552]	[0.502-0.552]	[0.467-0.528]	[0.004-0.104]	[0.503-0.563]	[0.498-0.561]	[0.502-0.553]	[0.461-0.539]	[0.511-0.596]
LDA	0.523 ± 0.148	0.523 ± 0.148	0.503 ± 0.182	0.045 ± 0.296 [-	0.545 ± 0.174	0.512 ± 0.170	0.521 ± 0.162	0.522 ± 0.232	0.524 ± 0.187
	[0.494-0.552]	[0.494-0.552]	[0.467-0.538]	0.012-0.103]	[0.511-0.579]	[0.478-0.545]	[0.489-0.552]	[0.476-0.567]	[0.487-0.560]

Logistic	0.513 ± 0.147 [0.484-0.542]	0.513 ± 0.148 [0.484-0.542]	0.495 ± 0.169 [0.462-0.528]	0.026 ± 0.295 [- 0.032-0.084]	0.520 ± 0.175 [0.486-0.555]	0.512 ± 0.181 [0.476-0.547]	0.518 ± 0.152 [0.488-0.548]	0.505 ± 0.203 [0.465-0.545]	0.521 ± 0.207 [0.480-0.562]
Lasso	0.505 ± 0.021 [0.501-0.510]	0.498 ± 0.014 [0.495-0.501]	0.325 ± 0.146 [0.297-0.354]	-0.004 ± 0.027 [-0.009-0.001]	0.506 ± 0.020 [0.502-0.510]	0.354 ± 0.029 [0.348-0.360]	0.496 ± 0.029 [0.490-0.502]	0.007 ± 0.053 [-0.004-0.017]	0.990 ± 0.079 [0.974-1.005]
Ridge	0.516 ± 0.142 [0.489-0.544]	0.516 ± 0.142 [0.488-0.544]	0.494 ± 0.167 [0.462-0.527]	0.032 ± 0.284 [- 0.024-0.087]	0.517 ± 0.165 [0.485-0.550]	0.514 ± 0.169 [0.481-0.547]	0.494 ± 0.168 [0.461-0.527]	0.502 ± 0.203 [0.462-0.541]	0.530 ± 0.216 [0.488-0.572]

Results were presented as mean ± standard [95% confidence interval].

TABLE 3. Performance of machine learning predictive models on external dataset

Group	Model_Name	Accuracy	Bal_Accuracy	F1 Score	Youden's J_index	NPV	PPV	AUC_ROC	Sensitivity	Specificity
HC_SECC	DTree	0.750	0.753	0.762	0.505	0.700	0.800	0.742	0.727	0.778
	Extra_Tree	0.650	0.641	0.696	0.283	0.625	0.667	0.727	0.727	0.556
	LightGBM	0.650	0.672	0.588	0.343	0.571	0.833	0.702	0.455	0.889
	Random Forest	0.650	0.641	0.696	0.283	0.625	0.667	0.646	0.727	0.556
	SVM_Poly	0.650	0.631	0.720	0.263	0.667	0.643	0.626	0.818	0.444
	SVM_Linear	0.450	0.439	0.522	-0.121	0.375	0.500	0.606	0.545	0.333
	KNN	0.600	0.566	0.714	0.131	0.667	0.588	0.586	0.909	0.222
	SVM_RBF	0.600	0.576	0.692	0.152	0.600	0.600	0.586	0.818	0.333
	XGBoost	0.500	0.505	0.500	0.010	0.455	0.556	0.586	0.455	0.556
	Ridge	0.450	0.439	0.522	-0.121	0.375	0.500	0.556	0.545	0.333
	MLP	0.600	0.606	0.600	0.212	0.545	0.667	0.556	0.545	0.667
	Lasso	0.500	0.495	0.545	-0.010	0.444	0.545	0.495	0.545	0.444
	NBC	0.350	0.328	0.480	-0.343	0.167	0.429	0.394	0.545	0.111
	LDA	0.400	0.374	0.538	-0.253	0.200	0.467	0.384	0.636	0.111
	Logistic	0.450	0.460	0.421	-0.081	0.417	0.500	0.374	0.364	0.556

SECC_SECCBS	MLP	0.556	0.584	0.556	0.169	0.714	0.455	0.617	0.714	0.455
	DTree	0.611	0.552	0.364	0.104	0.643	0.500	0.571	0.286	0.818
	KNN	0.722	0.695	0.615	0.390	0.750	0.667	0.545	0.571	0.818
	NBC	0.556	0.532	0.429	0.065	0.636	0.429	0.519	0.429	0.636
	SVM_Poly	0.500	0.461	0.308	-0.078	0.583	0.333	0.519	0.286	0.636
	Extra_Tree	0.500	0.487	0.400	-0.026	0.600	0.375	0.506	0.429	0.545
	Lasso	0.611	0.500	NUM!!	0.000	0.611	#NUM	0.500	0.000	1.000
	SVM_RBF	0.389	0.370	0.267	-0.260	0.500	0.250	0.494	0.286	0.455
	LightGBM	0.389	0.448	0.476	-0.104	0.500	0.357	0.481	0.714	0.182
	SVM_Linear	0.389	0.500	0.560	0.000	#NUM!	0.389	0.455	1.000	0.000
	LDA	0.500	0.487	0.400	-0.026	0.600	0.375	0.435	0.429	0.545
	Logistic	0.444	0.390	0.167	-0.221	0.538	0.200	0.390	0.143	0.636
	Random Forest	0.278	0.279	0.235	-0.442	0.375	0.200	0.364	0.286	0.273
	XGBoost	0.278	0.305	0.316	-0.390	0.333	0.250	0.338	0.429	0.182
	Ridge	0.556	0.481	0.200	-0.039	0.600	0.333	0.260	0.143	0.818

the external validation phase, the Dtree model emerged as the top performer with an ROC of 0.7424 (Fig. 3A), supported by an accuracy of 0.7550 and an F1-score of 0.7619 (Table 3).

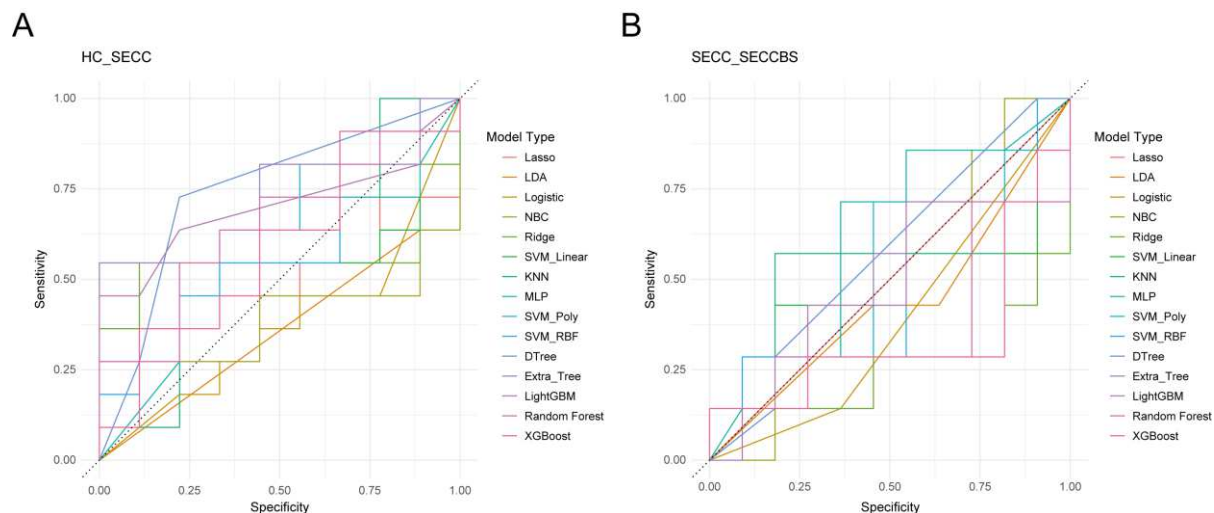


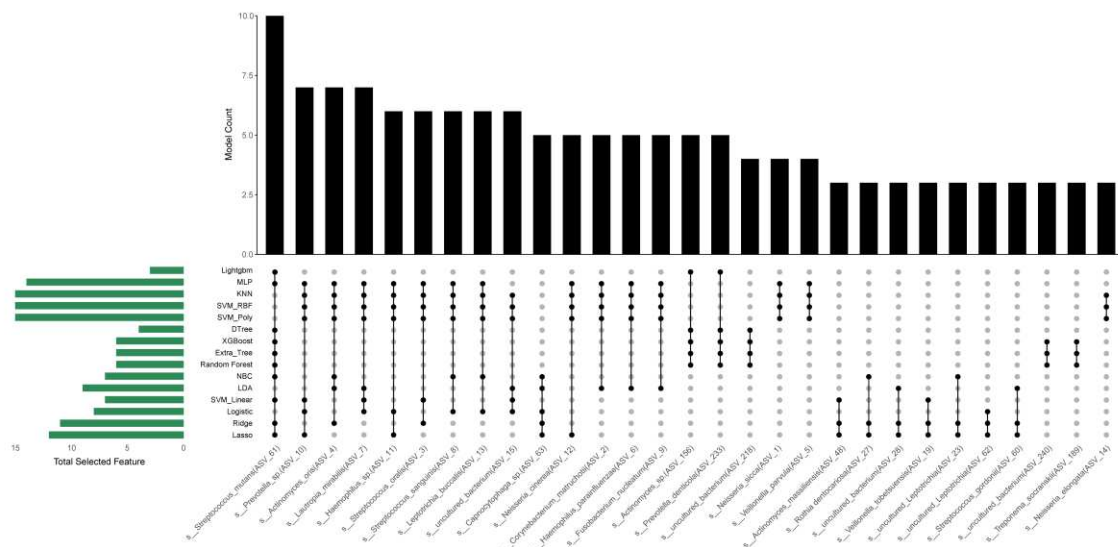
Figure 3

Furthermore, the MLP model exhibited notable performance consistency across both internal and external validations within the SECC-SECCBS group. Specifically, it recorded an internal validation ROC of 0.637 (95% CI: 0.608–0.666) (Table 2) and successfully maintained an external validation ROC of 0.6169 (Fig. 3B). Notably, the MLP model achieved a high sensitivity of 0.7143 in the external cohort (Table 3), suggesting its robust capability in identifying positive cases despite the complex nature of the SECC-SECCBS dataset. This retention of predictive power suggests that the MLP algorithm effectively captures complex non-linear features within this specific subgroup and possesses significant generalization potential for reliable feature identification in independent cohorts. Additionally, the accuracy for the HC-SECC and SECC-SECCBS groups during internal validation were 0.775 and 0.600, respectively, both of which surpassed the results obtained from the external validation cohorts (0.750 and 0.556, respectively).

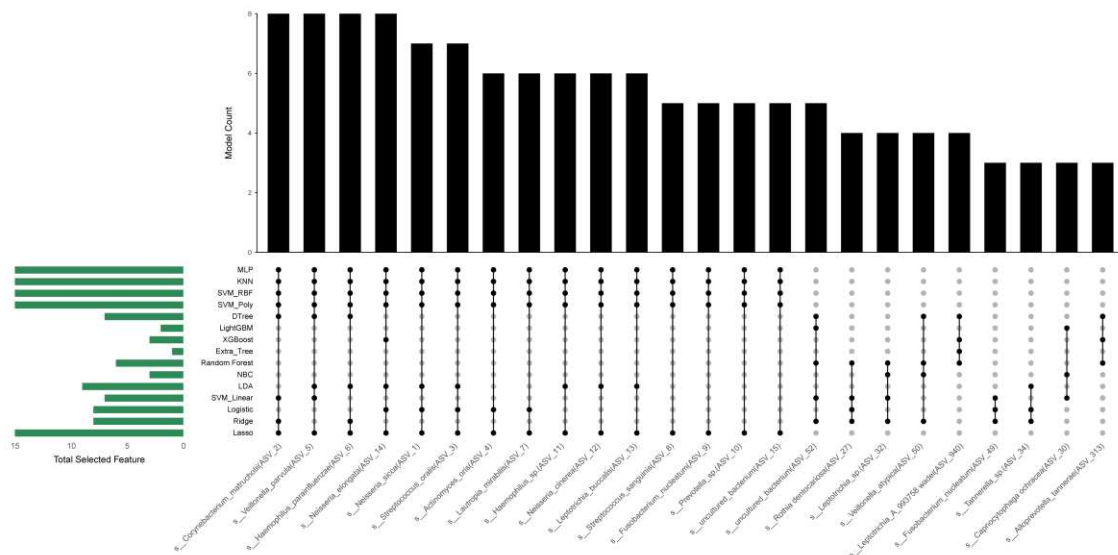
Consensus-Based Feature Selection via Machine Learning Ensembles

Among the 29 identified core features, *S.mutans* (ASV_61) emerged as the most consistent biomarker, identified by 10 out of 15 models as a top-tier predictor in the HC-SECC group. This was followed by a cluster of secondary biomarkers, including *Prevotella sp.*, *Actinomyces oris* (*A.oris*), and *Lautropia mirabilis*, which collectively established a high-confidence diagnostic

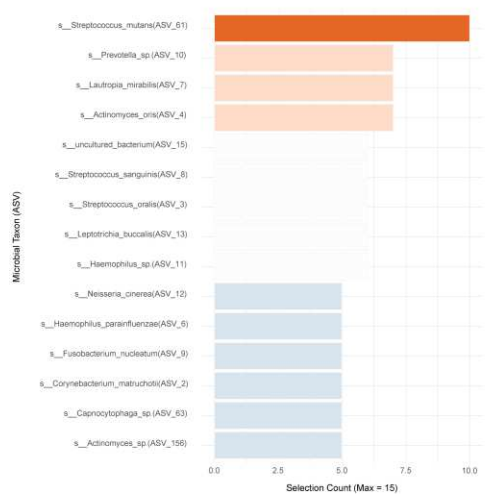
A



B



C



D

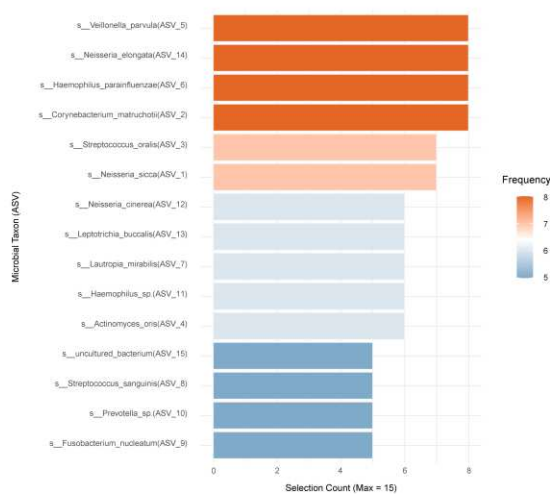


Figure 4

signature (Fig. 4A, Fig. 4C & Table 4). In the SECC-SECCBS comparison, four taxa achieved the maximum consensus frequency of 8 out of 15 models (Fig. 4B, Fig. 4D & Table 4): *Corynebacterium matruchotii* (*C.matruchotii*, ASV_2), *Veillonella parvula* (*V.parvula*, ASV_5), *Haemophilus parainfluenzae* (ASV_6), and *Neisseria elongata* (ASV_14). These species were predominantly selected by neural network architectures (MLP, KNN, SVM RBF, and SVM Poly) and specific linear models (Lasso, SVM Lin), indicating their roles as stable predictors of the BTS phenotype. A secondary consensus tier (7/15 models) identified *Neisseria sicca* (ASV_1) and *Streptococcus oralis* (ASV_3) as significant discriminatory features. Furthermore, a tertiary cluster of species including *A.oris*, *Lautropia mirabilis* (ASV_7), *Haemophilus sp.* (ASV_11), *Neisseria cinerea* (ASV_12), and *Leptotrichia buccalis* (ASV_13) was identified by 6 out of 15 models (Fig. 4D) .

Notably, 12 common microbial features were shared among the top 15 features, which were identified via model selection frequency in both aforementioned pairwise predictive groups (HC vs. SECC and SECC vs. SECCBS). This finding underscores the likely close association of these shared features with the pathogenesis of SECC, as well as their significant value for clinical diagnosis and prevention.

TABLE 4 Distribution of Consensus Microbiome Biomarkers across Ensemble Learning Models in the Two distinct comparison cohorts

Goup	Feature ID	Total Model Count	Linear Models		Neural/Non-Linear (SVM Poly, SVM RBF, KNN, MLP)	Tree-Based Models	
			(Lasso, Ridge, Log, SVM, LDA, NBC)			(RF, ExtraTree, XGB, DTree, LightGBM)	
HC-SECC	s__Streptococcus_mutans(ASV_61)	10	1	1, 0, 1, 0, 1	0, 0, 0, 1	1, 1, 1, 1, 1	
	s__Prevotella_sp.(ASV_10)	7	1	0, 1, 1, 0, 0	1, 1, 1, 1	0, 0, 0, 0, 0	
	s__Actinomyces_oris(ASV_4)	7	0	1, 0, 0, 1, 1	1, 1, 1, 1	0, 0, 0, 0, 0	
	s__Lautropia_mirabilis(ASV_7)	7	0	0, 1, 1, 1, 0	1, 1, 1, 1	0, 0, 0, 0, 0	
	s__Haemophilus_sp.(ASV_11)	6	1	0, 1, 0, 0, 0	1, 1, 1, 1	0, 0, 0, 0, 0	
	s__Streptococcus_oralis(ASV_3)	6	0	1, 0, 1, 0, 0	1, 1, 1, 1	0, 0, 0, 0, 0	

SECC- SECCBS	s__Streptococcus_sanguinis(ASV_8)	6	0, 0, 1, 0, 0, 1	1, 1, 1, 1	0, 0, 0, 0, 0
	s__Leptotrichia_buccalis(ASV_13)	6	0, 0, 1, 0, 0, 1	1, 1, 1, 1	0, 0, 0, 0, 0
	s__uncultured_bacterium(ASV_15)	6	0, 0, 1, 1, 1, 0	1, 1, 1, 0	0, 0, 0, 0, 0
	s__Capnocytophaga_sp.(ASV_63)	5	1, 1, 1, 0, 1, 1	0, 0, 0, 0	0, 0, 0, 0, 0
	s__Neisseria_cinerea(ASV_12)	5	1, 0, 0, 0, 0, 0	1, 1, 1, 1	0, 0, 0, 0, 0
	s__Corynebacterium_matruchotii(ASV_2)	5	0, 0, 0, 0, 1, 0	1, 1, 1, 1	0, 0, 0, 0, 0
	s__Haemophilus_parainfluenzae(ASV_6)	5	0, 0, 0, 0, 1, 0	1, 1, 1, 1	0, 0, 0, 0, 0
	s__Fusobacterium_nucleatum(ASV_9)	5	0, 0, 0, 0, 1, 0	1, 1, 1, 1	0, 0, 0, 0, 0
	s__Actinomyces_sp.(ASV_156)	5	0, 0, 0, 0, 0, 0	0, 0, 0, 0	1, 1, 1, 1, 1
	s__Prevotella_denticola(ASV_233)	5	0, 0, 0, 0, 0, 0	0, 0, 0, 0	1, 1, 1, 1, 1
	s__Corynebacterium_matruchotii(ASV_2)	8	1, 1, 0, 1, 0, 0	1, 1, 1, 1	0, 0, 0, 0, 1
	s__Veillonella_parvula(ASV_5)	8	1, 0, 0, 1, 1, 0	1, 1, 1, 1	0, 0, 0, 0, 1
	s__Haemophilus_parainfluenzae(ASV_6)	8	1, 1, 0, 0, 1, 0	1, 1, 1, 1	0, 0, 0, 0, 1
	s__Neisseria_elongata(ASV_14)	8	1, 0, 1, 0, 1, 0	1, 1, 1, 1	0, 0, 1, 0, 0
	s__Neisseria_sicca(ASV_1)	7	1, 0, 1, 0, 1, 0	1, 1, 1, 1	0, 0, 0, 0, 0
	s__Streptococcus_oralis(ASV_3)	7	1, 0, 1, 0, 1, 0	1, 1, 1, 1	0, 0, 0, 0, 0
	s__Actinomyces_oris(ASV_4)	6	1, 0, 1, 0, 0, 0	1, 1, 1, 1	0, 0, 0, 0, 0
	s__Lautropia_mirabilis(ASV_7)	6	1, 0, 1, 0, 0, 0	1, 1, 1, 1	0, 0, 0, 0, 0

Microbial Abundance Distribution and Consensus Validation

To identify a core signature of 20 microbial biomarkers, an integrated biomarker-performance analysis utilizing a consensus of 15 independent models were implemented with the current oral microbiome data. As shown in Fig. 5, the mean levels of *C. matruchotii* and *Leptotrichia buccalis* (ASV_13) were found to be higher in the HC and SECCBS groups compared to the SECC group.

Conversely, the enrichment of *V.parvula* was observed within the SECC cohort, showing lower relative expression in both the HC and SECCBS groups.

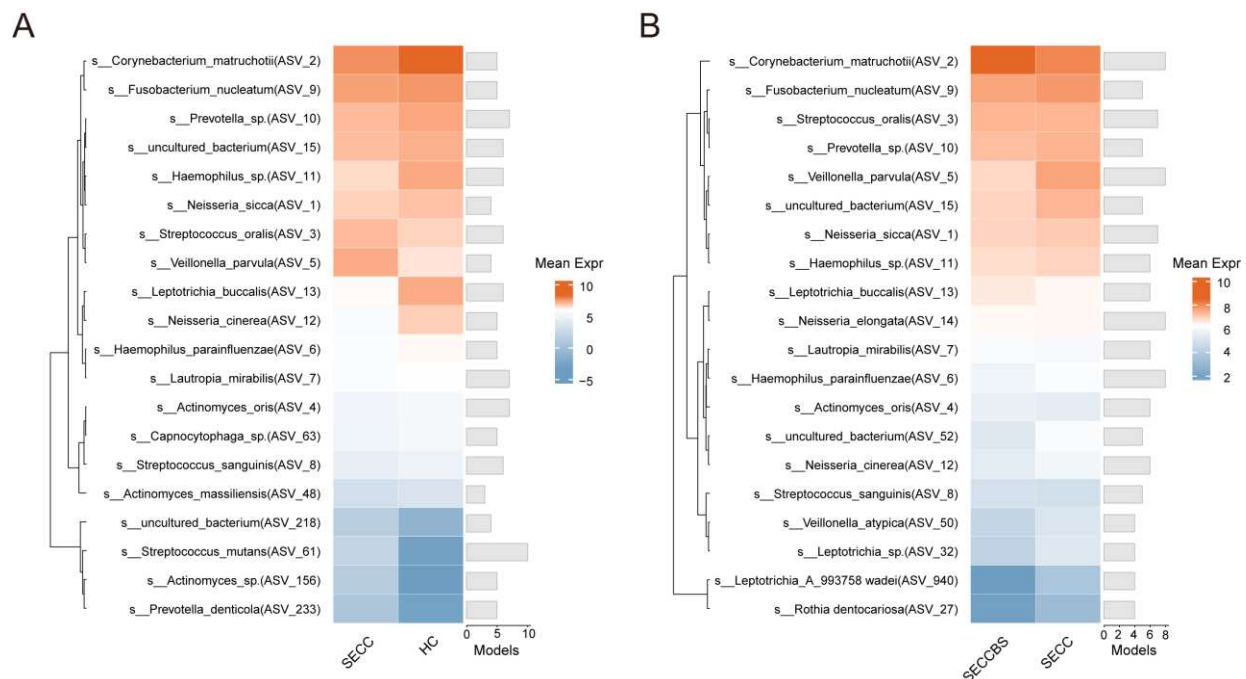


Figure 5

Clinical Utility

Our DCA framework was tailored based on the prior studies [20-22]. As depicted in Fig. 6 and Table 5, none of the ML models demonstrated a significant net benefit advantage over the "treat-all" strategy within the low-risk zone. Conversely, within the moderate-risk zone, the LightGBM (tree-based) and MLP models exhibited substantial net benefits over the 'treat all' and 'treat none' strategies (0.1073 vs. 0.0521, 0.1045 vs. 0.0521). In the high-risk zone, where the "treat-all" strategy yielded zero net benefit, the majority of ML models provided enhanced clinical utility. Notably, the SVM-RBF model achieved the highest net benefit among all evaluated algorithms (0.0491 vs. 0.000), suggesting its higher reliability for identifying individuals at high risk of SECC in clinical setting.

DCA Curve: 15-Model Comparison

Calibrated to 55.2% Prevalence for Active Treatment Scenarios

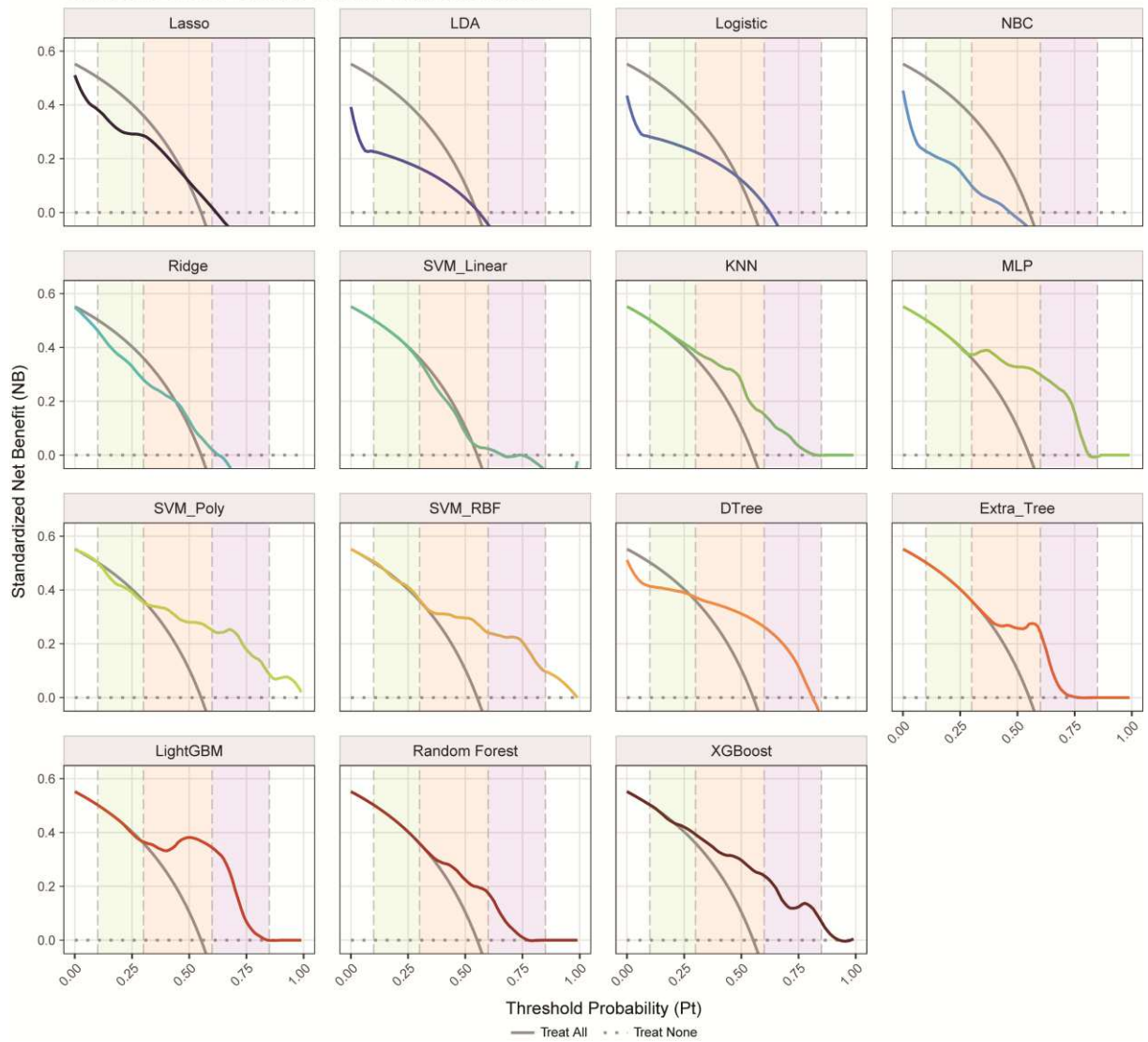


Figure 6

TABLE 5. Comparison of areas under Net_Benefit curve across 15 Models

Model Name	Areas under Net Benefit Curve			
	Low Risk Region (10% - 30%)	Moderate Risk Region (30% - 60%)	High Risk Region (60% - 85%)	Total Risk Region (10% - 85%)
Lasso	0.0634	0.0484	0.0001	0.1119
LDA	0.0398	0.0248	0.0000	0.0646
Logistic	0.0512	0.0436	0.0004	0.0951
NBC	0.0357	0.0083	0.0000	0.0439
Ridge	0.0737	0.0503	0.0002	0.1241
SVM_Linear	0.0870	0.0481	0.0017	0.1368
KNN	0.0888	0.0869	0.0146	0.1903
MLP	0.0874	0.1045	0.0411	0.2331
SVM_Poly	0.0841	0.0913	0.0490	0.2244
SVM_RBF	0.0879	0.0901	0.0491	0.2271
Dtree	0.0797	0.0974	0.0337	0.2107
Extra_Tree	0.0874	0.0848	0.0117	0.1840
LightGBM	0.0871	0.1073	0.0376	0.2320
Random Forest	0.0874	0.0780	0.0121	0.1775
XGBoost	0.0890	0.0937	0.0367	0.2194
Treat All	0.0874	0.0521	0.0000	0.1395

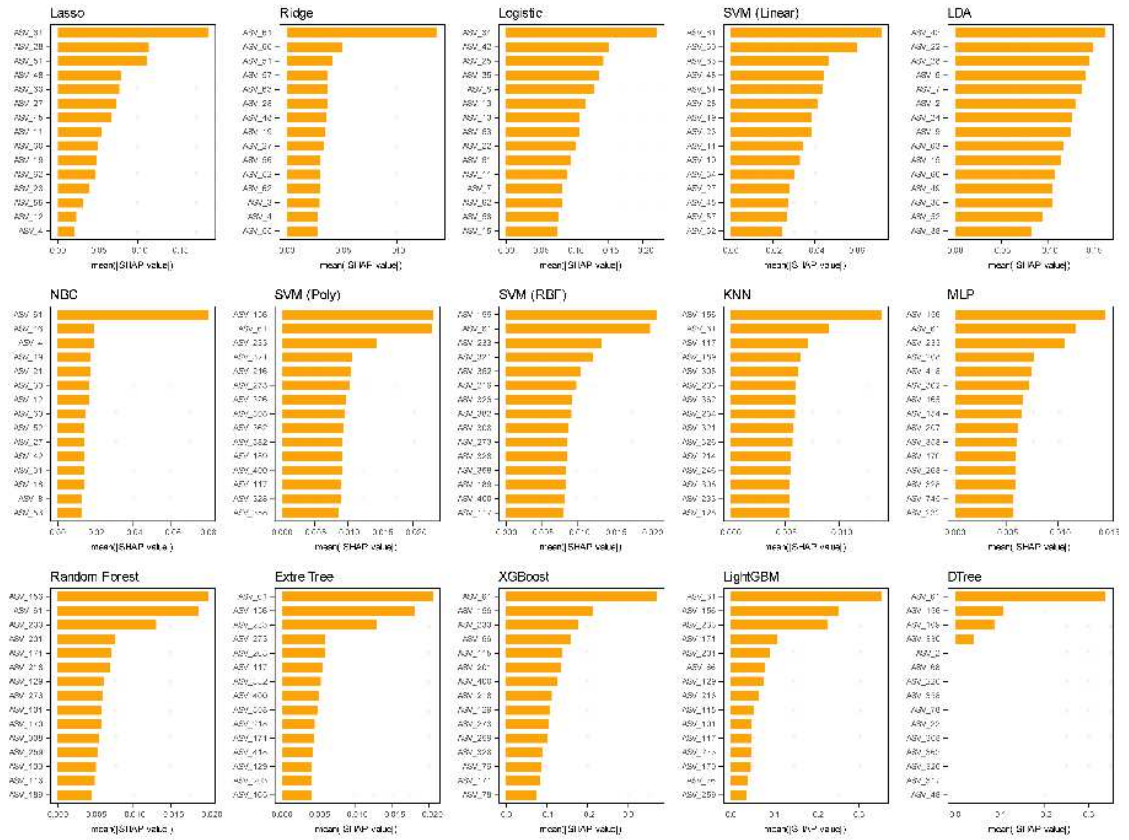
Feature Importance and Biomarker Identification via SHAP Analysis

To evaluate the contribution of microbial features to the predictive performance of our models, SHAP was employed to quantify the global importance of the top 15 ASVs across 15 ML algorithms. In the HC vs. SECC group (Fig. 7A), *S. mutans* and *Actinomyces_sp.* (ASV_156) were identified as the most robust discriminative features. Their relative importance was model-dependent: *S.*

mutans yielded the highest mean absolute SHAP values in seven models (Fig. 8A, 8C), including XGBoost and LightGBM, while *Actinomyces_sp.* (ASV_156) emerged as the lead feature in five models, such as RF and MLP (Fig. 8A, 8D).

Through SHAP analysis across 15 machine learning models in the SECC vs. SECCBS group, we identified several key microbial biomarkers for SECCBS prediction (Fig. 7B). Notably, *uncultured_bacterium* (ASV_224) emerged as the most influential feature in three distinct models (Fig. 8B, 8E), demonstrating its robust predictive power. Similarly, *Leptotrichia_A_993758 wadei* (ASV_940) was identified as the top-ranking feature in two models (Fig. 8B, 8F). The consistent high importance of these ASVs across multiple algorithmic frameworks underlines their potential as stable biological indicators for SECCBS.

A



B

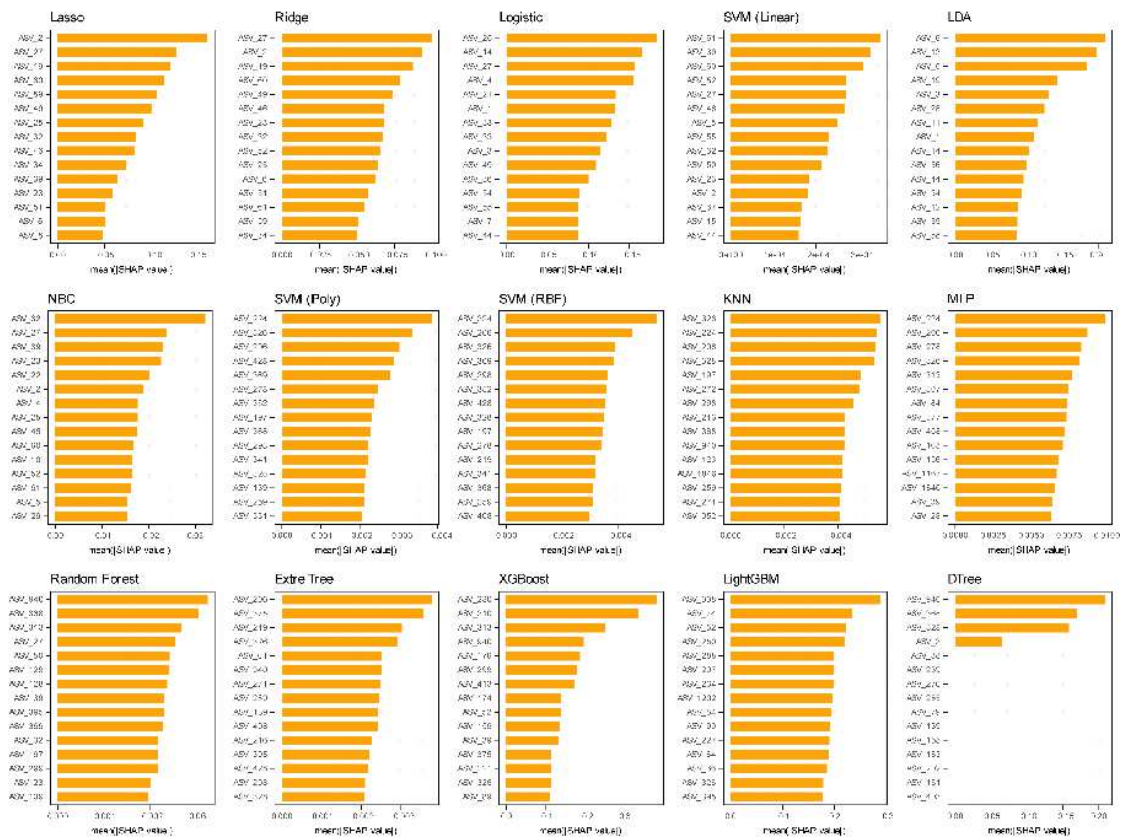


Figure 7

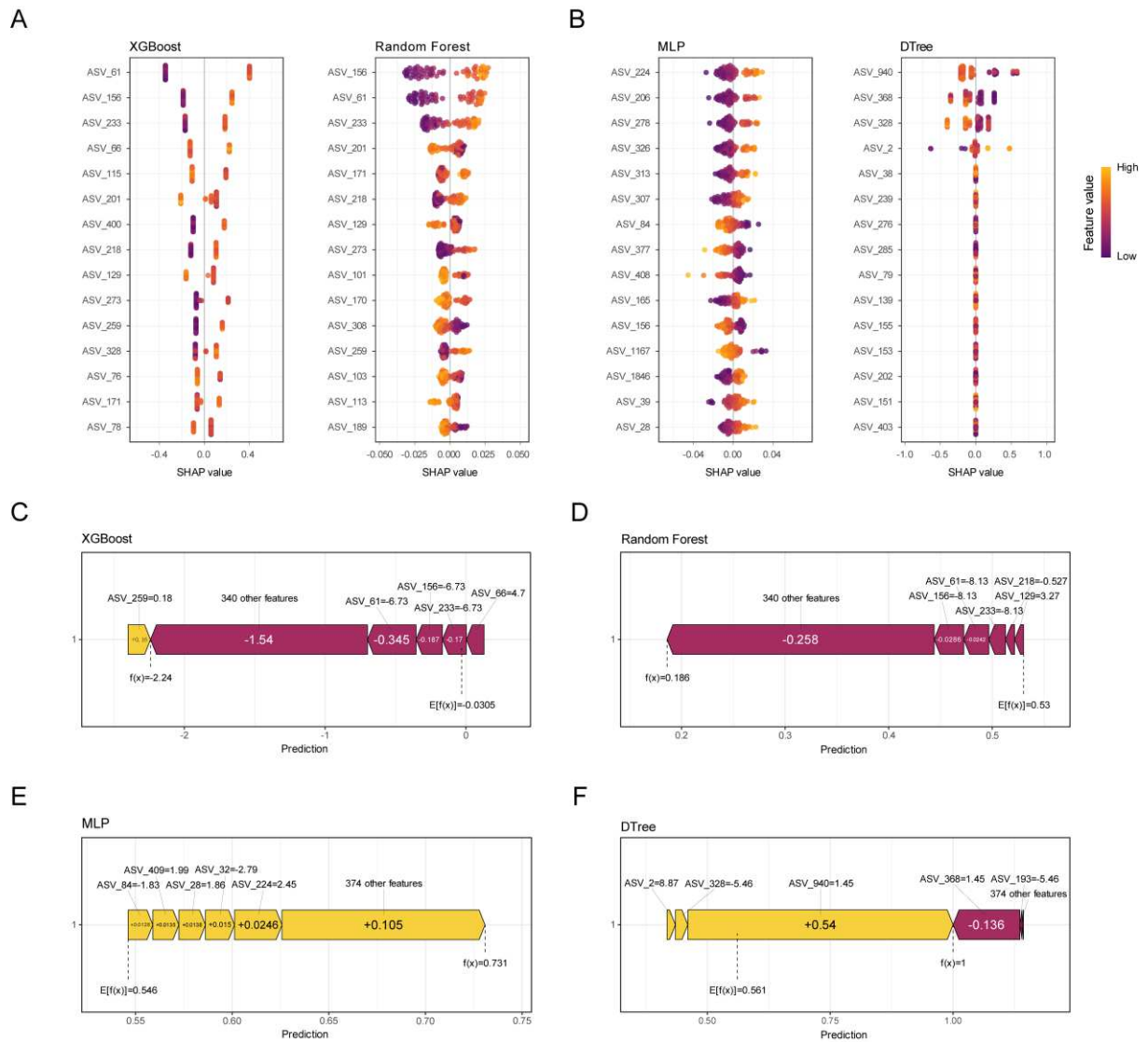


Figure 8

Differential Microbial Signatures and Key Biomarkers Across SECC Phenotypes

Fig.9A and Fig.9C showed that in the HC-SECC group, the consensus SHAP analysis of the HC-SECC cohort reveals a robust and highly consistent microbial shift, characterized by the dramatic enrichment of "classic" acidogenic pathogens. *S.mutans* demonstrated the highest predictive stability across the 15 ML models, reaffirming its status as the keystone pathogen in the development of SECC. This dominant signal is complemented by a secondary cluster of biomarkers, including *Actinomyces* sp. and *Prevotella denticola* (*P. denticola*), which underscores a synergistic successional pattern where early colonizers and anaerobic acid-producers co-construct a high-confidence diagnostic signature. The high intersectionality observed in the UpSet plot further suggests that the transition from eubiosis to SECC is driven by high-magnitude taxonomic changes

that are consistently captured across diverse algorithmic architectures (Fig.9B & Fig.9D), ranging from linear models to neural networks.

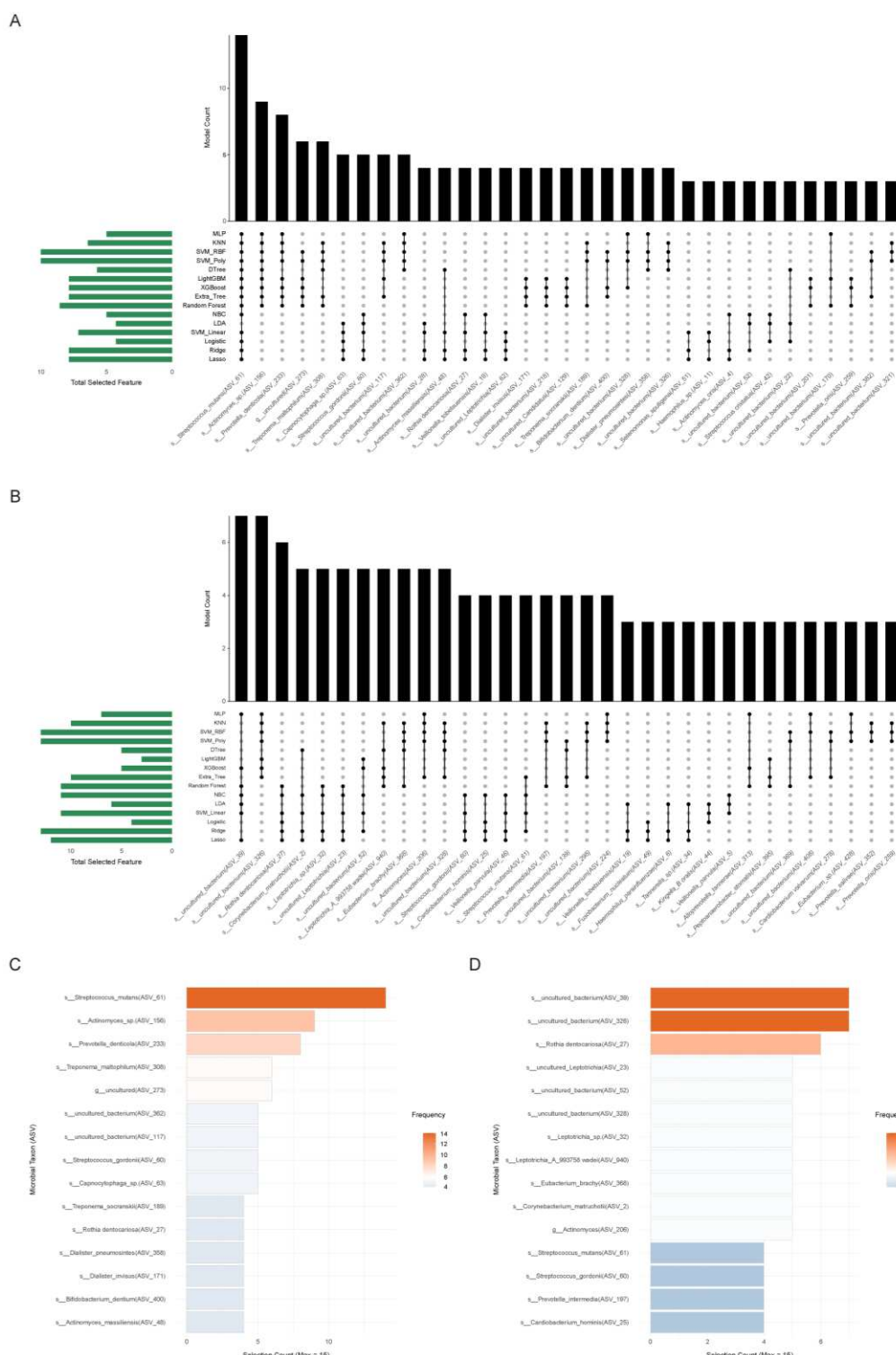


Figure 9

Furthermore, the differentiation between SECC and SECCBS highlights more subtle when compared to the HC vs. SECC group, niche-specific ecological modifications within the

cariogenic biofilm. The models prioritized a distinct set of features characterized by uncultured phylotypes and mineralization-associated taxa, notably *Rothia dentocariosa* (*R. dentocarios*) (Fig.9D), which is frequently implicated in nitrate reduction and the maintenance of a more alkaline or mineralized biofilm matrix. The prominence of these atypical biomarkers, coupled with a lower model consensus in the UpSet plot, suggests that the BTS phenotype does not represent a fundamentally different disease state, but rather a specialized functional sub-niche (Fig.9B). This microenvironment is likely shaped by the "rare biosphere", and the low-abundance taxa may possess unique metabolic pathways for iron sequestration or metallic-sulfide precipitation, providing a granular view of the microbial gradients within the SECC landscape.

Discussion

Early childhood caries (ECC) is a widespread public health issue with far-reaching implications for a child's growth, maturation, and life quality [23]. Effective caries management in early childhood relies heavily on proactive prevention. This requires the integration of validated diagnostic and prognostic tools to achieve precise, individualized risk profiling [24]. Despite its importance, no such efficacious instruments are available during the optimal preventive window of early childhood. Our study is to bridge this clinical gap by developing and validating high-performance predictive tools for dental caries in this vulnerable population.

Unlike traditional statistical methods, ML excels in dimensionality reduction, feature extraction, and pattern recognition within complex, unstructured datasets. By adapting to dynamic environments and achieving continuous self-optimization, ML can uncover latent heuristics beyond human-defined rules to enhance predictive precision [25]. However, the performance discrepancy observed between internal and external validation phases typically hinges on the inherent inter-dataset heterogeneity. Specifically, integration data from disparate sources often introduce substantial batch effects, which may artificially distinguishing features based on physical attributes or data types. Such spurious correlations can overshadow underlying biological mechanisms, thereby constraining the generalizability and robustness of the model when applied to external cohorts. It is critical to recognize that while our reported accuracy stems from internal data, this

reliance on intra-population sampling poses a risk of bias, as underscored by recent literature [26-28]. In this study, the model yielded higher internal validation accuracies (0.775 for HC-SECC and 0.600 for SECC-SECCBS) compared to external cohort. This trend aligns with findings by *Fu et al.* [29], and *Frenkel et al.* [26], where AUC and accuracy similarly attenuated during external testing. These consistent findings highlight the inherent performance decay upon transitioning to external datasets, reinforcing the imperative for independent external validation in assessing true clinical generalizability.

S. mutans is a well-established pathogen with potent cariogenic effects [30]. This was further substantiated in our study through ML ensemble modeling and SHAP analysis, which reinforced its classic role as a primary driver of acidification in SECC. Our findings are corroborated by a recent systematic review [31] that highlights the high relative abundance of *S. mutans* in caries-active cohorts.

The identification of *Actinomyces oris* (*A. oris*) in this study underscores the importance of inter-species coaggregation in the pathogenesis of dental caries. *A. oris* utilizes sortase-dependent pili to form a mutualistic partnership with oral streptococci [32], specifically through its type 2 pili and coaggregation factor CafA [33]. This structural synergism not only promotes the development of luxuriant, interdigitated microcolonies but also creates an acidic microenvironment conducive to the progression of SECC. The high predictive value of *A. oris* revealed by our SHAP analysis aligns with its biological role as a structural scaffold, reinforcing its status as a critical biomarker for early caries risk assessment.

Our study showed that *C. matruchotii* was relatively more abundant in the HC and SECCBS group. This result was consistent with the research conducted by *Qudeimat MA, et al* [34] using 16S rRNA sequencing and *Nezar Noor Al-Hebshi et al.* using metagenome sequencing [35]. It could be inferred that *C. matruchotii* might perform a key part in the formation and stabilization of healthy dental plaque [36]. The decreased abundance of *C. matruchotii* may be linked to the reduction of commensal species like *Streptococcus gordonii* [37] and *S. sanguinis* [38]. These commensals typically exert an antagonistic effect against *S. mutans* through hydrogen peroxide production; thus, their decline facilitates a microbial shift that allows *S. mutans* to dominate the oral ecosystem. This transition into a disease-adapted plaque disrupts acid-base metabolism, ultimately driving the

demineralization of tooth enamel. Consequently, *C. matruchotii* serves as a vital indicator of a caries-free state, where its diminishing presence signals the pathogenic evolution from a healthy biofilm to a cariogenic one.

Another key finding of this study was the identification of *V. parvula* as a dominant caries-associated bacterium. Lactate, a primary metabolite of *S. mutans*, serves as the preferred carbon source for most *Veillonella* spp. [39], establishing a syntrophic metabolic relationship between these two taxa. This metabolic consumption establishes a syntrophic "food chain" between these two taxa. Given that lactate is among the most potent acids produced by oral bacteria and a primary driver of dental tissue demineralization, its conversion into weaker volatile fatty acids by *Veillonella* spp. has been hypothesized to mitigate the low-pH environment and buffer against *S. mutans*-induced carious damage [40]. However, pH dynamic curves indicate that *V. parvula* may conversely stimulate acid production in *S. mutans*. Furthermore, the inherent acid tolerance of *S. mutans* enables its survival within such lethal, acidified microenvironments. In vitro experiments found that *V. parvula* can effectively promote *S. mutans* mature biofilm formation with enhanced acid resistance, hydrogen peroxide detoxicity, and biofilm virulence. Rodent model experiments revealed that *V. parvula* was incapable of causing disease on its own, but it significantly heightened the biofilm virulence of *S. mutans* when being co-infected and augmented the progression, quantity, and severity of dental caries [41]. Furthermore, the predictive value of unclassified phylotypes like ASV_15 highlights the utility of ensemble learning in uncovering 'rare biosphere' contributors. This aligns with the Extended Ecological Plaque Hypothesis, suggesting that SECC results from a functional shift across the entire microbial gradient rather than a single pathogen [42].

P. denticola significantly enhances the cariogenic potential of supragingival biofilms [43]. It facilitates the development of highly virulent biofilms on tooth surfaces, promoting caries by inducing oral microbial dysbiosis [44]. It could result in an increased volume of extracellular polysaccharides (EPS) and the up-regulation of *gtf* gene sets, indicating that this species actively induces EPS synthesis within the oral biofilm. The accumulation of the EPS matrix functions as a diffusion barrier, sequestering charged buffering ions and thereby impairing the neutralizing capacity of saliva. Conversely, uncharged solutes such as sucrose can readily diffuse through this matrix, where they are metabolized into organic acids. This selective permeability fosters a

localized, acidified microenvironment that promotes the demineralization of dental hard tissues [45]. Consistently, our study also identified *P.denticola* as a core consensus biomarker significantly associated with SECC. Through our ensemble machine learning framework, *P. denticola* was prioritized by multiple high-performance algorithms, reinforcing the premise that this species facilitates the transition from health-associated homeostatic states to cariogenic dysbiosis.

It is well-known that the 'core microbiome' in the plaque of patients with BS included *Actinomyces* (especially *A.naeslundii*), *Prevotella nigrescens*, *Pseudopropionibacterium*, *Rothia*, *Leptotrichia*, and *Neisseria* [4]. Consistently, our study confirmed that *Leptotrichia* acts as a key metabolic driver of acidogenesis and H₂S-related malodor [46], while facilitating the colonization of BTS-associated *Fusobacterium* [47]. Furthermore, our results substantiated the involvement of *Rothia* species in the BTS phenotype, supported by the identification of catechol-siderophore-like gene clusters in *R. mucilaginosa* and *R. dentocariosa* [48], which provides a molecular basis for the iron-sequestration and pigmentation observed in our SECCBS cohort. By integrating these high-dimensional signatures through ensemble ML, we demonstrate that the transition to a BTS state is driven by a distinct functional niche involving mineralization-associated taxa and the "rare biosphere.

Venn network diagrams represent robust analytical frameworks for elucidating interaction networks within genomics, proteomics, and metabolomics. By mapping complex interaction patterns, these visualizations provide pivotal insights into shared and distinct biological signatures across heterogeneous experimental datasets [49]. In the context of the present study, while traditional Venn diagrams are effective for illustrating overlaps between a limited number of sets, they become increasingly unwieldy and visually cluttered as the number of experimental groups or modeling iterations exceeds five. To circumvent these constraints, we implemented UpSet plots as a scalable, matrix-based alternative for visualizing the intricate intersections across our ensemble of 15 models. This approach was particularly instrumental in identifying consensus biomarkers that possessed both statistical robustness and cross-model predictive consistency. For instance, the UpSet matrix facilitated the clear identification of *S.mutans* as a top-tier predictor in the HC-SECC group, a finding validated by 10 out of 15 independent algorithms. By integrating UpSet plots with performance metrics, our research demonstrates that such advanced visualization strategies are

essential for distilling the most biologically significant signals from the high-dimensional 'rare biosphere' often encountered in dental microbiome research.

Conclusion

Our study demonstrates that the transition to SECC is driven by high-magnitude taxonomic shifts that are consistently captured across diverse algorithmic frameworks. By leveraging an ensemble of 15 machine learning models, we identified *S. mutans* as the keystone predictor with the highest stability, supported by a synergistic cluster including *A.oris*, *P.denticola*, and *V.parvula*. The superior predictive performance of the LightGBM, Dtree, and MLP models, further validated by significant clinical net benefits in DCA, underscores the robustness of these interpretable frameworks for the early identification of high-risk individuals in both SECC and SECCBS cohorts.

Limitations

This study also has several limitations. First, although SHAP analysis was employed to improve model interpretability, the directional interpretation of the identified feature contributions may still be influenced by label definitions and underlying feature distributions; the dimensionality of the predictive features remains relatively constrained; Second, the current models primarily focus on microbial taxa, excluding metabolites and specific clinical parameters. These factors likely play pivotal roles in the pathogenesis and prediction of caries. Future research could further enhance predictive performance through multi-omics integration or multimodal learning.

References

1. Stangvaltaite-Mouhat L, Skudutyte-Rysstad R, Ko H, Stankeviciene I, Aleksejuniene J, Puriene A: Co-occurrence of dental caries and periodontitis: multilevel modelling approach. *BMC Oral Health* 2024, 24(1):149.
2. Redwan SK, Alamoudi NM, Felemban OM, Al Tuwirqi AA, Alamoudi RA: Association of Early and Severe Early Childhood Caries with Oral Health-Related Quality of Life: A Cross-Sectional Survey. *Healthcare (Basel)* 2025, 13(23).
3. Yuan X, Chu Y, Cai W: Developing an Interpretable Machine Learning Framework to Predict and Analyse Early Childhood Caries in Children Aged 2 to 6 Years: A Single-centre Observational Study. *Int Dent J* 2026, 76(2):109419.
4. Dong X, Zhao W, Ma S, Li X, Li G, Zhang S: Oral microbial profiles of extrinsic black tooth

stain in primary dentition: A literature review. *J Dent Sci* 2024, 19(3):1369-1379.

5. Sangermano R, Pernarella S, Straker M, Lepanto MS, Rosa L, Cutone A, Valenti P, Ottolenghi L: The treatment of black stain associated with of iron metabolism disorders with lactoferrin: a litterature search and two case studies. *Clin Ter* 2019, 170(5):e373-e381.
6. Lopes MB, Coletti R, Duranton F, Glorieux G, Jaimes Campos MA, Klein J, Ley M, Perco P, Sampri A, Tur-Sinai A: The Omics-Driven Machine Learning Path to Cost-Effective Precision Medicine in Chronic Kidney Disease. *Proteomics* 2025, 25(11-12):e202400108.
7. Wu Y, Jia M, Fang Y, Duangthip D, Chu CH, Gao SS: Use machine learning to predict treatment outcome of early childhood caries. *BMC Oral Health* 2025, 25(1):389.
8. Casalegno F, Newton T, Daher R, Abdelaziz M, Lodi-Rizzini A, Schurmann F, Krejci I, Markram H: Caries Detection with Near-Infrared Transillumination Using Deep Learning. *J Dent Res* 2019, 98(11):1227-1233.
9. Wang H, Tian S, Bai S, Yang C, Jiang Z, Li N: Associations Between Thyroid Function and Periodontitis: A Machine Learning Approach Using NHANES. *Int Dent J* 2025, 75(5):100921.
10. Chubb JJ, Boyle AL, Albanese KI: Rational protein design. *Curr Opin Struct Biol* 2026, 97:103224.
11. Lee C, Gates KM: Automated machine learning for classification and regression: A tutorial for psychologists. *Behav Res Methods* 2025, 57(9):262.
12. Qiu W, Chen H, Dincer AB, Lundberg S, Kaeberlein M, Lee SI: Interpretable machine learning prediction of all-cause mortality. *Commun Med (Lond)* 2022, 2:125.
13. AlShehhi A, Alblooshi H, Fadul R, Tumzghi N, Tenaiji AA, Harbi MA, Al-Jasmi F: Comparison of machine learning models for mucopolysaccharidosis early diagnosis using UAE medical records. *Sci Rep* 2025, 15(1):28813.
14. Zou R, Lin Y, Da C, Liao G: Novel nomogram and decision curve analysis for predicting head and neck skin cancer risk. *Sci Rep* 2025, 15(1):41555.
15. Freeman SC, Saeedi E, Ordonez-Mena JM, Nevill CR, Hartmann-Boyce J, Caldwell DM, Welton NJ, Cooper NJ, Sutton AJ: Data visualisation approaches for component network meta-analysis: visualising the data structure. *BMC Med Res Methodol* 2023, 23(1):208.
16. Li Y, Yang Z, Cai T, Jiang D, Luo J, Zhou Z: Untargeted metabolomics of saliva in caries-

- active and caries-free children in the mixed dentition. *Front Cell Infect Microbiol* 2023, 13:1104295.
17. Gun Topcu AI, Erdogan Y, Elci SG: Biochemical Saliva Analysis and Caries Experience In Children With Black Tooth Stain. *Oral Health Prev Dent* 2025, 23:665-672.
 18. Zhang L, Chen Y, Xie J, Ding G, Wang GB, Zheng D, Lu Y: Comparative analysis of the dental plaque metabolome in tooth stain and caries within primary dentition. *BMC Oral Health* 2026, 26(1).
 19. Papolos DF, Teicher MH, Post RM: Treatment of Early-Onset Specified and Unspecified Bipolar Disorders: A Systematic Review and Strategies for Identifying and Managing a Thermally Dysregulated Subtype in Children. *Acta Psychiatr Scand* 2025, 152(3):156-179.
 20. Zhang J, Wang Y: Meta-analysis of prevalence and filling rate of dental caries in preschool children in China. *Hua Xi Kou Qiang Yi Xue Za Zhi* 2023, 41(5):573-581.
 21. Chen X, Zhan JY, Lu HX, Ye W, Zhang W, Yang WJ, Feng XP: Factors associated with black tooth stain in Chinese preschool children. *Clin Oral Investig* 2014, 18(9):2059-2066.
 22. Qiao C, Han R, Yang J, Huang H, Ma L: Black stain and dental caries in primary dentition of preschool children in Qingdao, China. *J Clin Pediatr Dent* 2024, 48(4):200-205.
 23. Derelioglu SS, Ozkan C, Gokdas BP, Yerli EB, Demirdogen EY, Kosan Z, Celikel P, Sarac F, Ozturk AB, Buyuksefil M: Impacts of early childhood caries on the mental development. *BMC Oral Health* 2025, 25(1):2005.
 24. Khazaei Y, Kodikara S, Butler CA, Messina NL, Le Cao KA, Dashper SG, Silva MJ: Development and validation of diagnostic and prognostic prediction tools for dental caries in young children through prospective and cross-sectional observational studies: a protocol. *BMJ Open* 2025, 15(10):e105145.
 25. Chen L, Wang X, Zhu K, Ren K, Wu Z: [Machine learning-based prediction model for caries in the first molars of 9-year-old children in Suzhou]. *Hua Xi Kou Qiang Yi Xue Za Zhi* 2025, 43(6):871-880.
 26. Frenkel E, Neumayr J, Schwarzmaier J, Kessler A, Ammar N, Schwendicke F, Kuhnisch J, Dujic H: Caries Detection and Classification in Photographs Using an Artificial

- Intelligence-Based Model-An External Validation Study. *Diagnostics (Basel)* 2024, 14(20).
27. Felsch M, Meyer O, Schlickerrieder A, Engels P, Schonewolf J, Zollner F, Heinrich-Weltzien R, Hesenius M, Hickel R, Gruhn V *et al*: Detection and localization of caries and hypomineralization on dental photographs with a vision transformer model. *NPJ Digit Med* 2023, 6(1):198.
 28. Park EY, Cho H, Kang S, Jeong S, Kim EK: Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. *BMC Oral Health* 2022, 22(1):573.
 29. Fu WT, Zhu QK, Li N, Wang YQ, Deng SL, Chen HP, Shen J, Meng LY, Bian Z: Clinically Oriented CBCT Periapical Lesion Evaluation via 3D CNN Algorithm. *J Dent Res* 2024, 103(1):5-12.
 30. Wang Y, Xie Y, Luo Y, Jia P, Wei J, Zhang J, Yan W, Huang J: iASMP: An interpretable in-silico predictive tool focusing on species-specific antimicrobial peptides. *J Pept Sci* 2023, 29(9):e3490.
 31. Mazurel D, Brandt BW, Boomsma M, Crielaard W, Lagerweij M, Exterkate RAM, Deng DM: Streptococcus mutans and Caries: A Systematic Review and Meta-Analysis. *J Dent Res* 2025, 104(6):594-603.
 32. Wu C, Mishra A, Yang J, Cisar JO, Das A, Ton-That H: Dual function of a tip fimbrillin of Actinomyces in fimbrial assembly and receptor binding. *J Bacteriol* 2011, 193(13):3197-3206.
 33. Yadav RK, Krishnan V: New structural insights into the PI-2 pilus from Streptococcus oralis, an early dental plaque colonizer. *FEBS J* 2022, 289(20):6342-6366.
 34. Qudeimat MA, Alyahya A, Karched M, Behbehani J, Salako NO: Dental plaque microbiota profiles of children with caries-free and caries-active dentition. *J Dent* 2021, 104:103539.
 35. Al-Hebshi NN, Baraniya D, Chen T, Hill J, Puri S, Tellez M, Hasan NA, Colwell RR, Ismail A: Metagenome sequencing-based strain-level and functional characterization of supragingival microbiome associated with dental caries in children. *J Oral Microbiol* 2019, 11(1):1557986.
 36. Li Q, Zhou F, Su Z, Li Y, Li J: Corynebacterium matruchotii: A Confirmed Calcifying Bacterium With a Potentially Important Role in the Supragingival Plaque. *Front Microbiol*

2022, 13:940643.

37. Agnello M, Marques J, Cen L, Mittermuller B, Huang A, Chaichanasakul Tran N, Shi W, He X, Schroth RJ: Microbiome Associated with Severe Caries in Canadian First Nations Children. *J Dent Res* 2017, 96(12):1378-1385.
38. Richards VP, Alvarez AJ, Luce AR, Bedenbaugh M, Mitchell ML, Burne RA, Nascimento MM: Microbiomes of Site-Specific Dental Plaques from Children with Different Caries Status. *Infect Immun* 2017, 85(8).
39. Scheiman J, Lubner JM, Chavkin TA, MacDonald T, Tung A, Pham LD, Wibowo MC, Wurth RC, Punthambaker S, Tierney BT *et al*: Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. *Nat Med* 2019, 25(7):1104-1109.
40. Wicaksono DP, Washio J, Abiko Y, Domon H, Takahashi N: Nitrite Production from Nitrate and Its Link with Lactate Metabolism in Oral Veillonella spp. *Appl Environ Microbiol* 2020, 86(20).
41. Wei Y, Zhang Y, Zhuang Y, Tang Y, Nie H, Haung Y, Liu T, Yang W, Yan F, Zhu Y: Veillonella parvula acts as a pathobiont promoting the biofilm virulence and cariogenicity of Streptococcus mutans in adult severe caries. *Microbiol Spectr* 2024, 12(11):e0431823.
42. Xiao J, Grier A, Faustoferri RC, Alzoubi S, Gill AL, Feng C, Liu Y, Quivey RG, Kopycka-Kedzierawski DT, Koo H *et al*: Association between Oral Candida and Bacteriome in Children with Severe ECC. *J Dent Res* 2018, 97(13):1468-1476.
43. Forssten SD, Bjorklund M, Ouwehand AC: Streptococcus mutans, caries and simulation models. *Nutrients* 2010, 2(3):290-298.
44. Li C, Qi C, Yang S, Li Z, Ren B, Li J, Zhou X, Cai H, Xu X, Peng X: F(0)F(1)-ATPase Contributes to the Fluoride Tolerance and Cariogenicity of Streptococcus mutans. *Front Microbiol* 2021, 12:777504.
45. Wu H, Li N, Yang S, Qiu J, Zhang M, Wang L, Gao R, Wu L, Yu Q, Cheng X: Prevotella denticola promotes caries by inducing oral microbial dysbiosis. *BMC Oral Health* 2026, 26(1).
46. Chen Y, Dou G, Wang D, Yang J, Zhang Y, Garnett JA, Chen Y, Wang Y, Xia B: Comparative Microbial Profiles of Caries and Black Extrinsic Tooth Stain in Primary

Dentition. *Caries Res* 2021, 55(4):310-321.

47. Li Y, Zou CG, Fu Y, Li Y, Zhou Q, Liu B, Zhang Z, Liu J: Oral microbial community typing of caries and pigment in primary dentition. *BMC Genomics* 2016, 17:558.
48. Hirtz C, Mannaa AM, Moulis E, Pible O, O'Flynn R, Armengaud J, Jouffret V, Lemaistre C, Dominici G, Martinez AY *et al*: Deciphering Black Extrinsic Tooth Stain Composition in Children Using Metaproteomics. *ACS Omega* 2022, 7(10):8258-8267.
49. Yang M, Chen T, Liu YX, Huang L: Visualizing set relationships: EVenN's comprehensive approach to Venn diagrams. *Imeta* 2024, 3(3):e184.

Abbreviations

ASVs: Amplicon Sequence Variants

ML: machine learning

SECC: severe early childhood caries

HC: health control

SECCBS: severe early childhood caries and extrinsic black stain

Shap: SHapley Additive exPlanations

DCA: Decision Curve Analysis

TSS: Total Sum Scaling

CLR: Centered Log-Ratio

MI: Mutual Information

VIF: Variance Inflation Factor

LR: Logistic Regression

LASSO: Ridge Regression, Least Absolute Shrinkage and Selection Operator

Linear SVM: Linear Support Vector Machine

LDA: Linear Discriminant Analysis

NBC: Naive Bayes Classifier

Dtree: Decision Tree

RF: Random Forest

Extra-Trees: Extremely Randomized Trees

XGBoost: Extreme Gradient Boosting

LightGBM: Light Gradient Boosting Machine

MLP: Multi-layer Perceptron

SVM-RBF: Support Vector Machine with Radial Basis Function kernel

KNN: K-Nearest Neighbors

SVM-Poly: Polynomial SVM

ROC-AUC: Area Under the ROC Curve

NPV: Negative predictive value

PPV: Positive Predictive Value

MCC: Matthews Correlation Coefficient

CI: Confidence Intervals

Declarations

The authors have no conflicts of interest to declare.

Ethics approval and consent to participate

The study protocol was reviewed and approved by the Shenzhen Children's Hospital ethics committee (No. 202319202) and conformed to the ethical standards for medical research involving human subjects, as laid out in the 1964 Declaration of Helsinki and its later amendments. The participants, including the children and their guardians, provided written informed consent prior to taking part in the study. All the subjects agreed to study par dataset.

Consent for Publication

Not applicable.

Data availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2025) in National Genomics Data Center (Nucleic Acids Res 2026), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA PRJCA067562) that are publicly accessible at <https://ngdc.cncb.ac.cn/gsa>.

Funding

This study was funded by the Shenzhen Science and Technology Program (No. JCYJ20230807093811023).

Competing interests

The authors declare no competing interests.

Graphical Abstract

Figure 1. Flow diagram of the study.

Figure 2. Schematic overview of the framework of this study.

Figure 3. ROC curves of multiple machine learning models in external validation cohorts. The predictive performance was evaluated across different model types for the HC_SECC (**A**) and the SECC_SECCBS (**B**) cohort.

Figure 4. Intersection and selection frequency of key microbial features across 15 independent machine learning models. (**A, B**) UpSet plot illustrating the feature intersection for the HC vs. SECC and the SECC vs. SECCBS cohort, respectively; (**C, D**) Horizontal bar chart illustrating selection frequency for the HC vs. SECC and the SECC vs. SECCBS cohort, respectively.

Figure 5. Multi-model consensus and relative abundance distribution of the core 20-biomarker signature across cohorts. (**A, B**) Hierarchical clustering heatmaps integrate the validation of 15 independent machine learning models with relative microbial expression for the HC vs. SECC group and the SECC vs. SECCBS group, respectively. In both panels, rows represent identified microbial taxa at the ASV level, columns represent clinical cohorts. The blue-to-orange gradient denotes mean expression levels (deep orange: high enrichment; deep blue: low expression), and horizontal gray bar plots quantify model consensus frequency.

Figure 6. Decision curve analysis (DCA) evaluating the clinical utility of 15 independent machine learning models for SECC risk stratification.

Figure 7. Global feature importance analysis based on SHAP values across 15 machine learning models for cohort differentiation. (**A, B**) Grid layout displaying the relative importance of the top 15 microbial ASVs across 15 distinct algorithmic frameworks for the HC vs. SECC group and the SECC vs. SECCBS group, respectively. In each individual subpanel, where the y-axis lists the specific ASVs, the x-axis quantifies the global contribution via the mean absolute SHAP value, and horizontal orange bars correspond to the magnitude of feature importance.

Figure 8. SHAP summary and force plots illustrating model-dependent feature contributions and individual predictions. **(A)** SHAP beeswarm plot for the XGBoost and RF model ranking the top 15 microbial features at the ASV level in the HC vs. SECC group, with dots representing individual samples and the purple-to-orange gradient corresponding to low-to-high relative feature values; **(B)** SHAP beeswarm plot for the MLP and DTree model in the SECC vs. SECCBS group; **(C,D)** Horizontal force plot visualizing individual sample predictions and tracking the shift from the base value to the final output for the XGBoost and RF model in the HC vs. SECC group, respectively. yellow and purple arrows indicate positive and negative contributing forces; **(E,F)** Horizontal force plot visualizing individual sample predictions and tracking the shift from the base value to the final output for the MLP model and DTree model in the SECC vs. SECCBS group, respectively.

Figure 9. UpSet plots and consensus frequency bar charts illustrating multi-model biomarker selection across SECC phenotypes. **(A, B)** Feature intersection matrix evaluating the robustness of microbial signatures across 15 independent machine learning algorithms for the HC vs. SECC group and the SECC vs. SECCBS group, respectively; **(C, D)** Rank-ordered selection counts for the HC vs. SECC group and the SECC vs. SECCBS group, respectively.