AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Human Microbiome | Research Article

Machine-learning analysis of cross-study samples according to the gut microbiome in 12 infant cohorts

Petri Vänni,[1] Mysore V. Tejesvi,[1,2] Niko Paalanne,[1,3] Kjersti Aagaard,[4] Gail Ackermann,[5] Carlos A. Camargo Jr.,[6] Merete Eggesbø,[7,8] Kohei Hasegawa,[6] Anne G. Hoen,[9] Margaret R. Karagas,[9] Kaija-Leena Kolho,[10] Martin F. Laursen,[11] Johnny Ludvigsson,[12] Juliette Madan,[13,14] Dennis Ownby,[15] Catherine Stanton,[16] Jakob Stokholm,[17,18] Terhi Tapiainen[1,4,19]

AUTHOR AFFILIATIONS See affiliation list on p. 16.

ABSTRACT Combining and comparing microbiome data from distinct infant cohorts has been challenging because such data are inherently multidimensional and complex. Here, we used an ensemble of machine-learning (ML) models and studied 16S rRNA amplicon sequencing data from 4,099 gut microbiome samples representing 12 prospectively collected infant cohorts. We chose the childbirth delivery mode as a starting point for such analysis because it has previously been associated with alterations in the gut microbiome in infants. In cross-study ensemble models, Bacteroides was the most important feature in all machine-learning models. The predictive capacity by taxonomy varied with age. At the age of 1–2 months, gut microbiome data were able to predict delivery mode with an area under the curve of 0.72 to 0.83. In contrast, ML models trained on taxa were not able to differentiate between the modes of delivery, in any of the cohorts, when the infants were between 3 and 12 months of age. Moreover, no ML model, alternately trained on the functional pathways of the infant gut microbiome, could consistently predict mode of delivery at any infant age. This study shows that infant gut microbiome data sets can be effectively combined with the application of ML analysis across different study populations.

IMPORTANCE There are challenges in merging microbiome data from diverse research groups due to the intricate and multifaceted nature of such data. To address this, we utilized a combination of machine-learning (ML) models to analyze 16S sequencing data from a substantial set of gut microbiome samples, sourced from 12 distinct infant cohorts that were gathered prospectively. Our initial focus was on the mode of delivery due to its prior association with changes in infant gut microbiomes. Through ML analysis, we demonstrated the effective merging and comparison of various gut microbiome data sets, facilitating the identification of robust microbiome biomarkers applicable across varied study populations.

KEYWORDS machine learning, bioinformatics, human microbiome, gut microbiome, random forest, infant, children, cross-study, ensemble

It has been suggested that childbirth delivery mode, Caesarean delivery, is associated with varying degrees of greater risk of non-communicable diseases later in life, notably asthma (1, 2), food allergies (3, 4), obesity (5), and diabetes (6) among offspring even though studies with high-quality designs have not given consistent results regarding these associations (7). In most studies, but not all, Caesarean delivery has been associated with an altered gut microbiome composition for neonates or infants (8) principally relatively lower abundances of *Escherichia-Shigella* (9) and *Parabacteroides* (10) and delayed colonization with *Bacteroides* (9–14) and *Bifidobacterium* (10, 14, 15) with a contrasting relative enrichment in *Clostridium* (10, 11)

Editor Renuka R. Nayak, University of California, San Francisco, California, USA

Address correspondence to Petri Vänni, petri.vanni@oulu.fi.

The authors declare no conflict of interest.

See the funding table on p. 17.

Received 14 April 2023
Accepted 13 September 2023
Published 24 October 2023

[This article was published on 24 October 2023 with typographical errors in the introduction and Materials and Methods. The text was corrected in the current version, posted on 30 October 2023.]

Most gut microbiome studies have used sequencing data of the bacterial 16S rRNA gene or less often whole-genome bacterial sequencing. As microbiome data are multidimensional and noisy, it is difficult to combine data from two or more populations for traditional statistical testing of a hypothesis (16). It has been suggested that machine learning (ML) models may help to overcome this limitation (17–19) because the model can train on specific data set and then be used on further data set, and its efficiency validated. To date, there have been a limited number of studies combining or comparing gut microbiome data from several available prospective cohort studies in neonates, infants, or children using the ML approach, such as random forest (17–20), support vector machine (17, 18), elastic net (17, 18), and gradient-boosted machine algorithms (18).

Here, we use an ensemble of ML models, including random forest (21) (RF), extremely randomized trees (22) (EXTRA), light gradient-boosting machine (23) (LGBM), and multilayer perceptron (MLP) predictive models across 12 prospective pediatric cohorts with gut microbiome data originating from 6 different countries. To evaluate the usefulness of ML algorithms in combining and comparing microbiome data across different cohort studies, we compared the association of delivery mode with gut microbiome composition in the cohorts.

MATERIALS AND METHODS

Literature search and data set recruitment

A systematic literature review was conducted in the Web of Science, Scopus, PubMed, and Google Scholar databases up to January 2020. Additionally, the clinicaltrials.gov website was searched for suitable studies (Fig. 1). We used the following terms to search through the titles, abstracts, and keywords of the literature in our set of materials: (infant AND cohort AND microbiome AND 16S) AND (fecal OR stool OR gut). Our inclusion criteria for data sets were that the studies should have more than 50 infant fecal samples with 16S microbiome data available from the first 12 months of life, with defined sampling times. Birth cohorts containing only preterm infants were excluded. The data set correspondents were invited to participate in this multicohort collaboration.

All the institutions' original data sets and protocols were approved by their institutional review boards and ethical committees, and all families of the infants provided their written informed consent. Only 16S rRNA amplicon sequence data and data on the delivery mode and breastfeeding were used here, and no individual personal data were transferred or used.

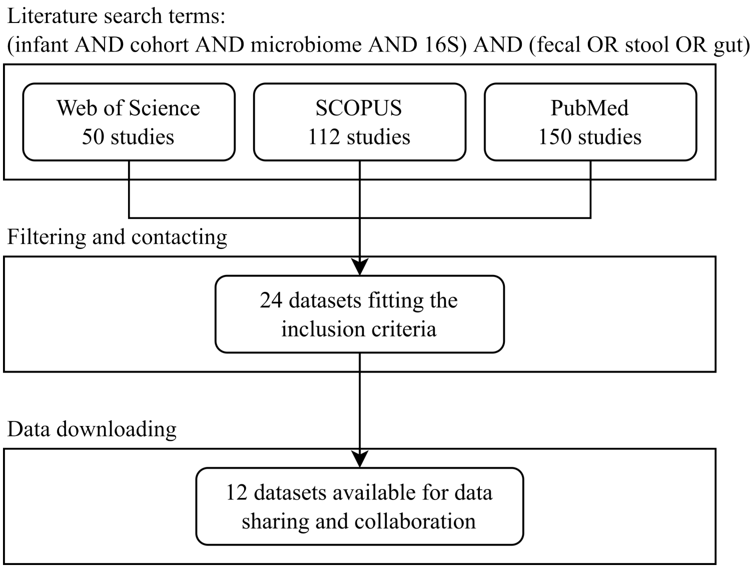

FIG 1 Flowchart of the literature search.

After contacting 24 research groups, 12 of these groups participated and provided access to 16S rRNA amplicon sequence data sets of fecal samples, along with data on the mode of delivery (Table 1). Microbiome development causes large shifts over time in the gut microbiome of infants in the first year of life (24, 25). As such, we analyzed data sets in three age groups: 1–2 months, 3–6 months, and 9–12 months. Seven infant cohorts had fecal samples available 1–2 months after birth, four cohorts had samples 3–6 months after birth, and eight had samples 9–12 months after birth. Altogether, we had 16S rRNA amplicon sequencing data available for 4,099 fecal samples, or 3,595 samples, after pre-processing and quality filtering. There were 1,457 samples collected at 1–2 months of age, of which 440 were from infants delivered by Caesarean and 1,017 were from infants delivered vaginally (Table 1). At 3–6 months of age, we had 473 samples, comprising 201 from infants delivered by Caesarean and 272 samples from vaginally delivered infants. At 9–12 months of age, we had 1,665 samples, of which 363 were from infants delivered by Caesarean and 1,302 were from infants delivered vaginally.

Sequence pre-processing

Before the data were analyzed using ML methods, each data set was prepared, quality filtered using similar methods, transformed into relative abundance information for each bacterial taxon or metabolic pathway in each sample, and presented in feature tables. The pre-processing pipeline is shown in Fig. 2. The sequences were downloaded from their repositories or acquired directly from the corresponding researchers (Table 1) before being imported into the Qiime2 (37) (version 2021.11) microbiome bioinformatics platform using the q2-tools module. The primer sequences were removed from each data set using the q2-cutadapt tool, and the open-source software package DADA2 (38) was used to de-noise the sequences into amplicon sequence variants (ASVs) using the q2-dada2 module, where the trunc-len parameter was set to zero. ASVs, found in fewer than 2 samples and in a total frequency of 10, were removed. Taxonomy was assigned using the SILVA (39) (version 138) database with a Naïve Bayes classifier. ASVs

TABLE 1 Study cohort characteristics

	Country of origin	Initial number of fecal samples	Available samples after pre-processing	Infants born via Caesarean delivery	Infants born via vaginal delivery	Mean age of infants (months)
Sampled at 1–2 months						
COPSAC (26)	Denmark	505	303	72	231	1
HOUSTON[a](27)	USA	52	48	11	37	1.5
INFANTMET (28)	Ireland	167	137	65	72	1
JORVI (29)	Finland	68	51	8	43	1
NHBCS (30)	USA	321	319	92	227	1.5
NOMIC (31)	Norway	485	485	159	326	1
WHEALS (32)	USA	130	114	33	81	1.2
Sampled at 3–6 months						
INFANTMET (28)	Ireland	152	152	86	66	5.5
JORVI (29)	Finland	68	62	10	52	6
MARC-43 (33)	USA	115	115	43	72	3.4
WHEALS (32)	USA	167	144	62	82	6.6
Sampled at 9–12 months						
ABIS (34)	Sweden	403	399	47	352	12
COPSAC (26)	Denmark	623	424	90	334	12
JORVI (29)	Finland	62	62	10	52	12
NHBCS (30)	USA	135	135	38	97	12
NOMIC (31)	Norway	340	340	103	237	12
OULU (35)	Finland	84	84	23	61	12
SKOT1 (36)	Denmark	115	115	16	99	9
SKOT2 (36)	Denmark	107	106	36	70	9

[a]Pregnant women prospectively enrolled in the early third trimester.

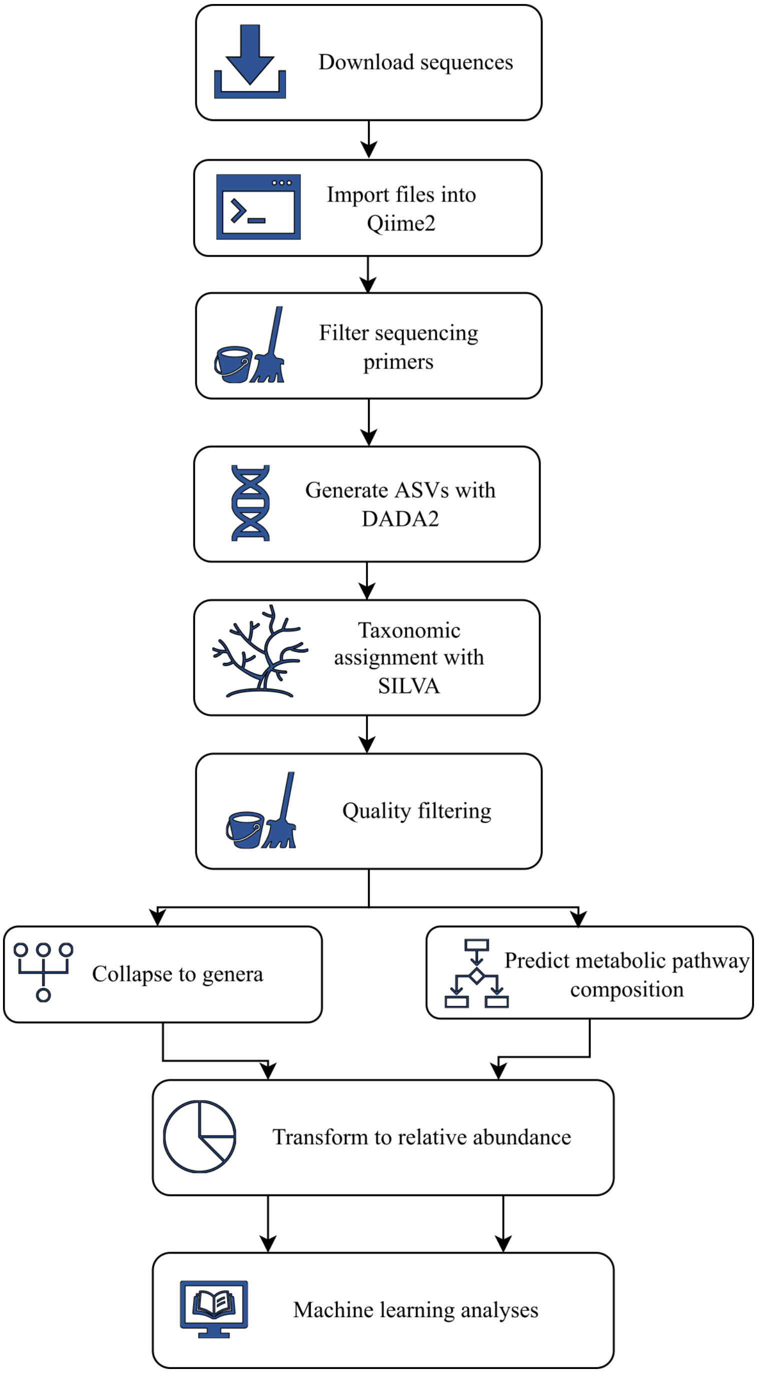
Download sequences
Import files into Qiime2
Filter sequencing primers
Generate ASVs with DADA2
Taxonomic assignment with SILVA
Quality filtering
Collapse to genera
Predict metabolic pathway composition
Transform to relative abundance
Machine learning analyses

FIG 2 Step-by-step flowchart of the pre-processing pipeline.

classified as mitochondria or chloroplasts were removed. The ASVs were collapsed to the taxonomic level of genera and transformed into relative abundances for downstream analyses. A predicted metabolic pathway composition was produced from the filtered ASV feature table using PICRUSt2 software (40). Additionally, a second set of data sets was generated with Greengenes (41) instead of SILVA for alternative ML model-building attempts. Each data set was pre-processed in the same way, except for the data from NOMIC, which were only available as a pre-processed ASV table with the taxonomic classification assigned with only Greengenes (41) database. A representative sequence file was prepared, and the data for pre-processing were input into the pre-processing pipeline as a taxonomic assignment step (Fig. 2), after which the protocol continued as for the other data sets.

Alpha and beta diversity of the microbiome between cohorts

Diversity indices were calculated from rarefied genera-collapsed feature tables to visualize cohort differences with the q2 diversity plugin. For each sampling interval, the rarefying depth was chosen with the following rules: (i) depth of 1,000 is the minimum and (ii) choose the next highest value without losing samples in the rarefying process. For the 1–2 months time period, the chosen depth was 1,027; for 3–6 months, it was 1,116; and for 9–12 months, it was 1,009. Cohort and sample diversity were analyzed with Shannon diversity index and Bray–Curtis dissimilarity index for the alpha and beta diversity, respectively. The Bray–Curtis dissimilarity index was further analyzed with PCoA. The results were plotted with Matplotlib (42), and Kruskal–Wallis H-tests were conducted to examine the statistical differences between infants born via Caesarean delivery and those born via vaginal delivery. *P*-values were adjusted for multiple testing using the Benjamini–Hochberg procedure.

Machine-learning analyses

Before training an ML model, a set of settings, i.e., "hyperparameters," needs to be defined, followed by a search for the optimal combination. These settings define the structure and behavior of the models, such as the depth of the decision trees or the number of layers in a neural network. A common way of tuning the hyperparameters is to use a nested cross-validation method, in which the hyperparameters are selected using the training fold with an additional cross-validation loop (43). Here, the ML models were trained, tested, and validated by means of a nested cross-validation scheme with 40 repetitions. Each data set was first split in an outer cross-validation loop, where each fold, in turn, was used as the validation fold and the rest of the data were used in an inner cross-validation loop. The model building and parameter tuning took place only in the inner cross-validation loop. The performance of the final model was validated using the outer cross-validation folds and then averaged and recorded. Previous microbiome studies have chosen different *k* values for *k*-fold cross validation, such as 5 (44) or 10 (17, 18) folds. Ten folds have been recommended for biomedical data with high dimensions (45). As such, to maximize number of samples used in model training, the number of folds was set to as close to 10 as possible. In both *k*-fold cross-validation loops, the number of folds was set at 10, except in the JORVI cohort (8 outer folds and 7 inner folds) and the HOUSTON cohort (10 outer folds and 9 inner folds) for fecal samples obtained at 1–2 months of age and the JORVI cohort (10 outer folds and 9 inner folds) for samples obtained at 3–6 and 9–12 months of age, as there must be at least one of each class (Caesarean delivery and vaginal delivery) in the testing and validation folds to calculate the receiver operating characteristic (ROC) curve.

The feature importance of the models was estimated using the scikit-learn (46) function termed permutation importance. The scikit function takes in a trained model and testing data set where each feature is shuffled among all the samples in the testing data. Feature importance is defined by how much the prediction performance of the model is lowered following the shuffling as compared with a situation in which the feature is included in the model. In brief, a higher feature importance value indicates a greater importance of the feature to the model.

The ML classifier performances were estimated using the area under the curve (AUC) for the ROC and precision-recall (PR) curves. The model performance and feature importance values were averaged over 40 nested cross-validation loop repetitions.

Feeding mode analyses

To better understand our findings, we ran additional analyses on feeding mode in the cohorts for which we had feeding mode data available at the 1–2 months time point. With these *post hoc* analyses, we attempted to control for the confounding effect of breastfeeding when predicting delivery mode. We produced additional analyses using the 1–2 months data sets from COPSAC, HOUSTON, INFANTMET, and JORVI, where infants fed only formula were removed.

We also employed the Fisher's exact test to examine if Caesarean delivery or vaginally delivered groups had statistically more exclusively breastfed, partially breastfed, or formula-fed only infants in each cohort where breastfeeding data were available.

Machine-learning hyperparameter tuning

Since decision tree-based algorithms have performed well in previous microbiome studies (17, 18, 20, 47, 48) and because neural networks (47, 49) show great promise for the analysis of several microbiome-related problems, we chose three decision tree algorithms and one deep learning algorithm for use here: RF (21), EXTRA (22), MLP, and the LGBM (23). The hyperparameters were tuned in the inner cross-validation loop using the scikit-learn RandomizedSearchCV function in which the n_iter parameter was set at 40 iterations. RandomizedSearchCV was used to tune the hyperparameter efficiently without having to go through all possible hyperparameter iterations (50).

The hyperparameters tuned for the random forest and extremely randomized trees algorithms were max_depth, max_features, class_weight, and bootstrap, the last-mentioned for random forest only. The hyperparameters searched for LGBM models were num_leaves, max_depth, n_estimators, reg_alpha, and learning_rate. The MLP models were trained using the "adam" solver in scikit-learn, the hyperparameters that were tuned being max_iter, alpha, learning_rate_init, and momentum. In the hyperparameter "hidden_layer_sizes," the number of layers ranged from 1 to 3, with 10, 30, 50, or 100 neurons in the first layer, while in the models with multiple layers, each subsequent layer had half the number of neurons than the previous layer. The MLP hyperparameters for parameter tuning were chosen based on previously published work (49). The exact hyperparameter values used for parameter tuning are shown in Table S1.

Cross-study machine learning using gut microbiome data

We then used a cross-study approach in which we aimed to test whether an ML model developed using certain given data sets is generalizable to other data sets with regard to the mode of delivery as an explanatory variable for gut microbiome composition. In addition, the mode of delivery was predicted in a cross-study manner so that each cohort's outer cross-validation samples were predicted using best-performing models for all the other cohorts. These models were collected into an ensemble classifier in which the delivery mode of a given validation sample was predicted based on the averaged prediction of the best models for all the other cohorts. In this way, the same testing samples can be used for both the within-study and cross-study methods, and the performances are more readily comparable.

PipelineSearch and control augmenting methods

There are countless combinations and ways in which to build ML models, and these can produce different results. Similarly, there are several options for each preprocessing step when handling 16S sequencing samples that affect downstream analyses, such as which software, taxonomic database, or collapsing level to choose (51, 52). Therefore, we developed "PipelineSearch" as a novel method to automate those choices. Instead of the researcher choosing which taxonomic database to use, such as SILVA or Greengenes, PipelineSearch they are chosen at the same time as hyperparameters in ML parameter tuning. During hyperparameter tuning, the models could select which taxonomic database was used in preprocessing, Greengenes or SILVA. Similarly, PipelineSearch could select between feature table types, predicted metabolic pathways or genera collapsed data.

We also used an approach referred to as control augmenting (20), in which additional control samples from outside data sets were added to the training data for the models.

Wirbel et al. (20) increased the number of control samples fivefold in each training fold. To achieve similar numbers, we considered COPSAC and INFANTMET the augmenting cohort, as they had the most control samples at their respective time points. We did not consider the NOMIC data set, as we had no control over its early pre-processing steps, and thus, in this approach, the additional control samples came from COPSAC (1–2 months group), INFANTMET (3–6 months group), and COPSAC (6–9 months group) for their respective sampling time points. The COPSAC and INFANTMET cohorts were not used for model building or cross-study validation at the sampling times, where they were used to augment all the other cohorts, as this would leak information between validation folds.

Reproducibility and code availability

The code used in the present ML analyses is available in the GitHub repository (https://github.com/pvanni/PipelineSearch). We reported our finding according to the Strengthening The Organization and Reporting of Microbiome Studies guidelines (53) and the checklist can be found in the GitHub repository. Installed Python packages are listed in (Table S2).

Random number generators were seeded to guarantee identical outer cross-validation splits for each algorithm choice in addition to rendering the results reproducible. In this way, each model was validated using the same validation samples, making direct comparison of their AUC values reliable in both within-study and cross-study predictions.

Full sequencing data for all cohorts used can be found from public data repositories, and their corresponding accession numbers can be found in (Table S3). Relative abundance feature tables from all cohorts used in genera and predicted pathway ML-analyses can be found in the supplemental material (Data S1 through S4) with delivery mode metadata linked to each sample as the last column.

RESULTS

Characteristics of the cohorts

The general characteristics of the populations and the 16S rRNA amplicon sequence data sets are presented in Table 1 and (Table S3). The microbiome data sets were further characterized by plotting alpha and beta diversity indices for each infant cohort (Fig. 3) and for each time point (Fig. S1). Shannon's diversity index was, on average, lower in the 1–2 months cohorts (mean = 1.8, SD = 0.26) than in the 3–6 months (mean = 2.16, SD = 0.37) or 9–12 months (mean = 2.46, SD = 0.58) cohorts. Alpha diversity did not differ significantly according to mode of delivery in any of the cohorts (Fig. 3). The Fisher's exact test showed no significant enrichment of breast- or formula-fed samples in either Caesarean delivery or vaginally delivered groups in any cohort (Tables S4 and S5).

The machine-learning models accurately predicted the delivery mode from the fecal microbiome taxonomic data at 1–2 months of age

Machine learning can be used to train models to predict target variables, such as the mode of delivery in the present case, from unknown samples using input variables such as the relative abundances of bacteria. In the initial training of the ML models, we used four algorithms (RF, EXTRA, LGBM, and MLP) to differentiate between children born by vaginal delivery and Caesarean delivery, using the gut microbiome data from fecal samples obtained for each cohort at each of the time points (Fig. 4).

The ML models were, indeed, effective in predicting the mode of delivery on this basis at 1–2 months of age. The ML models achieved high AUC values ranging from 0.73 to 0.82 in all cohorts, depending on the ML model selected (Fig. 4), while the RF models were the best in the COPSAC (AUC = 0.73), NHBCS (AUC = 0.79), NOMIC (AUC = 0.82), and WHEALS (AUC = 0.79) cohorts. The EXTRA models performed well in the INFANTMET (AUC = 0.80) and JORVI (AUC = 0.74) cohorts, but the LGBM model achieved the highest

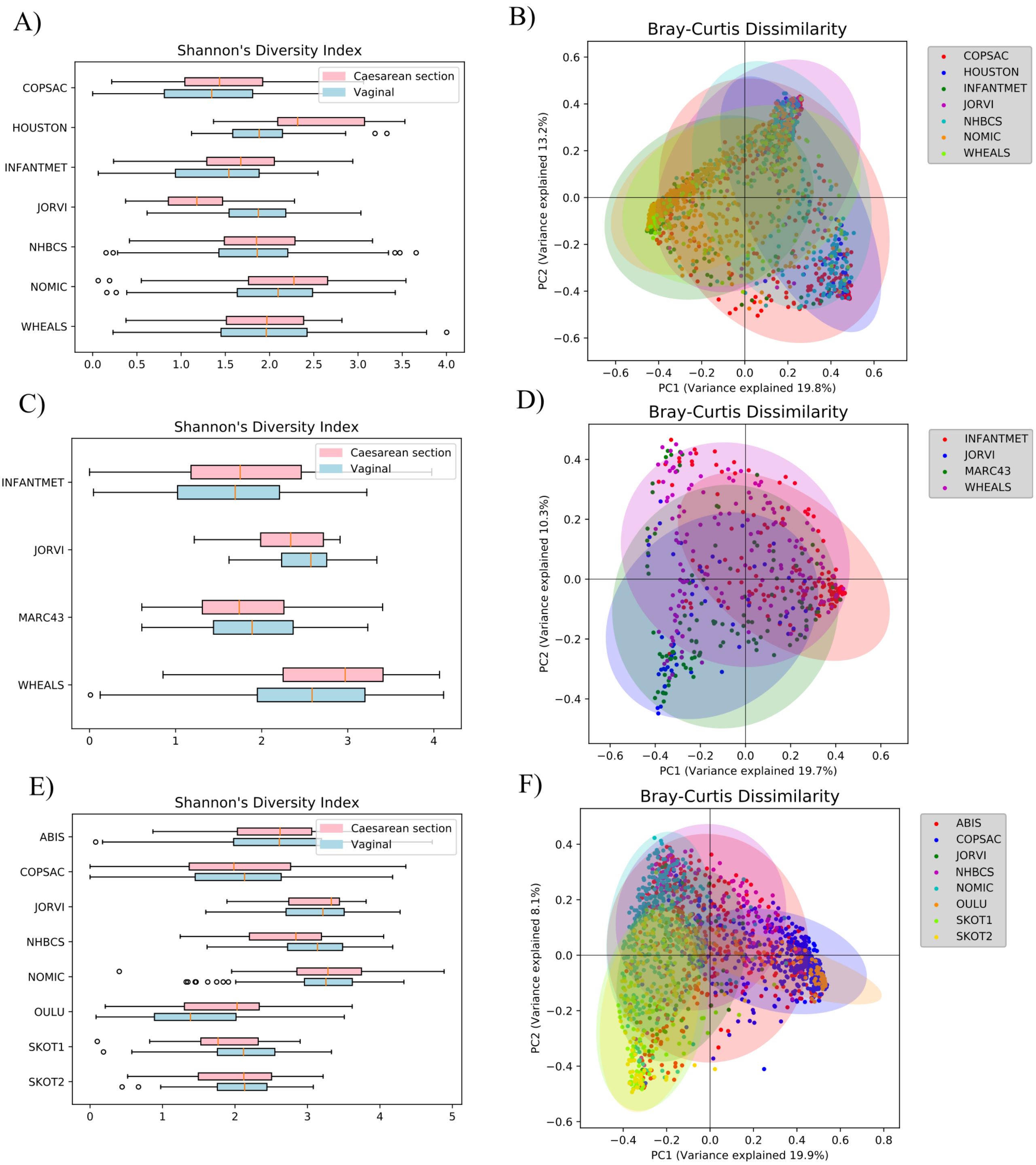

FIG 3 Alpha and beta diversity indices for each cohort. Within-sample diversity was analyzed using Shannon's diversity indices for gut microbiomes sampled at (A) 1–2 months, (C) 3–6 months, and (E) 9–12 months, and the results were visualized with boxplots, where Caesarean delivery samples (pink) and vaginal delivery samples (light blue) were plotted separately for each cohort. Outliers detected by the plotting software were drawn as circles. Between-sample diversity was analyzed using Bray-Curtis dissimilarity at (B) 1–2 months, (D) 3–6 months, and (F) 6–9 months sampling time points using principal coordinate analysis (PCoA). The samples from each cohort were drawn in a different color within the sampling time points, and the confidence ellipse was drawn using the Pearson correlation coefficient for each cohort.

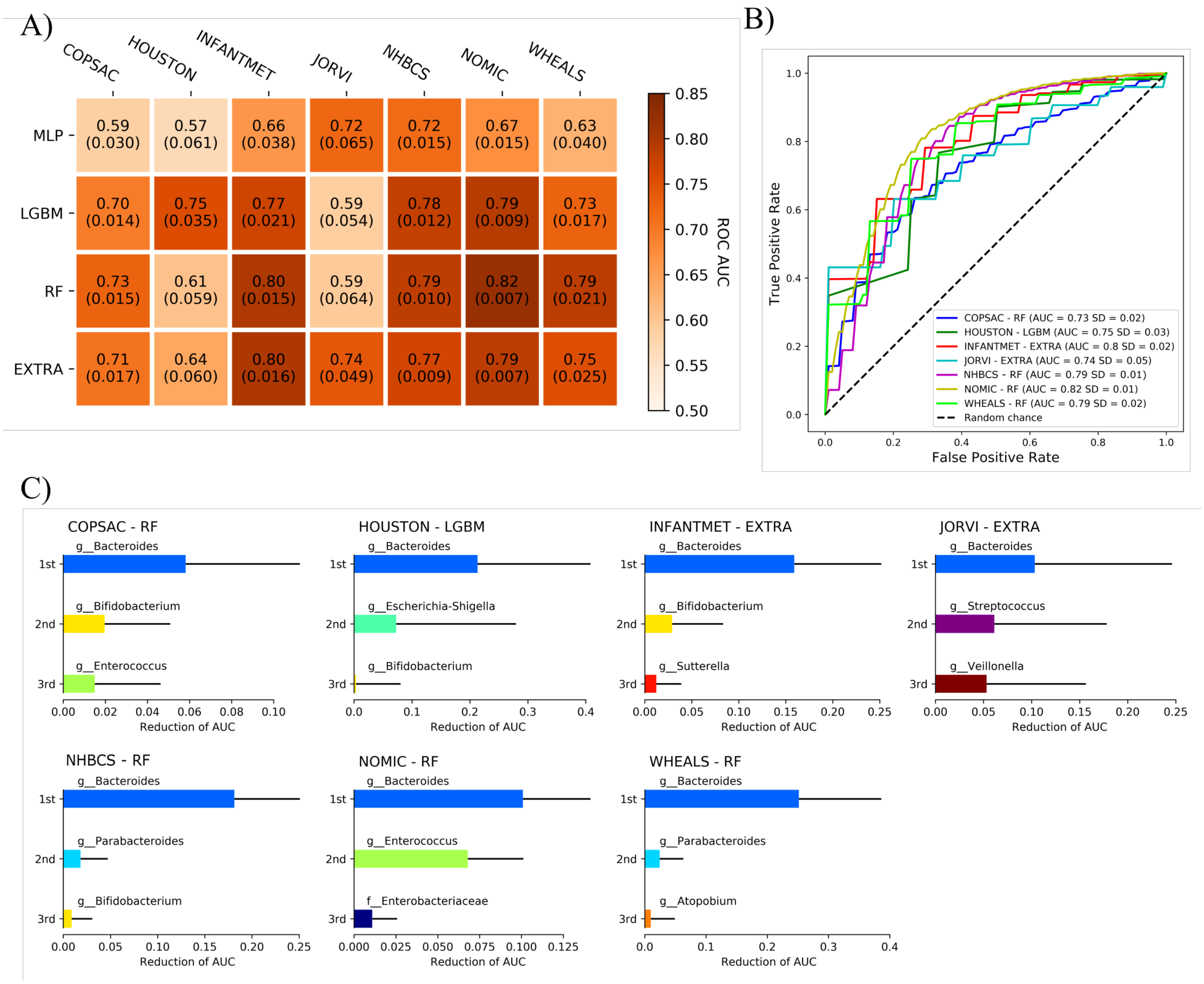

FIG 4 Machine-learning models can differentiate between the delivery modes of infants based on gut microbiome data at 1–2 months of age. The delivery modes were either vaginal delivery or Caesarean delivery, and the ML models used the gut microbiomes of the infants as assessed from fecal samples obtained approximately 1–2 months after birth in seven infant cohorts. (A) Best performances of the MLP, LGBM, RF, and EXTRA models, trained independently to differentiate between vaginal delivery and Caesarean delivery samples using the relative abundances of gut bacteria at 1–2 months after birth, shown separately for each cohort. (B) ROC curves for the best-performing models. The AUC values for these ROC curves indicate model performances that range between 0.5 and 1.0. Predictions from a model with a performance close to 0.5 are equivalent to a random choice, whereas a model with an AUC of 1.0 would hypothetically be a perfect model and classify all children correctly. (C) Permutation importance values for the best-performing models. The *x*-axis of each graph represents the reduction in AUC when the feature was randomized in the testing samples. Positive error bars indicate the standard deviation of the averaged importance values. Each feature is shown in the same color in all bar graphs.

AUC only in the HOUSTON cohort (AUC = 0.75). The MLP models achieved lower AUC values overall than the other models (Fig. 4B). PR curves can be found for all models in Fig. S2.

To control for the potentially confounding effect of breastfeeding, the same analyses were run separately in children receiving breastfeeding (Fig. S3). In COPSAC (AUC = 0.73) and INFANTMET (AUC = 0.75), the prediction performance remained the same or slightly lowered, while in HOUSTON (AUC = 0.62) and JORVI (0.63) cohorts, the prediction performance was much lower. There were only 29 children who were exclusively formula-fed at 1–2 months of age (Table S4), which did not allow separate ML analyses in this subgroup.

Bacteroides was the most important genus for the performance of machine-learning models trained on taxa at 1–2 months of age

Next, we determined which features of the gut microbiome data were most important for the performance of the ML models—i.e., in differentiating between the modes of delivery at 1–2 months of age. We identified the most important features of the ML models using the permutation importance method.

The relative abundance of *Bacteroides* in the gut microbiome had the greatest impact on the prediction performance at 1–2 months of age (Fig. 4C). This may be assessed by evaluating the decrease in the AUC when the *Bacteroides* feature is removed from the model. This reduced the model performance in multiple cohorts: 0.06 AUC in COPSAC, 0.21 in HOUSTON, 0.16 in INFANTMET, 0.10 in JORVI, 0.18 in NHBCS, 0.1 in NOMIC, and 0.25 in WHEALS (Fig. 4C). Other important features for differentiating between the delivery modes based on the gut microbiome data were *Bifidobacterium*, *Enterococcus*, the *Escherichia-Shigella* complex, *Streptococcus*, *Veillonella*, and *Parabacteroides* (Fig. 4C).

ML models were poor at accurately predicting the mode of delivery from microbiome taxonomic data recorded at 3–6 months or 9–12 months

When using gut microbiome taxonomic data from fecal samples taken at 3–6 months and 9–12 months of age, the ML models were not able to differentiate accurately between the modes of delivery of the children in any of the cohorts (Fig. 5; Fig. S4 and S5). The AUC of the best models ranged from 0.61 to 0.62 in four cohorts with gut microbiome data available at 3–6 months of age (Fig. 5A and C). Similarly, models trained using fecal samples collected 9–12 months after birth were unable to differentiate reliably between children born vaginally or via a Caesarean delivery (Fig. 5B and D).

Machine-learning models failed to recognize the delivery mode using predicted metabolic pathway features at any infant age

Next, we used 16S rRNA gene sequences to infer the metabolic pathway composition of each sample with PICRUSt2, which can be used to generate an estimation of metabolic pathway composition based on the 16S rRNA amplicon sequencing data and a reference database. The performance of the ML models trained with predicted metabolic pathways in differentiating Caesarean delivery samples from vaginal delivery samples was comparable to that observed for the genera collapsed models in some cohorts, while in others, the performance values were much lower (Fig. S6). The ML models that used predicted metabolic pathways could not accurately predict the mode of delivery from gut microbiome samples collected 3–6 months or 9–12 months after birth, except in the Oulu (AUC = 0.72, SD = 0.03) and Jorvi (AUC = 0.75, SD = 0.07) cohorts at 9–12 months group (Fig. S7).

The ML models showed several metabolic pathways in the gut microbiome that were important for the performance of the models, such as carbohydrate degradation, nucleotide degradation, and fermentation of the pyruvate metabolic pathways. At 1–2 months of age, the ML models achieved an AUC of 0.7–0.8 for the COPSAC, NHBCS, and NOMIC cohorts, with the carbohydrate degradation pathway (PWY-7456) emerging as the most important performance feature (Fig. S6C).

The ML models for the INFANTMET cohort achieved moderate AUC scores, but unlike the other ML models in the three previously mentioned cohorts, they had pathways related to pyruvate fermentation and amino acid degradation as the top performance features. The best ML models in the HOUSTON, JORVI, and WHEALS cohorts, which all achieved low AUC scores (0.61–0.66), had a variety of pathways as their most important features (Fig. S6C).

The mean relative abundances and standard deviations of metabolic pathways can be found in the Supplementary Table (Table S6).

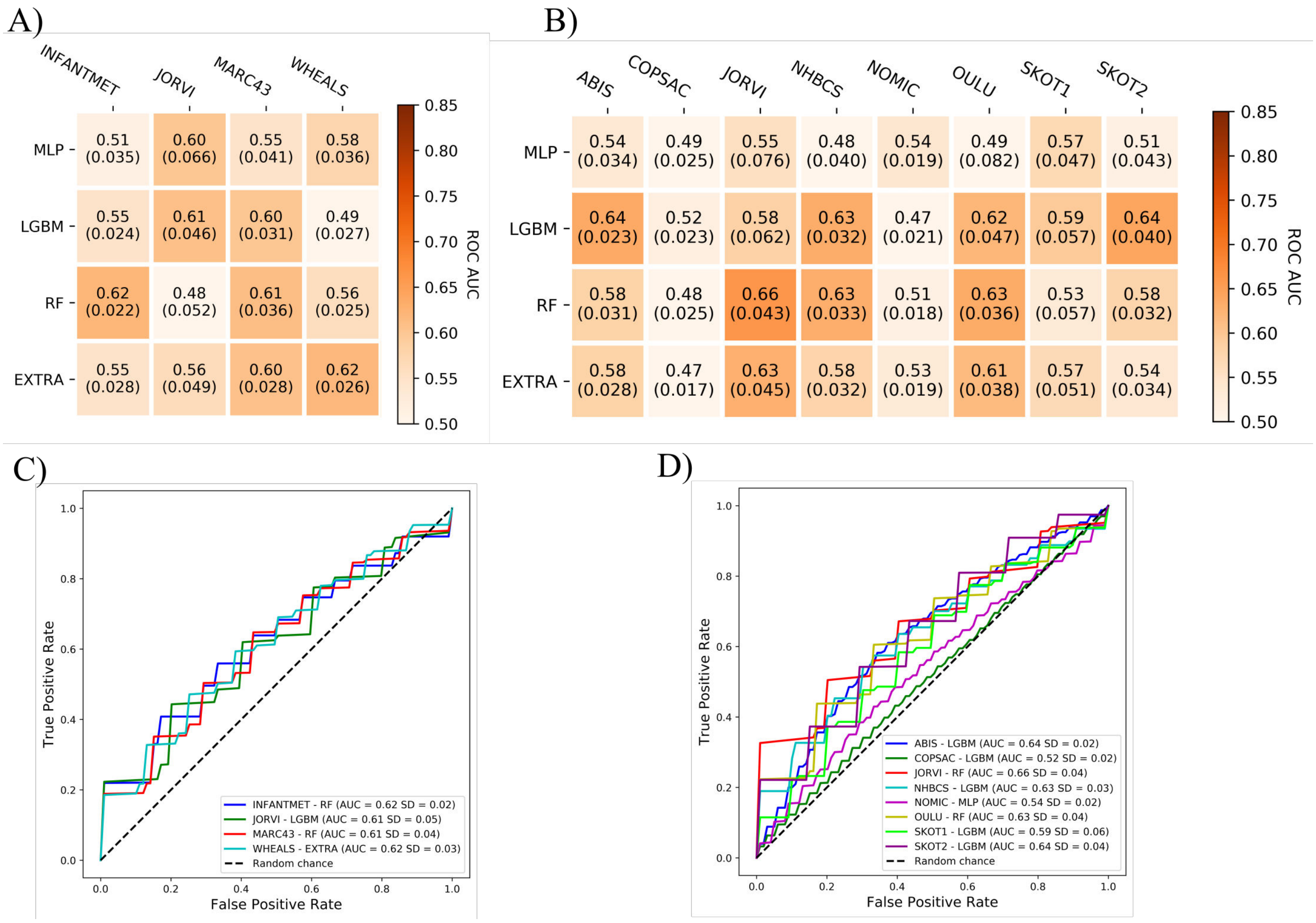

FIG 5 Performance of the ML model when predicting the mode of delivery from samples taken at 3–6 months or 9–12 months after birth. The MLP, LGBM, RF, and EXTRA models were trained independently to differentiate between vaginal delivery and Caesarean delivery samples using the relative abundances of gut bacteria at (A) 3–6 months and (B) 9–12 months after birth. ROC curves for the best-performing models were drawn for (C) 3–6 months and (D) 9–12 months after birth. The AUC values range between 0.5 and 1.0, where predictions from a model with a performance close to 0.5 would be equivalent to a random guess, and those from a model with a performance of 1.0 would always be correct.

Cross-study machine-learning models trained on taxonomy, but not on function, performed well when identifying the delivery mode at 1–2 months of infant age

The ML models trained with all the other cohorts and then tested on the remaining cohort performed well with all cohorts at 1–2 months after birth (Fig. 6), and the test samples from HOUSTON (AUC 0.83, SD 0.05), JORVI (AUC 0.79, SD 0.04), and WHEALS (AUC 0.81, SD 0.02) were predicted more accurately by the cross-study ML models than were those originally trained on the cohort's own training samples (Fig. 6B). The cross-study ML models achieved fairly high accuracy when applied to the COPSAC (AUC 0.72, SD 0.01), INFANTMET (AUC 0.75, SD 0.02), NHBCS (AUC 0.78, SD 0.01), and NOMIC (AUC 0.77, SD 0.01) cohorts (Fig. 6B).

Bacteroides was the most important feature when studying samples from other cohorts at 1–2 months after birth

Next, every feature was removed from the cross-study testing data one at a time by the permutation importance method; meanwhile, the average reduction of prediction performance was recorded for each feature. The most important feature when predicting the mode of delivery using gut microbiome data at 1–2 months of age was *Bacteroides*

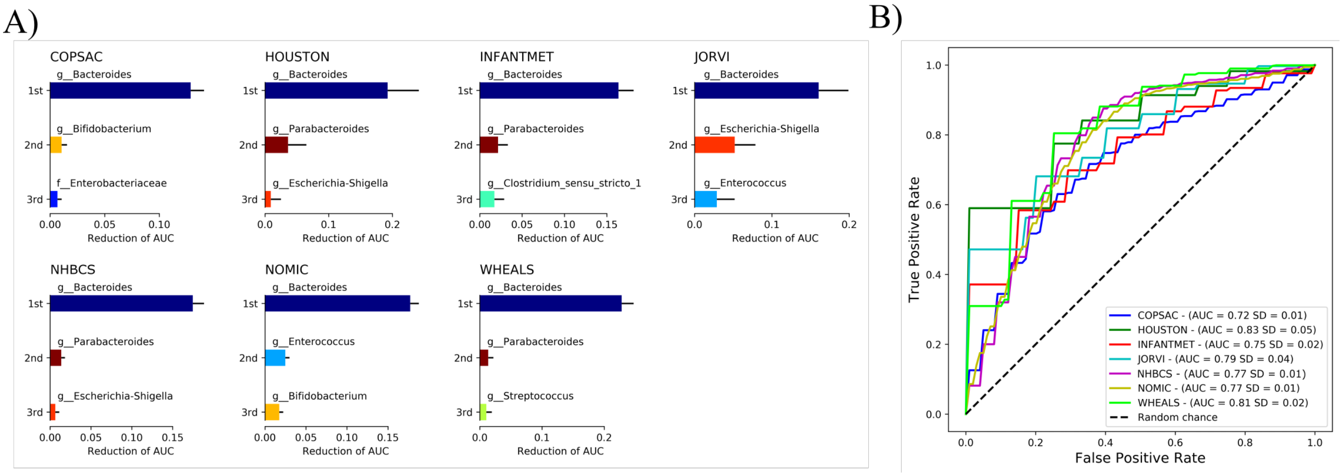

FIG 6 ML models were used to predict the mode of delivery when testing samples from other cohorts in a cross-study manner 1–2 months after birth. The mode of delivery was predicted for the test samples in each cohort by combining the best ML models from each of the other cohorts to form an ensemble classifier. The ML models had no previous knowledge of the other cohorts. (A) The permutation importance of the ensemble classifiers are visualized, with the *x*-axis of the graphs representing the reduction in the AUC when the feature is randomized in the test samples. Positive error bars were plotted to represent the standard deviation of the averaged importance values. Each feature is shown in the same color in all bar graphs. (B) ROC curves were drawn for the ensemble classifiers. The AUC values ranged between 0.5 and 1.0. Predictions from a model with a performance close to 0.5 are equivalent to a random guess, while a model with 1.0 is always correct.

(Fig. 6A), while *Escherichia-Shigella*, *Parabacteroides*, *Bifidobacterium*, and *Enterococcus* had a lesser impact on the performance of the ML model. Removing *Bacteroides* from the testing data reduced the prediction capability of the model by more than half in every cohort.

Bacteroides is relatively enriched in vaginally delivered infants at 1–2 months of age

To investigate why *Bacteroides* was shown as the most important feature in the ML models, we calculated the mean relative abundance of *Bacteroides* in each cohort in both the Caesarean delivery and vaginal delivery groups and plotted them side by side (Fig. 7). *Bacteroides* had a higher mean relative abundance in children born via Caesarean delivery in all cohorts at 1–2 months of age (Fig. 7A; Table S6), and similarly in all samples collected at roughly 3–6 months (Fig. 7B), while only a few cohorts had a higher mean relative abundance in vaginally delivered infants at 9–12 months after birth (Fig. 7C). The standard deviation for the relative abundance of *Bacteroides* was nevertheless very high at all three time points, indicating that the proportion of this genus differed greatly from one infant to another. The mean relative abundances and standard deviations of all genera shown as important by ML analyses can be found in Supplementary Table (Table S6).

Alternative model building approaches did not improve the cross-study model performance at 1–2 months of age

Since there is no gold standard for building ML models or pre-processing microbiome data, we tested how a few different methods affected the cross-study predictions at 1–2 months. Alternative methods used were PipelineSearch and control augmenting. Using the genera-collapsed feature table to train the models and then combining all the other models into an ensemble voting classifier to predict the samples in one cohort was the strategy that performed best in all the cohorts except INFANTMET (Fig. S8). The Control augmenting (AUC = 0.80, SD = 0.045) and Pathway ensemble (AUC = 0.76, SD = 0.022) methods both achieved a higher AUC than the genera ensemble method (AUC = 0.75, SD

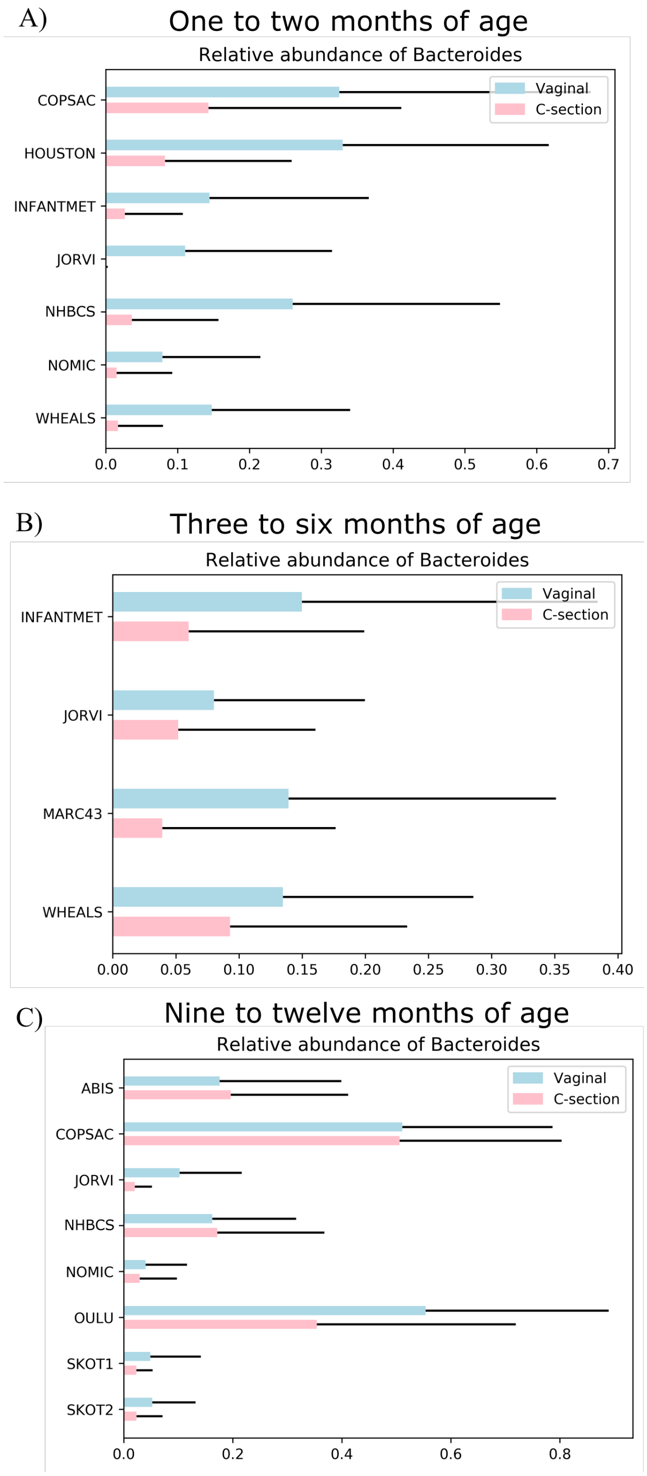

FIG 7 Combined mean relative abundances of sequences classified into the genus *Bacteroides* in each cohort. Mean relative abundance and standard deviation of *Bacteroides* in cohorts sampled approximately (A) 1–2 months, (B) 3–6 months, and (C) 9–12 months after birth, partitioned by mode of delivery. Light blue bars denote vaginally delivered infants, and pink bars denote those born by Caesarean delivery. The black lines are positive error bars (standard deviation).

= 0.018) when predicting delivery in the case of the INFANTMET samples. PipelineSearch did not have the highest AUC when predicting delivery mode in a cross-study way.

DISCUSSION

We experimented with several ML model-building strategies to identify the best approach to combine and compare gut microbiome composition from 12 pediatric cohort studies. We chose mode of delivery as the exposure factor since it is well known that C-section delivery has an effect on infant gut microbiome. We showed that ML models, including MLP, LGBM, RF, and EXTRA models, were able to identify the mode of delivery of infants based on their gut microbiome taxonomic data at 1–2 months of age. *Bacteroides*, enriched in the gut microbiome of infants born by the vaginal route, was the most important feature in the ML models for identifying the mode of delivery. When the infants were older, all ML models performed poorly. Similarly, all ML models performed poorly when trained on predicted microbiome function at all ages.

In the present study, we used ML analysis of gut microbiome composition retrieved from high-quality prospective cohorts from Europe and USA in predicting the mode of delivery. Previously, Le Goallec et al. aggregated 1,570 samples from 300 infants included in 4 European studies to form a single data set from the first 3 years of life (18). Their models were used to predict host characteristics, such as age, sex, country of origin, antibiotic usage, delivery mode, and breastfeeding status. In their study, adding microbiome data to the model increased the prediction performance from an AUC of 0.59 to 0.76 as compared to demographic factors alone in predicting delivery mode (18). Here, we show that based on microbiome taxonomic data alone, cross-study ML models were able to predict delivery mode accurately in infants under 3 months of age with AUC ranging from 0.72 to 0.83 depending on the cohort and the algorithm used. However, when using gut microbiome data from fecal samples taken at 3–6 months and 9–12 months of age, or alternately training on the predicted functional metabolic pathways at any age, the ML models were not able to differentiate accurately between the modes of delivery of the children in any of the cohorts.

There are only a few other earlier studies of cross-study ML in gut microbiome research (17, 20, 54). In a study investigating the role of the gut microbiome in patients with type 2 diabetes, random forest models trained on cross-study data were able to predict type 2 diabetes status (17). In another study examining gut microbiome composition in adult obesity with a cross-study design using 10 data sets, the median accuracy of the ML analyses in distinguishing obesity based on the gut microbiome data was close to that of a random chance classifier (54). In a large study using a cross-disease design, ML models trained to predict one disease lost their accuracy when naively transferred to predict samples from other disease data sets (20). The authors of the study, however, suggested a method called "control augmenting," in which control samples from outside cohorts are added to the training data to increase portability between data sets by the data augmenting method.

The subsequent health of children after Caesarean delivery has been reported in several previous epidemiological observational studies, associating with asthma (1, 2), food allergies (3, 4), obesity (5), and diabetes (6). Furthermore, Caesarean delivery has been associated with alterations in the gut microbiome composition in infants, with the relative abundance of *Bacteroides* (9–14), *Escherichia-Shigella* (9), and *Parabacteroide*s (10) being lower than in vaginal deliveries. Similarly, we found that *Bacteroides* was the most important taxonomic feature when predicting the mode of delivery of infants based on the gut microbiome at 1–2 months in all cohorts studied here. In addition, we found that the mean relative abundance of *Bacteroides* was lower at 1–2 months of age in the Caesarean delivery infants than in those born by vaginal delivery. When using the predicted metabolic pathways of the gut microbiome, the carbohydrate degradation pathways were of greater relative importance in classifying by mode of delivery. However, no pathway could consistently predict mode of delivery at any gestational age which is similar to the results previously reported by Chu et al. (27). Caesarean deliveries, performed for multiple underlying maternal and fetal indications, are associated with varying rates of success at exclusive breastfeeding (55). In the present study, we planned to perform a sensitivity analysis stratified on feeding mode. However, due to low number of infants who received exclusive formula feeding, among those in whom we had information on mode of feeding, we were not able to perform this analysis.

Previous microbiome studies have used various algorithms, such as random forest (17, 18, 20), support vector machine (17, 18), elastic net (17, 18), and gradient boosted machine (18, 56). Decision tree-based algorithms such as random forest and LGBM have consistently been among the top performers in studies employing multiple data sets (17, 18, 20). Additionally, deep learning approaches have shown promise (49), and consequently, we selected three decision tree-based algorithms and one neural network for our analyses. Our results suggest that even relatively similar decision tree algorithms perform differently in each data set, so each algorithm needs to be validated on a data set-by-data set basis. Interestingly, multilayer perceptron performed poorly relative to the decision tree-based algorithms in our study.

Pre-processing choices made before training the ML models affected the downstream analyses. The choice of a taxonomic database, the quality filtering parameters, and collapsing to a specific taxonomic level are all likely to affect the downstream ML analyses. We, therefore, tested four model-building approaches using the same cross-validation folds for each method. As a baseline, we built models on the genera (SILVA [39] database) collapsed feature tables. Second, we built models based on predicted metabolic pathway feature tables, and the third alternative method was to use independent control samples to augment each cohort, as presented in a previous study (20). Lastly, we used a novel method called "PipelineSearch," in which each model could select data from various pre-processing routes—e.g., Greengenes instead of SILVA as a taxonomic database and predicted metabolic pathway features instead of genera-collapsed features. Interestingly, the PipelineSearch method could not achieve the same prediction performance as the baseline genera-collapsed models even though the models could select the same data to be used. This could be explained by the volatility of the microbiome data and the relatively low number of available samples in each cohort. Nevertheless, PipelineSearch is useful in cases where researchers lack the necessary domain knowledge to make optimal pre-processing choices; instead, they can supply the PipelineSearch model with a variety of methods even though some of those pre-processing choices may be suboptimal.

The present study has several strengths. The use of ML in gut microbiome analysis in a cross-study way, although it has been employed in previous studies (17, 20, 54), is still a novel approach. We had gut microbiome data from 4,099 samples representing 12 infant cohorts in their first year of life, and by using ML models, we were able to show predictable differences on the composition of the gut microbiome appears to have certain universal characteristics in cohorts of 1- to 2-month-old infants born by Caesarean delivery across populations. However, this was limited to relative abundance differences in a single taxa, Bacteroides, and was not accompanied by changes in the predicted functional metagenome. Furthermore, the use of active data sharing and collaboration enabled the analysis of a varied collection of data sets spanning Europe and the United States. Finally, we have successfully combined two research fields: clinical medicine and computational biology.

Nevertheless, there are some limitations to our study. To assign taxonomy, we used the SILVA database (version 138). *Bacteroides* has been reclassified as *Phocaeicola* (57); consequently, the genera names shown here may change in the future release of SILVA database. ML analyses do not allow for direct controlling for various factors. We did, however, perform ML analyses separately in subgroups depending on breastfeeding status. We used the PICRUSt2 bioinformatic tool to predict microbial metabolic pathway composition data, which might not correspond to the actual metagenomic data produced by whole-genome sequencing. Furthermore, the study results are generalizable to term infants because we excluded cohorts with mainly preterm infants. Finally, the cohorts recruited for this study were not created solely to study the effects of the mode of delivery. As such, the cohort structures and designs varied from one cohort to another. Furthermore, we were not able to investigate emergency C-section and elective C-section groups separately due to low sample sizes and lack of required data.

Our study provides a new perspective on microbiome research, as it shows that ML enables data analyses in gut microbiome research by comparing and combining data sets from multiple cohorts collected in different countries across diverse patient populations. Furthermore, there is a crucial need to shift the research paradigm from merely retrospective predictions to a more proactive approach, where extensive investigation is directed toward anticipating the health outcomes of infants and children through the analysis of the gut microbiome. This proactive stance could provide a deeper understanding of how the gut microbiome influences the well-being of infants and children and potentially lead to more effective strategies for promoting their optimal health and development.

ACKNOWLEDGMENTS

Financial support was received from the Academy of Finland (T.T.), the Finnish Paediatric Research Foundation (T.T. and N.P.), the University of Oulu Graduate School (P.V.), the Päivikki and Sakari Sohlberg Foundation (P.V. and M.V.T.), the Orion Research Foundation (P.V.), and government funding for clinical research available at Oulu University Hospital, Finland (T.T.). Additional support was provided by NIH-NICHD (R01HD091731) (K.A.), NIH-NIDDK (R01DK089201) (K.A.), NIH UG3/UH3 (OD023253) (K.H.), NIH UG3/UH3 (OD023275) (A.G.H., M.R.K., and J.M.), and NIH-NLM (R01LM012723) (A.G.H.).

All authors declare no conflict of interest relevant to this article.

AUTHOR AFFILIATIONS

[1]Research Unit of Clinical Medicine, University of Oulu, Oulu, Finland
[2]Ecology and Genetics, Faculty of Science, University of Oulu, Oulu, Finland
[3]Department of Pediatrics and Adolescent Medicine, Oulu University Hospital, University of Oulu, Oulu, Finland
[4]Department of Obstetrics & Gynecology, Division of Maternal-Fetal Medicine, Baylor College of Medicine and Texas Children's Hospital, Houston, Texas, USA
[5]Department of Pediatrics, University of California, San Diego, California, USA
[6]Department of Emergency Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[7]Department of Climate and Environmental Health, Norwegian Institute of Public Health, Oslo, Norway
[8]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway
[9]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA
[10]Children's Hospital, University of Helsinki and HUS, Helsinki, Finland
[11]National Food Institute, Technical University of Denmark, Lyngby, Denmark
[12]Crown Princess Victoria Children's Hospital and Division of Pediatrics, Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden
[13]Department of Psychiatry, Dartmouth Hitchcock Medical Center, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA
[14]Department of Pediatrics, Dartmouth Hitchcock Medical Center, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA
[15]Medical College of Georgia, Augusta, Georgia, USA
[16]Teagasc Food Research Centre & APC Microbiome Ireland, Moorepark, Fermoy, Co. Cork, Ireland
[17]Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark
[18]Department of Food Science, University of Copenhagen, Copenhagen, Denmark
[19]Biocenter Oulu, University of Oulu, Oulu, Finland

AUTHOR ORCIDs

Petri Vänni http://orcid.org/0000-0003-0100-2545
Martin F. Laursen http://orcid.org/0000-0001-6017-7121

FUNDING

Funder	Grant(s)	Author(s)
Academy of Finland (AKA)		Terhi Tapiainen
Lastentautien Tutkimussäätiö (Foundation for Pediatric Research)		Niko Paalanne Terhi Tapiainen
University of Oulu Graduate School		Petri Vänni
Päivikki and Sakari Solhberg Foundation		Petri Vänni Mysore V. Tejesvi
Orionin Tutkimussäätiö (Orion Research Foundation)		Petri Vänni
Oulun Yliopistollinen Sairaala (OYS)		Terhi Tapiainen
HHS | NIH | Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)	R01HD091731	Kjersti Aagaard
HHS | NIH | National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)	R01DK089201	Kjersti Aagaard
NIH UG3/UH3	OD023253	Kjersti Aagaard
NIH UG3/UH3	OD023275	Anne G. Hoen Margaret R. Karagas Juliette Madan
HHS | NIH | U.S. National Library of Medicine (NLM)	R01LM012723	Anne G. Hoen

AUTHOR CONTRIBUTIONS

Petri Vänni, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing | Mysore V. Tejesvi, Conceptualization, Data curation, Methodology, Supervision, Writing – original draft, Writing – review and editing | Niko Paalanne, Data curation, Writing – review and editing | Kjersti Aagaard, Data curation, Writing – review and editing | Gail Ackermann, Data curation, Writing – review and editing | Carlos A. Camargo Jr., Data curation, Writing – review and editing | Merete Eggesbø, Data curation, Writing – review and editing | Kohei Hasegawa, Data curation, Writing – review and editing | Anne G. Hoen, Data curation, Writing – review and editing | Margaret R. Karagas, Data curation, Writing – review and editing | Kaija-Leena Kolho, Data curation, Writing – review and editing | Martin F. Laursen, Data curation, Writing – review and editing | Johnny Ludvigsson, Data curation, Writing – review and editing | Juliette Madan, Data curation, Writing – review and editing | Dennis Ownby, Data curation, Writing – review and editing | Catherine Stanton, Data curation, Writing – review and editing | Jakob Stokholm, Data curation, Writing – review and editing | Terhi Tapiainen, Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – original draft, Writing – review and editing

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Data S1 (mSystems00364-23-s0001.xlsx). Genera collapsed relative abundance feature tables of each cohort used in ML analyses.

Data S2 (mSystems00364-23-s0002.xlsx). Predicted metabolic pathway relative abundance feature tables used in ML analyses at 1-2 months.

Data S3 (mSystems00364-23-s0003.xlsx). Predicted metabolic pathway relative abundance feature tables used in ML analyses at 3-6 months.

Data S4 (mSystems00364-23-s0004.xlsx). Predicted metabolic pathway relative abundance feature tables used in ML analyses at 9-12 months.

Supplemental Information (mSystems00364-23-s0005.docx). Figures S1 to S8.

Supplemental Tables (mSystems00364-23-s0006.xlsx). Tables S1 to s6.

REFERENCES

1. Darabi B, Rahmati S, HafeziAhmadi MR, Badfar G, Azami M. 2019. The association between Caesarean section and childhood asthma: an updated systematic review and meta-analysis. Allergy Asthma Clin Immunol 15:62. https://doi.org/10.1186/s13223-019-0367-9
2. Thavagnanam S, Fleming J, Bromley A, Shields MD, Cardwell CR. 2008. A meta-analysis of the association between Caesarean section and childhood asthma. Clin Exp Allergy 38:629–633. https://doi.org/10.1111/j.1365-2222.2007.02780.x
3. Bager P, Wohlfahrt J, Westergaard T. 2008. Caesarean delivery and risk of atopy and allergic disesase: meta-analyses. Clin Exp Allergy 38:634–642. https://doi.org/10.1111/j.1365-2222.2008.02939.x
4. Eggesbø M, Botten G, Stigum H, Nafstad P, Magnus P. 2003. Is delivery by cesarean section a risk factor for food allergy?J Allergy Clin Immunol 112:420–426. https://doi.org/10.1067/mai.2003.1610
5. Darmasseelane K, Hyde MJ, Santhakumaran S, Gale C, Modi N. 2014. Mode of delivery and offspring body mass index, overweight and obesity in adult life: a systematic review and meta-analysis. PLoS One 9:e87896. https://doi.org/10.1371/journal.pone.0087896
6. Cardwell CR, Stene LC, Joner G, Cinek O, Svensson J, Goldacre MJ, Parslow RC, Pozzilli P, Brigis G, Stoyanov D, Urbonaite B, Sipetić S, Schober E, Ionescu-Tirgoviste C, Devoti G, de Beaufort CE, Buschard K, Patterson CC. 2008. Caesarean section is associated with an increased risk of childhood-onset type 1 diabetes mellitus: a meta-analysis of observational studies. Diabetologia 51:726–735. https://doi.org/10.1007/s00125-008-0941-z
7. Samuelsson U, Lindell N, Bladh M, Åkesson K, Carlsson A, Josefsson A. 2015. Caesarean section per se does not increase the risk of offspring developing type 1 diabetes: a Swedish population-based study. Diabetologia 58:2517–2524. https://doi.org/10.1007/s00125-015-3716-3
8. Stinson LF, Payne MS, Keelan JA. 2018. A critical review of the bacterial baptism hypothesis and the impact of cesarean delivery on the infant microbiome. Front Med (Lausanne) 5:135. https://doi.org/10.3389/fmed.2018.00135
9. Azad MB, Konya T, Maughan H, Guttman DS, Field CJ, Chari RS, Sears MR, Becker AB, Scott JA, Kozyrskyj AL, CHILD Study Investigators. 2013. Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. Can Med Assoc J 185:385–394. https://doi.org/10.1503/cmaj.121189
10. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, Khan MT, Zhang J, Li J, Xiao L, Al-Aama J, Zhang D, Lee YS, Kotowska D, Colding C, Tremaroli V, Yin Y, Bergman S, Xu X, Madsen L, Kristiansen K, Dahlgren J, Wang J. 2015. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17:690–703. https://doi.org/10.1016/j.chom.2015.04.004
11. Penders J, Thijs C, Vink C, Stelma FF, Snijders B, Kummeling I, van den Brandt PA, Stobberingh EE. 2006. Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics 118:511–521. https://doi.org/10.1542/peds.2005-2824
12. Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, D Lieber A, Wu F, Perez-Perez GI, Chen Y, Schweizer W, Zheng X, Contreras M, Dominguez-Bello MG, Blaser MJ. 2016. Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci Transl Med 8:343ra82. https://doi.org/10.1126/scitranslmed.aad7121
13. Jakobsson HE, Abrahamsson TR, Jenmalm MC, Harris K, Quince C, Jernberg C, Björkstén B, Engstrand L, Andersson AF. 2014. Decreased gut microbiota diversity, delayed *Bacteroidetes* colonisation and reduced Th1 responses in infants delivered by Caesarean section. Gut 63:559–566. https://doi.org/10.1136/gutjnl-2012-303249
14. Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, Kumar N, Stares MD, Rodger A, Brocklehurst P, Field N, Lawley TD. 2019. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature 574:117–121. https://doi.org/10.1038/s41586-019-1560-1
15. Korpela K. 2021. Impact of delivery mode on infant gut microbiota. Ann Nutr Metab:1–9. https://doi.org/10.1159/000518498
16. Antosca K, Hoen AG, Palys T, Hilliard M, Morrison HG, Coker M, Madan J, Karagas MR. 2020. Reliability of stool microbiome methods for DNA yields and sequencing among infants and young children. Microbiologyopen 9:e1018. https://doi.org/10.1002/mbo3.1018
17. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 12:e1004977. https://doi.org/10.1371/journal.pcbi.1004977
18. Le Goallec A, Tierney BT, Luber JM, Cofer EM, Kostic AD, Patel CJ. 2020. A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. PLoS Comput Biol 16:e1007895. https://doi.org/10.1371/journal.pcbi.1007895
19. Stanislawski MA, Dabelea D, Wagner BD, Iszatt N, Dahl C, Sontag MK, Knight R, Lozupone CA, Eggesbø M. 2018. Gut microbiota in the first 2 years of life and the association with body mass index at age 12 in a Norwegian birth cohort. mBio 9:e01751-18. https://doi.org/10.1128/mBio.01751-18
20. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, Bork P, Sunagawa S, Zeller G. 2021. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol 22:93. https://doi.org/10.1186/s13059-021-02306-1
21. Breiman L. 2001. Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
22. Geurts P, Ernst D, Wehenkel L. 2006. Extremely randomized trees. Mach Learn 63:3–42. https://doi.org/10.1007/s10994-006-6226-1
23. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. 2017. LightGBM: a highly efficient gradient boosting decision tree In Proceedings of the International Conference on Neural Information Processing Systems (NIPS'17), p 3149–3157Curran Associates Inc, Red Hook, NY, USA
24. Wampach L, Heintz-Buschart A, Hogan A, Muller EEL, Narayanasamy S, Laczny CC, Hugerth LW, Bindl L, Bottu J, Andersson AF, de Beaufort C, Wilmes P. 2017. Colonization and succession within the human gut microbiome by archaea, bacteria, and microeukaryotes during the first year of life. Front Microbiol 8:738. https://doi.org/10.3389/fmicb.2017.00738
25. Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, Belzer C, Delgado Palacio S, Arboleya Montes S, Mancabelli L, Lugli GA, Rodriguez

JM, Bode L, de Vos W, Gueimonde M, Margolles A, van Sinderen D, Ventura M. 2017. The first microbial colonizers of the human gut: composition, activities, and health implications of the infant gut microbiota. Microbiol Mol Biol Rev 81:e00036-17. https://doi.org/10.1128/MMBR.00036-17

26. Stokholm J, Blaser MJ, Thorsen J, Rasmussen MA, Waage J, Vinding RK, Schoos A-M, Kunøe A, Fink NR, Chawes BL, Bønnelykke K, Brejnrod AD, Mortensen MS, Al-Soud WA, Sørensen SJ, Bisgaard H. 2018. Maturation of the gut microbiome and risk of asthma in childhood. Nat Commun 9:704. https://doi.org/10.1038/s41467-017-02573-2

27. Chu DM, Ma J, Prince AL, Antony KM, Seferovic MD, Aagaard KM. 2017. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. Nat Med 23:314–326. https://doi.org/10.1038/nm.4272

28. Hill CJ, Lynch DB, Murphy K, Ulaszewska M, Jeffery IB, O'Shea CA, Watkins C, Dempsey E, Mattivi F, Tuohy K, Ross RP, Ryan CA, O'Toole PW, Stanton C. 2017. Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort. Microbiome 5:21. https://doi.org/10.1186/s40168-017-0240-3

29. Jokela R, Korpela K, Jian C, Dikareva E, Nikkonen A, Saisto T, Skogberg K, de Vos WM, Kolho K-L, Salonen A. 2022. Quantitative insights into effects of intrapartum antibiotics and birth mode on infant gut microbiota in relation to well-being during the first year of life. Gut Microbes 14:2095775. https://doi.org/10.1080/19490976.2022.2095775

30. Lundgren SN, Madan JC, Emond JA, Morrison HG, Christensen BC, Karagas MR, Hoen AG. 2018. Maternal diet during pregnancy is related with the infant stool microbiome in a delivery mode-dependent manner. Microbiome 6:109. https://doi.org/10.1186/s40168-018-0490-8

31. Iszatt N, Janssen S, Lenters V, Dahl C, Stigum H, Knight R, Mandal S, Peddada S, González A, Midtvedt T, Eggesbø M. 2019. Environmental toxicants in breast milk of Norwegian mothers and gut bacteria composition and metabolites in their infants at 1 month. Microbiome 7:34. https://doi.org/10.1186/s40168-019-0645-2

32. Levin AM, Sitarik AR, Havstad SL, Fujimura KE, Wegienka G, Cassidy-Bushrow AE, Kim H, Zoratti EM, Lukacs NW, Boushey HA, Ownby DR, Lynch SV, Johnson CC. 2016. Joint effects of pregnancy, sociocultural, and environmental factors on early life gut microbiome structure and diversity. Sci Rep 6:31775. https://doi.org/10.1038/srep31775

33. Robinson A, Fiechtner L, Roche B, Ajami NJ, Petrosino JF, Camargo CA, Taveras EM, Hasegawa K. 2017. Association of maternal gestational weight gain with the infant fecal microbiota. J Pediatr Gastroenterol Nutr 65:509–515. https://doi.org/10.1097/MPG.0000000000001566

34. Russell JT, Roesch LFW, Ördberg M, Ilonen J, Atkinson MA, Schatz DA, Triplett EW, Ludvigsson J. 2019. Genetic risk for autoimmunity is associated with distinct changes in the human gut microbiome. Nat Commun 10:3621. https://doi.org/10.1038/s41467-019-11460-x

35. Tapiainen T, Koivusaari P, Brinkac L, Lorenzi HA, Salo J, Renko M, Pruikkonen H, Pokka T, Li W, Nelson K, Pirttilä AM, Tejesvi MV. 2019. Impact of intrapartum and postnatal antibiotics on the gut microbiome and emergence of antimicrobial resistance in infants. Sci Rep 9:10635. https://doi.org/10.1038/s41598-019-46964-5

36. Laursen MF, Andersen LBB, Michaelsen KF, Mølgaard C, Trolle E, Bahl MI, Licht TR. 2016. Infant gut microbiota development is driven by transition to family foods independent of maternal obesity. mSphere 1:e00069-15. https://doi.org/10.1128/mSphere.00069-15

37. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:1091. https://doi.org/10.1038/s41587-019-0252-6

38. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869

39. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596. https://doi.org/10.1093/nar/gks1219

40. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. 2020. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol 38:685–688. https://doi.org/10.1038/s41587-020-0548-6

41. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072. https://doi.org/10.1128/AEM.03006-05

42. Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55

43. Cawley GC, Talbot NLC. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107.

44. Zhou Y-H, Gallins P. 2019. A review and tutorial of machine learning methods for microbiome host trait prediction. Front Genet 10:579. https://doi.org/10.3389/fgene.2019.00579

45. Pang H, Jung S-H. 2013. Sample size considerations of prediction-validation methods in high-dimensional data for survival outcomes. Genet Epidemiol 37:276–282. https://doi.org/10.1002/gepi.21721

46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. SciKit-learn: machine learning in python. J Mach Learn Res 12:2825–2830.

47. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, Berland M, Gruca A, Hasic J, Hron K, Klammsteiner T, Kolev M, Lahti L, Lopes MB, Moreno V, Naskinova I, Org E, Paciência I, Papoutsoglou G, Shigdel R, Stres B, Vilne B, Yousef M, Zdravevski E, Tsamardinos I, Carrillo de Santa Pau E, Claesson MJ, Moreno-Indias I, Truu J. 2021. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Front Microbiol 12:634511. https://doi.org/10.3389/fmicb.2021.634511

48. Bokulich NA, Dillon MR, Bolyen E, Kaehler BD, Huttley GA, Caporaso JG. 2018. q2-sample-classifier: machine-learning tools for microbiome classification and regression. J Open Res Softw 3:934. https://doi.org/10.21105/joss.00934

49. Oh M, Zhang L. 2020. DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep 10:6026. https://doi.org/10.1038/s41598-020-63159-5

50. Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305.

51. Marizzoni M, Gurry T, Provasi S, Greub G, Lopizzo N, Ribaldi F, Festari C, Mazzelli M, Mombelli E, Salvatore M, Mirabelli P, Franzese M, Soricelli A, Frisoni GB, Cattaneo A. 2020. Comparison of bioinformatics pipelines and operating systems for the analyses of 16S rRNA gene amplicon sequences in human fecal samples. Front Microbiol 11:1262. https://doi.org/10.3389/fmicb.2020.01262

52. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. Nat Rev Microbiol 16:410–422. https://doi.org/10.1038/s41579-018-0029-9

53. Mirzayi C, Renson A, Zohra F, Elsafoury S, Geistlinger L, Kasselman LJ, Eckenrode K, van de Wijgert J, Loughman A, Marques FZ, MacIntyre DA, Arumugam M, Azhar R, Beghini F, Bergstrom K, Bhatt A, Bisanz JE, Braun J, Bravo HC, Buck GA, Bushman F, Casero D, Clarke G, Collado MC, Cotter PD, Cryan JF, Demmer RT, Devkota S, Elinav E, Escobar JS, Fettweis J, Finn RD, Fodor AA, Forslund S, Franke A, Furlanello C, Gilbert J, Grice E, Haibe-Kains B, Handley S, Herd P, Holmes S, Jacobs JP, Karstens L, Knight R,

Knights D, Koren O, Kwon DS, Langille M, Lindsay B, McGovern D, McHardy AC, McWeeney S, Mueller NT, Nezi L, Olm M, Palm N, Pasolli E, Raes J, Redinbo MR, Rühlemann M, Balfour Sartor R, Schloss PD, Schriml L, Segal E, Shardell M, Sharpton T, Smirnova E, Sokol H, Sonnenburg JL, Srinivasan S, Thingholm LB, Turnbaugh PJ, Upadhyay V, Walls RL, Wilmes P, Yamada T, Zeller G, Zhang M, Zhao N, Zhao L, Bao W, Culhane A, Devanarayan V, Dopazo J, Fan X, Fischer M, Jones W, Kusko R, Mason CE, Mercer TR, Sansone S-A, Scherer A, Shi L, Thakkar S, Tong W, Wolfinger R, Hunter C, Segata N, Huttenhower C, Dowd JB, Jones HE, Waldron L, Genomic Standards Consortium, Massive Analysis and Quality Control Society. 2021. Reporting guidelines for human microbiome research: the STORMS checklist. Nat Med 27:1885–1892. https://doi.org/10.1038/s41591-021-01552-x

54. Sze MA, Schloss PD. 2016. Looking for a signal in the noise: revisiting obesity and the microbiome. mBio 7:e01018-16. https://doi.org/10.1128/mBio.01018-16

55. Sassin AM, Johnson GJ, Goulding AN, Aagaard KM. 2022. Crucial nuances in understanding (mis)associations between the neonatal microbiome and Cesarean delivery. Trends Mol Med 28:806–822. https://doi.org/10.1016/j.molmed.2022.07.005

56. Gou W, Ling C-W, He Y, Jiang Z, Fu Y, Xu F, Miao Z, Sun T-Y, Lin J-S, Zhu H-L, Zhou H, Chen Y-M, Zheng J-S. 2021. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. Diabetes Care 44:358–366. https://doi.org/10.2337/dc20-1536

57. García-López M, Meier-Kolthoff JP, Tindall BJ, Gronow S, Woyke T, Kyrpides NC, Hahnke RL, Göker M. 2019. Analysis of 1,000 type-strain genomes improves taxonomic classification of *Bacteroidetes*. Front Microbiol 10:2083. https://doi.org/10.3389/fmicb.2019.02083