frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 02 August 2022
DOI 10.3389/fmicb.2022.912853

Predicting preterm birth through vaginal microbiota, cervical length, and WBC using a machine learning model

Sunwha Park[1†], Jeongsup Moon[2†], Nayeon Kang[2], Young-Han Kim[3], Young-Ah You[1], Eunjin Kwon[1], AbuZar Ansari[1], Young Min Hur[1], Taesung Park[2,4]* and Young Ju Kim[1]*

[1]Department of Obstetrics and Gynecology, College of Medicine, Ewha Medical Research Institute, Ewha Womans University, Seoul, South Korea, [2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea, [3]Department of Obstetrics and Gynecology, College of Medicine, Yonsei University, Seoul, South Korea, [4]Department of Statistics, Seoul National University, Seoul, South Korea

An association between the vaginal microbiome and preterm birth has been reported. However, in practice, it is difficult to predict premature birth using the microbiome because the vaginal microbial community varies highly among samples depending on the individual, and the prediction rate is very low. The purpose of this study was to select markers that improve predictive power through machine learning among various vaginal microbiota and develop a prediction algorithm with better predictive power that combines clinical information. As a multicenter case–control study with 150 Korean pregnant women with 54 preterm delivery group and 96 full-term delivery group, cervicovaginal fluid was collected from pregnant women during mid-pregnancy. Their demographic profiles (age, BMI, education level, and PTB history), white blood cell count, and cervical length were recorded, and the microbiome profiles of the cervicovaginal fluid were analyzed. The subjects were randomly divided into a training (n=101) and a test set (n=49) in a two-to-one ratio. When training ML models using selected markers, five-fold cross-validation was performed on the training set. A univariate analysis was performed to select markers using seven statistical tests, including the Wilcoxon rank-sum test. Using the selected markers, including *Lactobacillus* spp., *Gardnerella vaginalis*, *Ureaplasma parvum*, *Atopobium vaginae*, *Prevotella timonensis*, and *Peptoniphilus grossensis*, machine learning models (logistic regression, random forest, extreme gradient boosting, support vector machine, and GUIDE) were used to build prediction models. The test area under the curve of the logistic regression model was 0.72 when it was trained with the 17 selected markers. When analyzed by combining white blood cell count and cervical length with the seven vaginal microbiome markers, the random forest model showed the highest test area under the curve of 0.84. The GUIDE, the single tree model, provided a more reasonable biological interpretation, using the 10 selected markers (*A. vaginae*, *G. vaginalis*, *Lactobacillus crispatus*, *Lactobacillus fornicalis*, *Lactobacillus gasseri*, *Lactobacillus iners*, *Lactobacillus jensenii*, *Peptoniphilus grossensis*, *P. timonensis*, and *U. parvum*), and the covariates

Check for updates

OPEN ACCESS

EDITED BY
Hyun-Seob Song,
University of Nebraska-Lincoln,
United States

REVIEWED BY
Baiba Vilne,
Riga Stradiņš University,
Latvia
Ryszard Kubinski,
Phyla Technologies,
Canada

*CORRESPONDENCE
Taesung Park
tspark@stats.snu.ac.kr
Young Ju Kim
kkyj@ewha.ac.kr

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to
Systems Microbiology,
a section of the journal
Frontiers in Microbiology

RECEIVED 07 April 2022
ACCEPTED 05 July 2022
PUBLISHED 02 August 2022

CITATION
Park S, Moon J, Kang N, Kim Y-H, You Y-A, Kwon E, Ansari A, Hur YM, Park T and Kim YJ (2022) Predicting preterm birth through vaginal microbiota, cervical length, and WBC using a machine learning model. *Front. Microbiol.* 13:912853.
doi: 10.3389/fmicb.2022.912853

produced a tree with a test area under the curve of 0.77. It was confirmed that the association with preterm birth increased when *P. timonensis* and *U. parvum* increased (AUC=0.77), which could also be explained by the fact that as the number of *Peptoniphilus lacrimalis* increased, the association with preterm birth was high (AUC=0.77). Our study demonstrates that several candidate bacteria could be used as potential predictors for preterm birth, and that the predictive rate can be increased through a machine learning model employing a combination of cervical length and white blood cell count information.

KEYWORDS

preterm birth, vaginal microbiome, pregnancy, 16s ribosomal RNA metagenome sequencing, cervicovaginal fluid, machine learning, microbial-marker

Introduction

Preterm birth (PTB) is defined as delivery at less than 37 weeks of gestation, and prematurity from PTB is a major cause of morbidity and mortality among infants (Goldenberg et al., 2008). The risk factors for PTB are influenced by ethnicity, low socioeconomic status, maternal weight, smoking, periodontal status, and underlying diseases (Koullali et al., 2016). Due to the increase in elderly pregnant women and pregnant women with various underlying diseases, PTB is increasing, and efforts are being made to predict and prevent it (Newnham et al., 2017; Ananth et al., 2018). Among PTB, spontaneous PTB accounts for 70–75% of all cases, and one-third of them are caused by intra-amniotic infection, an infection of the tissues surrounding the fetus (Goldenberg et al., 2008; Chan, 2014). Microorganisms that cause these intra-amniotic infections (*Ureaplasma* spp., *Gardnerella vaginalis*) show similar patterns to those of the lower genital tract, and are known to induce uterine contractions and premature rupture of membranes due to an inflammatory response caused by the ascending infection (Romero et al., 2014; Bennett et al., 2020; Park et al., 2020). Therefore, methods to evaluate the risk of PTB by microscopy, culture, and polymerase chain reaction (PCR) are being carried out in clinical practice. Furthermore, with the development of 16s rRNA metagenome sequencing, it has become possible to analyze not only pathogens but also the microbial community, that is, the microbiome (Yoo et al., 2016; Hur et al., 2021).

In pregnant women, an increase in *Lactobacillus* is known to be associated with term birth (TB), whereas an increase in *G. vaginalis*, *Ureaplasma* spp., *Prevotella* spp., *Atopobium vaginae, Peptoniphilus* spp., *Staphylococcus aureus, Streptococcus* spp., and *Bacteroides* spp. are known to increase PTB (Fettweis et al., 2019). Vaginal dysbiosis, a state of imbalance in the microbial community in the vagina, is related to PTB (Fettweis et al., 2019; Bennett et al., 2020; Kumar et al., 2021). However, the results of microbiome analysis using 16s rRNA metagenome sequencing are difficult to interpret, and since there are many individual differences, it is very difficult to predict PTB using this method.

Many researchers have created prediction models using logistic regression (LR) with PTB-associated clinical information and microbiome data (Hyman et al., 2014; Kumar et al., 2021). Various other machine-learning methods have been applied to classify PTB, such as the random forest (RF) and support vector machine (SVM; Della Rosa et al., 2021; Urushiyama et al., 2021). However, PTB prediction modeling techniques that use intersect markers from several metagenomic analyses have not been studied, and research is lacking on how reproducible the developed models are when applied in practice. Therefore, in this study, we aimed to select markers by analyzing vaginal microbiome data from pregnant women and develop a model with a high predictive rate by combining clinical information.

Materials and methods

Study subjects and CVF collection

In this case–control study, subjects were recruited from Yonsei University Severance Hospital and Ewha Womans University Mokdong Hospital between 2018 and 2020. This study was approved by the Ethical Research Committees of Yonsei University Severance Hospital (no. 4-2018-0564) and Ewha Womans University Mokdong Hospital (no. 2018-07-007). All the participants provided written informed consent. The subjects included singleton pregnant women with a gestational age between 17 and 32 weeks. CVF samples were collected from the posterior vaginal fornix using sterile cotton before vaginal examination or clinical treatment, including antibiotics, steroids, and progesterone. For all study subjects, baseline demographic information and health-related characteristics including age, pre-pregnancy body mass index, education level, and maternal PTB history were collected. At the time of CVF sample collection, cervical lengths (CL) were measured, and the white blood cell (WBC) count of the blood test was recorded. After delivery, delivery mode, gestational age at birth (GAB), birth weight of newborn, appearance, pulse, grimace, activity, and respiration

(APGAR) scores were evaluated. Subjects diagnosed with gestational diabetes mellitus, preeclampsia, or with insufficient medical records were excluded.

Metagenome analysis using 16s rRNA gene sequencing

Amplification of the V3-4 region of 16S-rRNA gene sequencing for identification of the taxonomy

For microbiome analysis, the collected CVF samples were subjected to bacterial DNA extraction using the NucleoSpin Tissue Kit (Macherey–Nagel, Düren, Germany), following the manufacturer's instructions. Sequencing of 16S rRNA was performed according to the 16S metagenomic sequencing library preparation protocol, targeting the V3 and V4 hypervariable regions. For PCR and purification of the PCR product, the KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, United States) and Agencourt AMPure XP system (Beckman Coulter Genomics, Brea, United States) were used. The initial PCR was performed with 12 ng template DNA using region-specific primers (Supplementary Table 1). After magnetic bead-based purification, a second PCR was performed using primers from the Nextera XT Index Kit (Illumina). Purified PCR products were visualized by gel electrophoresis and quantified using DropSense96 (Trinean, Gentbrugge, Belgium). For quality analysis, the pooled samples were run on an Agilent 2,100 Bioanalyzer (Agilent, Santa Clara, CA, United States). Using the CFX96 Real-Time System, Libraries were quantified by qPCR. After normalization, sequencing of the prepared library was conducted using the MiSeq system (Illumina, San Diego, CA, United States) with 300 bp paired-end reads.

Bioinformatics analysis and marker selection

Microbiome sequence and composition analysis

Generated paired-end reads were analyzed using DADA2 pipeline (version 1.19.1) to build an amplicon sequence variant (ASV) table (Callahan et al., 2016). Primers which were truncated and the reads with ambiguous bases or more than two expected errors were dropped. The forward and reverse reads were trimmed to 285 and 225, respectively, ensuring a 20 bp overlapping region for the merging step. Taxonomies are assigned to ASV using exact string matching against EzBioCloud 16S database (Yoon et al., 2017). Then, unassigned ASVs were taxonomically identified using NCBI Blast search with 99% sequence similarity. Lastly, ASVs with unidentified taxonomy and low prevalence (<0.005%) were filtered out. Using the Shannon index, the α-diversity was computed to understand the richness and diversity of the microbiome species in the TB and PTB groups (Shannon, 1997). The α-diversity was compared using the Wilcoxon rank-sum test between the two groups. Furthermore, the β-diversity using the Bray-Curtis distance was examined to compare the divergence in the microbiome community between the two groups (Bray and Curtis, 1957).

Marker selection

Before developing the prediction models, marker selection was performed to reduce the size of marker set. Samples were split into a training set and a test set in a two-to-one ratio. On the training set, the following seven statistical methods for marker selection were applied: zero-inflated Gaussian mixture model (ZIG), zero-inflated beta regression (ZIBSeq), analysis of microbiome composition (ANCOM), centered log-ratio transformation, and permutation logistic regression model (CLR Permutation), Wilcoxon rank-sum test (Wilcoxon), DESeq2, and edgeR. First, the markers with frequency less than 25% and mean proportion less than 0.001% were filtered out. Then, the markers whose *p-values* were less than 0.05 were selected.

Among these selected 15 markers, we further investigate whether or not these markers are preterm and genital infection based on the literature searches (Supplementary Table 2). Finally, 10 markers were selected as a marker set. In addition, we also considered the following seven markers that were detected by at least two statistical analyses with its frequencies between 10 and 25%: *Bifidobacterium breve*, *Dialister propionicifaciens*, *Lactobacillus paracasei*, *Mobiluncus curtisii*, *Prevotella disiens*, *S. aureus*, and *Streptococcus anginosus*. Hence, two sets of markers, containing 10 species and 17 species, and the entire marker set were used for the multiple marker selection step (Figure 1).

We performed multiple marker selection in two ways: one from pre-selected sets and the other from the whole marker set. The pre-selected sets were the two sets of statistically significant markers from single marker selection, with or without already reported PTB-related markers. First, for the two pre-selected feature sets, two different feature selection methods were used: exhaustive search and forward selection. We applied five-fold cross-validation (CV) to the training set (Kim et al., 2021). Second, we also performed feature selection from the whole marker set. However, we excluded exhaustive search since the computational cost of exhaustive search increased exponentially on the whole marker set. Instead, we applied stepwise selection and lasso penalization along with forward selection (Tibshirani, 1996; Kim et al., 2019). The detailed multiple marker selection methods are described in Figure 1; Supplementary Figure 1.

Prediction model development

LR, RF, XGB, SVM, and GUIDE (version 38.0) were used to develop prediction models (Loh, 2009; Chen and Guestrin, 2016). Hyperparameters of RF, XGB, and SVM were tuned from the training set using five-fold cross-validation. Training set was randomly divided into five separate sets. By using each one as validation set, we trained the model with four other sets and calculated model performance on each of the validation set. The hyperparameters with the greatest mean validation AUCs were chosen. The test set was only used in evaluating the

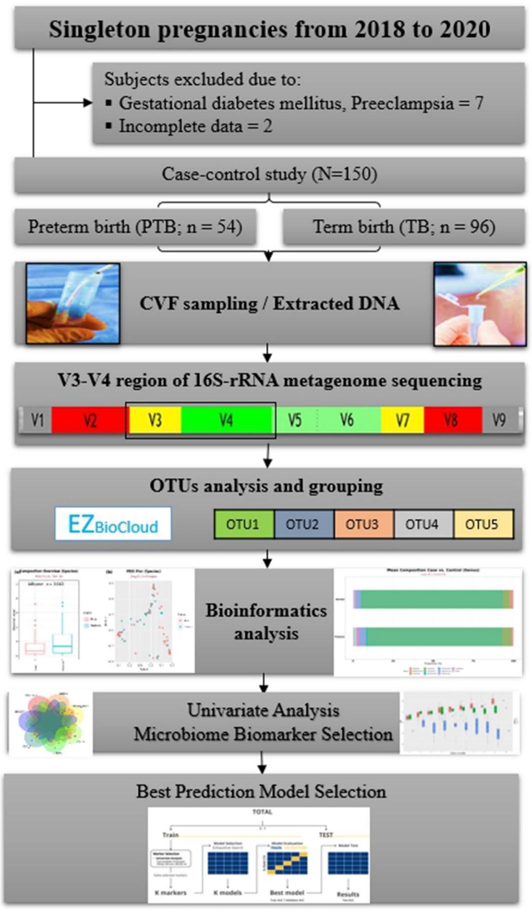
Singleton pregnancies from 2018 to 2020
Subjects excluded due to:
▪ Gestational diabetes mellitus, Preeclampsia = 7
▪ Incomplete data = 2
Case-control study (N=150)
Preterm birth (PTB; n = 54)
Term birth (TB; n = 96)
CVF sampling / Extracted DNA
V3-V4 region of 16S-rRNA metagenome sequencing
V1
V2
V3
V4
V5
V6
V7
V8
V9
OTUs analysis and grouping
EZBioCloud
OTU1
OTU2
OTU3
OTU4
OTU5
Bioinformatics analysis
Univariate Analysis
Microbiome Biomarker Selection
Best Prediction Model Selection
TOTAL
Train
TEST
K markers
K models
Best model
Results

FIGURE 1
Flowchart of the study. CVF, cervicovaginal fluid; rRNA, ribosomal ribonucleic acid; OTUs, operational taxonomic units.

model performances. GUIDE, a single-decision tree-based method, reduced the variable selection bias by choosing significant variables from Chi-square tests (Loh, 2011). The selected split point minimized the node impurity measure. The final tree was pruned using five-fold CV to minimize the misclassification cost. The performance of all the models was measured using the AUC. Then, using the test set, the AUCs of each model were compared to identify a better-performing model.

Results

Clinical characteristics

A total of 150 women participated in this case–control study: 54 in the PTB group and 96 in the TB group (Figure 2). There were no significant differences between the characteristics of the PTB and TB groups, except for the history of sPTB, WBC count, CL, GAB, birth weight, and APGAR score (Table 1).

Association between bacteria and preterm birth

Differences in microbial diversity between PTB and TB groups

A total of 365 bacteria were detected at the species level. In the diversity analysis of the microbial community, the α-diversity using the Shannon index, an indicator of species diversity, was significantly higher in PTB (Figure 3A). However, there was no significant difference in β-diversity (Figure 3B).

Marker selection using univariate analysis

The samples were randomly split into a training set ($n = 101$) and test set ($n = 49$; Supplementary Table 3). Seven different metagenomic analyses were performed on the training set to identify differentially distributed species between the PTB and TB groups. When the marker selection step with mean proportion and frequency was applied, there were 15 markers that showed significance in more than two statistical tests (Figure 1). Additional filtering and literature search were applied to select the following ten species: *Lactobacillus crispatus*, *Lactobacillus fornicalis*, *Lactobacillus gasseri*, *Lactobacillus iners*, *Lactobacillus jensenii*, *G. vaginalis*, *Ureaplasma parvum*, *A. vaginae*, *Prevotella timonensis*, and *Peptoniphilus grossensis* (Supplementary Tables 2, 4). Moreover, seven additional species that could be associated with PTB were appended: *B. breve, D. propionicifaciens, L. paracasei, M. curtisii, P. disiens, S. aureus, and S. anginosus* (Fettweis et al., 2019; Dunlop et al., 2021).

Multiple marker selection

LR, exhaustive search, and forward selection were independently applied to the 10 and 17 pre-selected markers to identify the best marker sets. Forward selection, stepwise selection, and LASSO were applied to 365 markers. In addition to these markers, WBC count was included as a covariate. With multiple marker selection, the minimum number of selected markers was one, and the maximum number of selected markers was 49 (Supplementary Table 5). Because the selection result may depend on how the training and test datasets are split, we repeated the entire splitting process 100 times independently. Marker selection was consistent without showing any outlying results (Supplementary Figures 2, 3).

Further analysis was performed on participants with CL information: 67 with TBs and 42 with PTBs. As in the previous analysis, the remaining samples were randomly divided into a training set and a test set with a two-to-one ratio, and multiple marker selections were applied on 10, 17, and the whole marker set. The minimum number of selected variables was 5 and the maximum number of selected variables was 19 (Table 2). We independently repeated the entire splitting process 100 times. Marker selection was consistent without showing any outlying results (Supplementary Figures 4, 5).

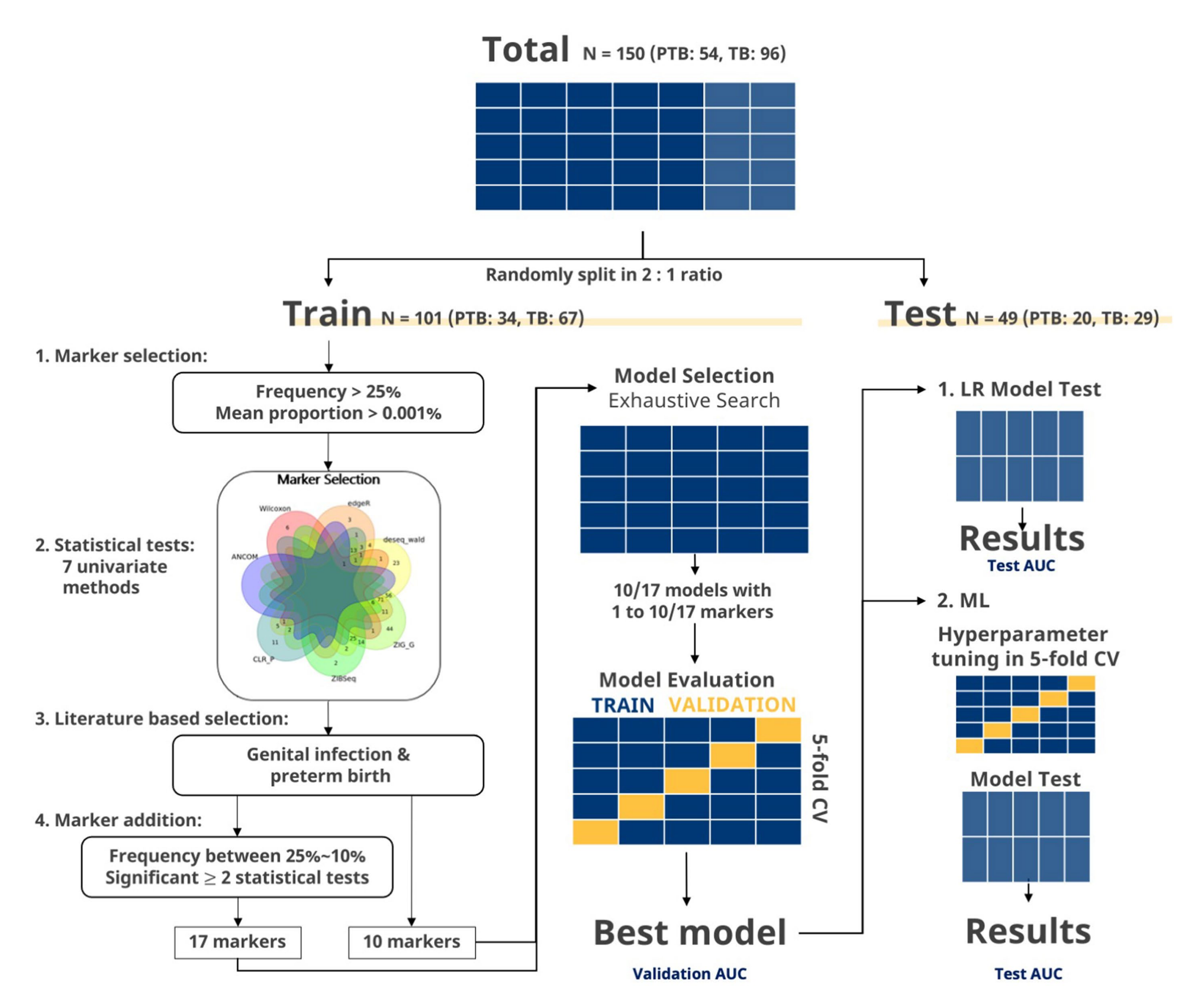
Total N = 150 (PTB: 54, TB: 96)
Randomly split in 2 : 1 ratio
Train N = 101 (PTB: 34, TB: 67)
Test N = 49 (PTB: 20, TB: 29)
1. Marker selection:
Frequency > 25%
Mean proportion > 0.001%
2. Statistical tests: 7 univariate methods
Marker Selection
Wilcoxon
edgeR
deseq_wald
ANCOM
CLR_P
ZIBSeq
ZIG_G
3. Literature based selection:
Genital infection & preterm birth
4. Marker addition:
Frequency between 25%~10%
Significant ≥ 2 statistical tests
17 markers
10 markers
Model Selection
Exhaustive Search
10/17 models with 1 to 10/17 markers
Model Evaluation
TRAIN VALIDATION
5-fold CV
Best model
Validation AUC
1. LR Model Test
Results
Test AUC
2. ML
Hyperparameter tuning in 5-fold CV
Model Test
Results
Test AUC

FIGURE 2
Flowchart of marker selection and evaluation in exhaustive search. The data were split to a training set and test set in a two-to-one ratio. Markers with frequency more than 25% and mean proportion more than 0.001% were selected. Then, markers, showing significant *p* values in two or more statistical tests, were selected. Venn diagram of significant markers (p <0.05) after seven statistical methods (ZIG, ZIBSeq, ANCOM, CLR permutation, Wilcoxon rank-sum test, DESeq2, and edgeR) is shown. Additional filtering steps were applied to the selected markers to finalize the set of 10 and 17 markers. For the given set of markers, exhaustive search was applied to every possible combination of markers using LR. Best marker sets for each number of combinations were selected using AUC from the training set. The global best marker set among these selected sets was chosen as the one that showed the highest AUC from the five-fold CV. Then, the final marker set was select based on the test set. Lastly, the final marker sets were used in building machine learning (ML) models.

Prediction model using machine learning algorithms

Using 18 differently selected marker sets, the PTB prediction models were trained based on the following five machine-learning methods. When trained without the covariates, the SVM model using the six markers selected from the 17 pre-selected markers showed the highest test AUC 0.70 (Supplementary Table 5). The RF model using the 17 preselected markers showed a similar AUC of 0.75. *Lactobacillus* spp. and *U. parvum* were reported to be important features for predicting PTB in the RF model (Figures 4A,B; Kataoka et al., 2006; Petricevic et al., 2014; Tabatabaei et al., 2019). When WBC was added as a covariate, most machine-learning methods showed improved prediction performance (Supplementary Table 5).

When PTB prediction models were trained using subjects with CL, those with both WBC and CL generally showed higher test AUCs than those without covariates (Table 2). With the increase of test AUCs, other metrics, such as f1-score and MCC, also increased when the covariates were added to the models (Supplementary Table 6). The RF model using the seven forward selected markers from the total markers showed the highest AUC of 0.84. This model showed a sensitivity of 0.79 when the specificity was 0.83 (Figures 4C,D). In addition, the model's precision and recall were 0.77 and 0.71, respectively. These precision and recall produced a high f1-score of 0.74, which was

TABLE 1 Clinical characteristics of the study subjects.

Characteristics	Preterm birth (*n* = 56)	Term birth (*n* = 99)	*P*-value
Maternal age (year)	32.5 (±3.8)	33.0 (±4.0)	0.427
Pre-pregnancy BMI (kg/m^2)	21.4 (±3.2)	21.4 (±2.7)	0.938
Education level			>0.999
High school graduation or below	4 (16.0)	11 (15.3)	
University graduates	21 (84.0)	61 (84.7)	
History of PTB			<0.002*
No	42 (85.7)	93 (98.9)	
Yes	7 (14.3)	1 (1.1)	
WBC (1 × 10^3/μl)	11.20 (8.8–13.2)	9.30 (8.0–10.5)	<0.001*
GAS (wks)	26.8 (22.8–30.4)	25.8 (22.1–30.5)	0.262
Cervical lengths (mm)	22.7 (13.6–31.9)	30.4 (26.6–36.0)	<0.001*
CST type			0.106
I, II, V	18 (36.0)	47 (54.1)	
III	10 (20.0)	19 (21.8)	
IV	22 (44.0)	21 (24.1)	
GAB (wks)	30.6 (27.5–34.1)	38.9 (38.1–39.6)	<0.001*
Delivery mode			0.055
ND	25 (44.6)	60 (60.6)	
CS	31 (55.4)	39 (39.4)	
Birth Weight (g)	1738.6 (±885.7)	3234.9 (±323.0)	<0.001*
APGAR score at 1 min	6.23 (3–9)	9.35 (9–10)	<0.001*
APGAR score at 5 min	7.55 (6–10)	9.76 (10–10)	<0.001*

Categorical variables were expressed as frequencies (percentages) and analyzed using the Chi-square test and Fisher's exact test. Continuous variables were expressed as mean ± standard deviation (SD) or median (interquartile range) and were compared using the *t*-test or Mann–Whitney *U* test. BMI, body mass index; PTB, preterm birth; WBC, white blood cell; GAS, gestational age at sampling; CST, community-state type; ND, normal delivery; CS, cesarean section; GAB, gestational age at birth; APGAR, appearance, pulse, grimace, activity, respiration; NICU, neonatal intensive care unit. *p < 0.05, considered statistically significant.

the harmonic mean of them. Lastly, the model's MCC value of 0.59 indicated that there was a positive correlation between model's prediction and the true value (Supplementary Table 6).

The following three GUIDE models yielded the highest test AUC of 0.77 using (1) 10 pre-selected markers, (2) 7 markers forward selected from the total markers, and (3) 17 markers selected from the total markers *via* LASSO (Table 2). In the GUIDE method, when 10 selected markers and CL were used, cases with a CL < 17.5 mm were highly related to PTB, and in cases of CL > 17.5 mm, when *Ureaplasma* and *Prevotella* increased, there was a tendency toward PTB (Figure 5A). Among the markers selected by forward selection and LASSO, an increase in *Peptoniphilus lacrimalis* showed a high association with PTB when CL <17.5 mm (Figure 5B).

Discussion

This is the first study using a machine learning technique to predict PTB using vaginal microbiome, blood WBC, and CL, suggesting that the generated prediction model could be used to predict PTB considering model validation. In this study, the vaginal microbiome was analyzed using 16s rRNA metagenome sequencing.

In most microbiome studies with 16s rRNA metagenome sequencing, an operational taxonomic unit (OTU) is commonly used. OTU is derived from a cluster of similar sequences, while ASV is inferred from a unique sequence. Hence, ASVs can define sequences in one nucleotide difference and provide finer resolution (Callahan et al., 2017). In addition, their representations of sequences do not depend on the choice of reference database, because the inference is not computed based on the reference database but on the sequences (Callahan et al., 2017). AST inference is computed by separating technical errors and biological differences. Learning error rates and denoising errors are essential in making an ASV table. We chose to use DADA2 pipeline, because it is most popularly used to estimate error rates and statistically denoise the errors in the sequences (Callahan et al., 2016).

Candidate markers, used in training machine learning models, were selected through a combination of various methods, such as literature search and numerous statistical tests. As we focused on building machine learning models that can accurately predict preterm births, we chose a wider range of markers for the models by using a less stringent threshold in the statistical tests. To avoid any possible positive errors, we selected differentially abundant markers commonly detected by at least two statistical methods. Then, 17 markers were selected using a multiple marker selection method that can predict PTB (AUC = 0.78). Markers selected using forward selection, stepwise selection, and LASSO showed significant performance (AUC = 0.84).

We chose use AUC because it has advantages of comparing models with a combination of sensitivity and specificity in all decision thresholds. As AUC represents area under a curve that is drawn from all possible combinations of sensitivity and specificity, models with high AUC will have high sensitivity and specificity.

In model comparison, the models using various marker selection methods from the total markers set resulted in best predictions across a large portion of the model (Table 2). It is because these models utilized not only the statistically significant markers from 10 and 17 markers sets but also additional markers not selected in the statistical tests. These additional markers were chosen through forward/stepwise selection and lasso as they enhanced model performances. For example, *Moraxella osloensis* and *P. lacrimalis* were filtered out according to our filtering criteria but they were selected in the marker selection from the total markers set. As a result, the models using the total marker set resulted in the best predictions across a large portion of the models.

This study validated the prediction model by evaluating its performance on the test set. Using the test set, it was possible to predict the model performance for future patients. Furthermore, using the GUIDE method, it was possible to determine the role of each microbiome. Through the presentation of the tree using the GUIDE method, it was confirmed that the association with PTB was high when *Prevotella* and *Ureaplasma* increased, and it could also

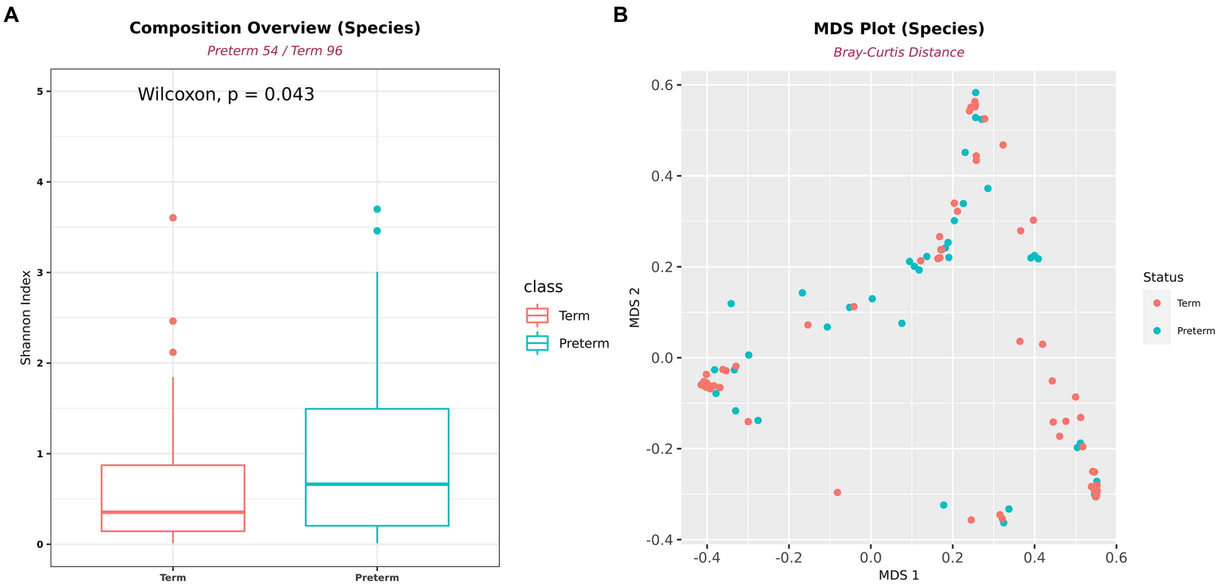

FIGURE 3
Differences in alpha- and beta-diversity between PTB and TB groups. **(A)** Shannon's alpha diversity was significantly higher in the PTB group (PTB, *n*=54; TB, *n*=96). **(B)** Multidimensional scaling plot. Boxes show median and interquartile ranges, and whiskers extend from minimum to maximum values.

TABLE 2 Performances of different multiple marker selection methods and test AUC comparison in prediction models.

		Variables	Train AUC	Validation AUC	Test AUC	LR	RF	XGB	SVM	GUIDE
10 Markers	Best Subset[1]	5	**0.84**	**0.79**	0.68	0.68	**0.74**	**0.77**	**0.74**	**0.73**
	Forward[2]	7	**0.83**	**0.83**	0.66	0.66	0.68	**0.77**	0.66	**0.73**
	Total	12	**0.87**	**0.71**	0.70	0.70	0.63	**0.77**	**0.75**	**0.77**
17 Markers	Best Subset[3]	7	**0.88**	**0.81**	0.57	0.57	0.63	**0.78**	0.59	**0.73**
	Forward[4]	7	**0.83**	**0.83**	0.65	0.65	**0.71**	**0.78**	0.61	**0.73**
	Total	19	**0.95**	0.55	0.60	0.60	0.60	**0.76**	0.60	0.57
365 Markers	Forward[5]	9	**0.98**	**1**	**0.81**	**0.81**	**0.84**	**0.83**	**0.82**	**0.77**
	Stepwise[6]	5	**0.96**	**0.92**	**0.78**	**0.78**	**0.72**	**0.83**	**0.81**	0.63
	Lasso[7]	19	**0.99**	**0.78**	**0.78**	**0.78**	**0.79**	**0.78**	**0.76**	**0.77**

Models with a higher AUC (>0.70) are shown in bold font. Logistic regression (LR), random forest (RF), XGBoost (XBG), support vector machine (SVM), and generalized unbiased interaction detection and estimation (GUIDE) were used to develop the prediction model. CLR-transformed data were used in the LR model and SVM, and proportional data were used for RF and XGB. The markers selected from the different methods are as follows:

[1] WBC, cervix length, *Lactobacillus fornicalis, Ureaplasma parvum, Prevotella timonensis.*

[2] WBC, cervical length, *U. parvum, P. timonensis, L. fornicalis, Lactobacillus crispatus gallinarum, Atopobium vaginae.*

[3] WBC, cervical length, *L. crispatus gallinarum, L. fornicalis, U. parvum, Lactobacillus paracasei, and Dialister propionicifaciens.*

[4] WBC, cervical length, *U. parvum, P. timonensis, L. fornicalis, D. propionicifaciens, and Mobiluncus curtisii.*

[5] WBC, cervical length, *Ureaplasma urealyticum, Alistipes finegoldii, Ruminococcus bromii, PAC001524_s, Peptoniphilus lacrimalis, L. crispatus gallinarum, and Lactobacillus jensenii.*

[6] WBC, cervix length, *U. urealyticum, A. finegoldii, R. bromii.*

[7] WBC, cervical length, *Prevotella disiens, A. finegoldii, Alistipes putredinis, PAC001031_s, PAC001524_s, Peptostreptococcus anaerobius, DQ905423_s, PAC001247_s, PAC001402_s, Anaerococcus tetradius, KQ960143_s, P. lacrimalis, Paracoccus marcusii hibiscisoli carotinifaciens, Moraxella osloensis, Pseudomonas glareae benzenivorans, U. parvum, and U. urealyticum.*

be explained that, as the number of *P. lacrimalis* increased, the association with PTB was higher (Figure 4).

The findings of this study showed similarity to those of various vaginal microbiome studies that used 16s rRNA metagenome sequencing. There have been several reports showing that *Lactobacillus* is related to PTB, and it is known that the risk of PTB is high in the group in which *Lactobacillus* is not dominant (Fettweis et al., 2019). If the bacterial diversity is high, the distribution of other pathogens increases, resulting in an increased risk of PTB (Fox and Eichelberger, 2015; Chu et al., 2018; Dominguez-Bello, 2019; Oliver et al., 2020). Various bacteria related to PTB have been reported (Freitas et al., 2018; Romero et al., 2019; You et al., 2019; Payne et al., 2020; Sprong et al., 2020); however, this study suggested that *Lactobacillus* spp. *U. parvum, M. curtisii, S. aureus,* and *Peptoniphilus grossensis* played a more important role, and the GUIDE method explained that *P. timonensis, U. parvum,* and *P. lacrimalis* played the

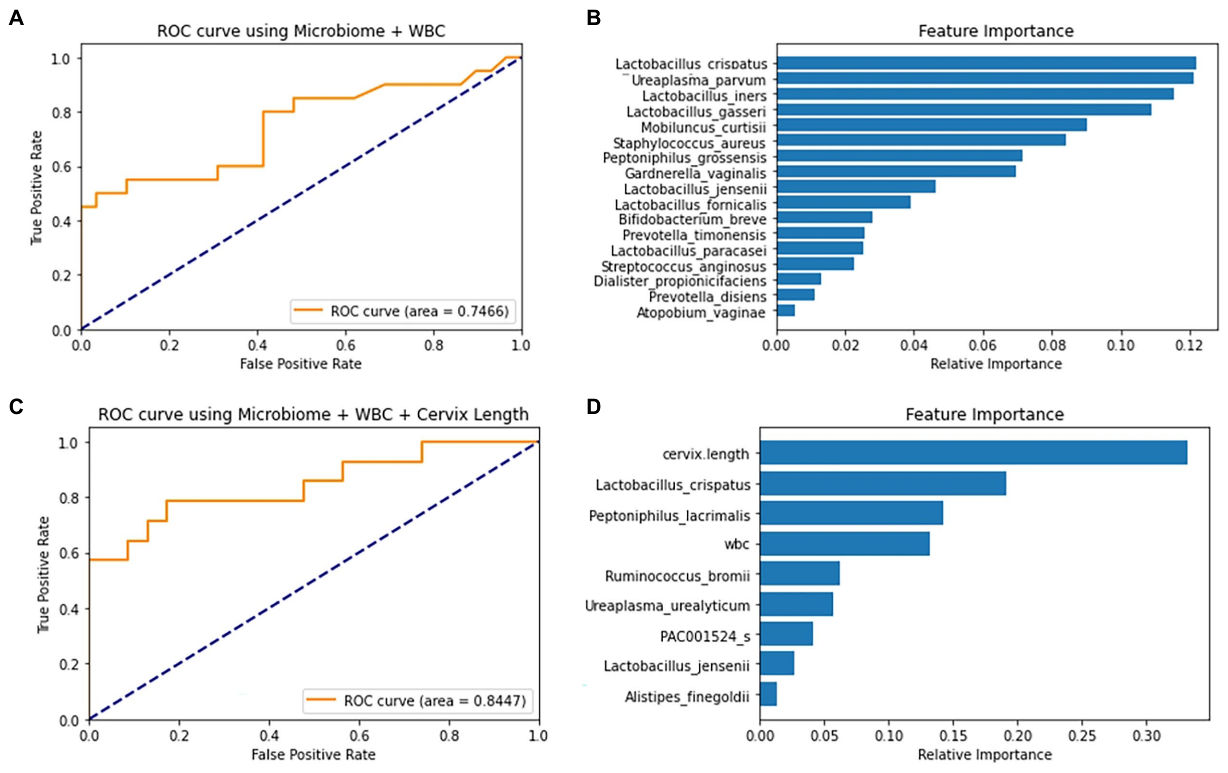

FIGURE 4
ROC curve and feature importance plot of the Random Forest (RF) models using covariates and selected markers. **(A)** RF model's ROC curve on test data using 10 selected markers and WBC **(B)** RF model's feature importance plot using 10 selected markers and WBC. **(C)** RF model's ROC curve on a test using forward-selected markers, WBC and cervical length. **(D)** RF model's Feature Importance plot using forward-selected markers, WBC and cervical length.

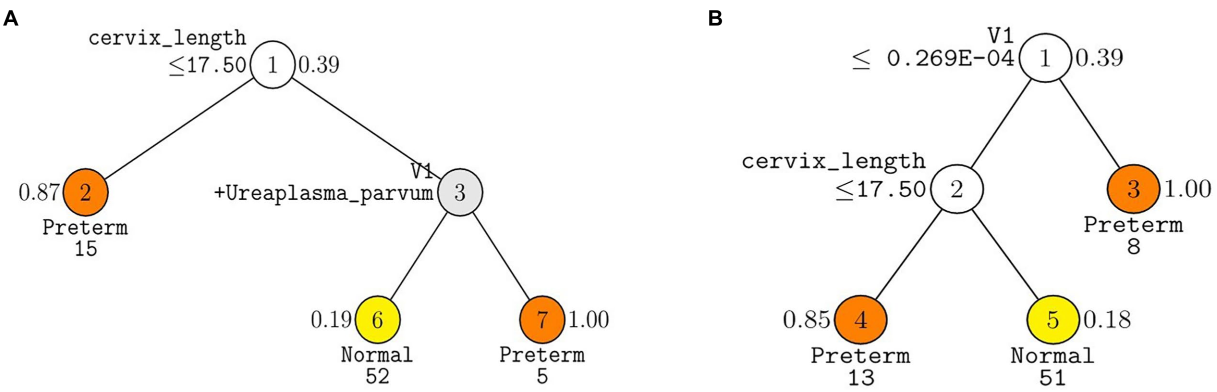

FIGURE 5
Decision trees made from GUIDE algorithm using covariates, cervix length and WBC with **(A)** ten pre-selected markers and **(B)** seven markers forward selected from the total markers. GUIDE v.38.0 classification tree for predicting Y using estimated priors and unit misclassification costs. Tree constructed with 109 observations. Pruning parameter α was 0.02 for **A** and 0.03 for **B**. At each split, an observation goes to the left branch if and only if the condition is satisfied. Predicted classes and sample sizes printed below terminal nodes; class sample proportion for Y=Preterm beside nodes. In **(A)**, V1 stands for *Prevotella timonensis*. In **(B)** V1 stands for *Peptoniphilus lacrimalis*.

most important role in PTB. The measurement of CL was performed to predict the risk of PTB in the second trimester globally. If it is shorter than 25 mm, it is considered high-risk, and if it is less than 15 mm, it is recommended to be hospitalized (Suff et al., 2019). The standard suggested by GUIDE in this study was 17.5 mm, and similarly, if it was shorter than the standard, the risk of PTB was high.

In this study, as a noninvasive method, the possibility of developing a PTB prediction model using the microbiome analysis results of CVF, blood test results, and CL measurement was shown. In addition, our study showed better predictive power than the existing method of PTB prediction. Fetal fibronectin (fFN) is commonly used in clinical applications for PTB prediction (Heng et al., 2015). However, the sensitivity of fFN is only 0.56 and that of phosphorylated insulin-like growth factor binding protein-1 (phIGFBP1) is 0.33 (Conde-Agudelo et al., 2011). In comparison, the PTB prediction model developed in this study showed a better sensitivity of 0.79 and specificity of 0.83. As a result, the application of this prediction model, based on the most important microbiome, could be clinically useful and cost-effective.

In this study, we compared the best prediction methods using various marker selection techniques and machine-learning methods. Although RF and XGB provide better prediction performance, they lack reasonable biological interpretation. For instance, in XGB model with forward selected markers from entire marker set, it is possible to observe each feature's impact by applying SHAP method (Supplementary Figure 6). However, it is difficult to interpret the model in relation to whole features. On the other hand, a simple single-tree model is easier to interpret, but is also known to have lower performance than other tree ensemble models (Hasan et al., 2020). The GUIDE, an enhanced version of a single tree, can still be used for the sake of interpretability with improved performance (Figure 4).

In future research, prediction models can be applied in clinical practice through a method that can more quantitatively evaluate the microbiome relationship, or it may be useful to substitute PCR tests that can utilize whether mRNA is expressed from DNA (Loh, 2014; Payne et al., 2020). To confirm the biological mechanism, it may be necessary to study proteomics and metabolomics in addition to genomics. Furthermore, studies on changes in cytokines or immune activation to determine how this microbiome acts with the host should be conducted.

Previous studies that presented the predictive power of machine learning models neglected to present the test AUC (Pedregosa et al., 2011; Hyman et al., 2014; Kumar et al., 2021; Urushiyama et al., 2021). This study confirmed the predictive power of the models on the test set. Therefore, this study presents a more accurate predictive power.

As the causes of PTB are very diverse, this was an attempt to increase predictive power by taking a combined approach that looked at the patient's blood test results and changes in CL, rather than a simple microbiome analysis (Park et al., 2021). In the marker selection process, we implemented an evidence-based medicine method through the existing literature review to select markers related to preterm birth, which can properly complement the machine-learning method using the data-driven hypothesis (Lamont et al., 2020). This study has strengths as it was a large-scale, multicenter study targeting pregnant Korean women. In addition, because the microbiome differs between races, it was possible to identify species related to PTB in Korea.

However, the limitations of this study are that the entire microbiota was not analyzed, including strain level measurements for *U. parvum*, despite recent studies showing that the pathogenicity of *Ureaplasma* differs depending on serovar levels. As a limitation of the method itself, 16s rRNA metagenome sequencing can analyze all colonized microbiomes of the vagina with high sensitivity, but it is difficult to identify the actual activity and pathogenicity of the microbiome. In addition to measuring CL to predict preterm birth, recently, a method of predicting preterm birth using elastography has been widely used (Seol et al., 2020), but this method was not applied in this study.

Our study demonstrates that several candidate microbiota could be used as potential predictors for PTB, and we confirmed that the predictive rate can be increased through a machine learning model based on the cervical length and WBC count.

Data availability statement

The data presented in the study are deposited in the SRA repository, accession number PRJNA845012.

Ethics statement

The studies involving human participants were reviewed and approved by this study was approved by the Ethical Research Committee of Ewha Womans University Mokdong Hospital (no. 2018-07-007) and Yonsei University Severance Hospital (no. 4-2018-0564), and all participants provided written informed consent. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SP enrolled subjects and wrote and edited the manuscript. JM wrote the manuscript and analyzed data. NK analyzed data and interpreted data. Y-HK enrolled subjects and designed the study. Y-AY developed the extraction of protocols and interpreted analyzed data. EK and AA developed the extraction of protocols and performed the experiments. YH enrolled subjects. YK designed the study, obtained funding, and enrolled subjects. TP designed the study and obtained funding. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT; grant no. 2020R1A2C3011850) and the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF). This work was also supported by a grant from the Korea Health Technology R&D Project through the Korea Health

Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant numbers: HI16C2037).

Acknowledgments

We would like to thank Editage (www.editage.co.kr) for English language editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.912853/full#supplementary-material

References

Ananth, C. V., Friedman, A. M., Goldenberg, R. L., Wright, J. D., and Vintzileos, A. M. (2018). Association between temporal changes in neonatal mortality and spontaneous and clinician-initiated deliveries in the United States, 2006–2013. *JAMA Pediatr.* 172, 949–957. doi: 10.1001/jamapediatrics.2018.1792

Bennett, P. R., Brown, R. G., and MacIntyre, D. A. (2020). Vaginal microbiome in preterm rupture of membranes. *Obstet. Gynecol. Clin. North Am.* 47, 503–521. doi: 10.1016/j.ogc.2020.08.001

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325–349. doi: 10.2307/1942268

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Chan, R. L. (2014). Biochemical markers of spontaneous preterm birth in asymptomatic women. *Biomed. Res. Int.* 2014:164081. doi: 10.1155/2014/164081

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; August 13–17, 785–794.

Chu, D. M., Seferovic, M., Pace, R. M., and Aagaard, K. M. (2018). The microbiome in preterm birth. *Best Pract. Res. Clin. Obstet. Gynaecol.* 52, 103–113. doi: 10.1016/j.bpobgyn.2018.03.006

Conde-Agudelo, A., Papageorghiou, A. T., Kennedy, S. H., and Villar, J. (2011). Novel biomarkers for the prediction of the spontaneous preterm birth phenotype: a systematic review and meta-analysis. *BJOG* 118, 1042–1054. doi: 10.1111/j.1471-0528.2011.02923.x

Della Rosa, P. A., Miglioli, C., Caglioni, M., Tiberio, F., Mosser, K. H. H., Vignotto, E., et al. (2021). A hierarchical procedure to select intrauterine and extrauterine factors for methodological validation of preterm birth risk estimation. *BMC Pregnancy Childbirth* 21:306. doi: 10.1186/s12884-021-03654-3

Dominguez-Bello, M. G. (2019). Gestational shaping of the maternal vaginal microbiome. *Nat. Med.* 25, 882–883. doi: 10.1038/s41591-019-0483-6

Dunlop, A. L., Satten, G. A., Hu, Y. J., Knight, A. K., Hill, C. C., Wright, M. L., et al. (2021). Vaginal microbiome composition in early pregnancy and risk of spontaneous preterm and early term birth among African American women. *Front. Cell. Infect. Microbiol.* 11:641005. doi: 10.3389/fcimb.2021.641005

Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., et al. (2019). The vaginal microbiome and preterm birth. *Nat. Med.* 25, 1012–1021. doi: 10.1038/s41591-019-0450-2

Fox, C., and Eichelberger, K. (2015). Maternal microbiome and pregnancy outcomes. *Fertil. Steril.* 104, 1358–1363. doi: 10.1016/j.fertnstert.2015.09.037

Freitas, A. C., Bocking, A., Hill, J. E., and Money, D. M. (2018). Increased richness and diversity of the vaginal microbiota and spontaneous preterm birth. *Microbiome* 6:117. doi: 10.1186/s40168-018-0502-8

Goldenberg, R. L., Culhane, J. F., Iams, J. D., and Romero, R. (2008). Epidemiology and causes of preterm birth. *Lancet* 371, 75–84. doi: 10.1016/s0140-6736(08)60074-4

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., and Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access* 8, 76516–76531. doi: 10.1109/ACCESS.2020.2989857

Heng, Y. J., Liong, S., Permezel, M., Rice, G. E., Di Quinzio, M. K., and Georgiou, H. M. (2015). Human cervicovaginal fluid biomarkers to predict term and preterm labor. *Front. Physiol.* 6, 151. doi: 10.3389/fphys.2015.00151

Hur, Y. M., Kang, M. N., and Kim, Y. J. (2021). Vaginal health in women and the possibility of predicting preterm birth through microbiome analysis. *J. Korean Med. Assoc.* 64, 833–840. doi: 10.5124/jkma.2021.64.12.833

Hyman, R. W., Fukushima, M., Jiang, H., Fung, E., Rand, L., Johnson, B., et al. (2014). Diversity of the vaginal microbiome correlates with preterm birth. *Reprod. Sci.* 21, 32–40. doi: 10.1177/1933719113488838

Kataoka, S., Yamada, T., Chou, K., Nishida, R., Morikawa, M., Minami, M., et al. (2006). Association between preterm birth and vaginal colonization by mycoplasmas in early pregnancy. *J. Clin. Microbiol.* 44, 51–55. doi: 10.1128/jcm.44.1.51-55.2006

Kim, J. R., Han, K., Han, Y., Kang, N., Shin, T. S., Park, H. J., et al. (2021). Microbiome markers of pancreatic Cancer based on Bacteria-derived extracellular vesicles acquired from blood samples: a retrospective propensity score matching analysis. *Biology* 10:219. doi: 10.3390/biology10030219

Kim, S. I., Song, M., Hwangbo, S., Lee, S., Cho, U., Kim, J. H., et al. (2019). Development of web-based nomograms to predict treatment response and prognosis of epithelial ovarian Cancer. *Cancer Res. Treat.* 51, 1144–1155. doi: 10.4143/crt.2018.508

Koullali, B., Oudijk, M. A., Nijman, T. A., Mol, B. W., and Pajkrt, E. (2016). Risk assessment and management to prevent preterm birth. *Semin. Fetal Neonatal Med.* 21, 80–88. doi: 10.1016/j.siny.2016.01.005

Kumar, M., Murugesan, S., Singh, P., Saadaoui, M., Elhag, D. A., Terranegra, A., et al. (2021). Vaginal microbiota and cytokine levels predict preterm delivery in Asian women. *Front. Cell. Infect. Microbiol.* 11:639665. doi: 10.3389/fcimb.2021.639665

Lamont, R. F., Richardson, L. S., Boniface, J. J., Cobo, T., Exner, M. M., Christensen, I. B., et al. (2020). Commentary on a combined approach to the problem of developing biomarkers for the prediction of spontaneous preterm labor that leads to preterm birth. *Placenta* 98, 13–23. doi: 10.1016/j.placenta.2020.05.007

Loh, W.-Y. (2009). Improving the precision of classification trees. *Ann. Appl. Stat.* 3, 1710–1737. doi: 10.1214/09-AOAS260

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdiscip. Rev.* 1, 14–23. doi: 10.1002/widm.8

Loh, W. Y. (2014). Fifty years of classification and regression trees. *Int. Stat. Rev.* 82, 329–348. doi: 10.1111/insr.12016

Newnham, J. P., Kemp, M. W., White, S. W., Arrese, C. A., Hart, R. J., and Keelan, J. A. (2017). Applying precision public health to prevent preterm birth. *Front. Public Health* 5, 66. doi: 10.3389/fpubh.2017.00066

Oliver, A., LaMere, B., Weihe, C., Wandro, S., Lindsay, K. L., Wadhwa, P. D., et al. (2020). Cervicovaginal microbiome composition is associated with metabolic profiles in healthy pregnancy. *mBio* 11:e01851-20. doi: 10.1128/mBio.01851-20

Park, S., Oh, D., Heo, H., Lee, G., Kim, S. M., Ansari, A., et al. (2021). Prediction of preterm birth based on machine learning using bacterial risk score in cervicovaginal fluid. *Am. J. Reprod. Immunol.* 86:e13435. doi: 10.1111/aji.13435

Park, S., You, Y. A., Yun, H., Choi, S. J., Hwang, H. S., Choi, S. K., et al. (2020). Cervicovaginal fluid cytokines as predictive markers of preterm birth in symptomatic women. *Obstet. Gynecol. Sci.* 63, 455–463. doi: 10.5468/ogs.19131

Payne, M. S., Newnham, J. P., Doherty, D. A., Furfaro, L. L., Pendal, N. L., Loh, D. E., et al. (2020). A specific bacterial DNA signature in the vagina of Australian women in midpregnancy predicts high risk of spontaneous preterm birth (the Predict1000 study). *Am. J. Obstet. Gynecol.* 224, 206.e1–206.e23. doi: 10.1016/j.ajog.2020.08.034

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Petricevic, L., Domig, K. J., Nierscher, F. J., Sandhofer, M. J., Fidesser, M., Krondorfer, I., et al. (2014). Characterisation of the vaginal *Lactobacillus microbiota* associated with preterm delivery. *Sci. Rep.* 4:5136. doi: 10.1038/srep05136

Romero, R., Dey, S. K., and Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science* 345, 760–765. doi: 10.1126/science.1251816

Romero, R., Gomez-Lopez, N., Winters, A. D., Jung, E., Shaman, M., Bieda, J., et al. (2019). Evidence that intra-amniotic infections are often the result of an ascending invasion - a molecular microbiological study. *J. Perinat. Med.* 47, 915–931. doi: 10.1515/jpm-2019-0297

Seol, H. J., Sung, J. H., Seong, W. J., Kim, H. M., Park, H. S., Kwon, H., et al. (2020). Standardization of measurement of cervical elastography, its reproducibility, and analysis of baseline clinical factors affecting elastographic parameters. *Obstet. Gynecol. Sci.* 63, 42–54. doi: 10.5468/ogs.2020.63.1.42

Shannon, C. E. (1997). The mathematical theory of communication. 1963. *MD Comput.* 14, 306–317.

Sprong, K. E., Mabenge, M., Wright, C. A., and Govender, S. (2020). Ureaplasma species and preterm birth: current perspectives. *Crit. Rev. Microbiol.* 46, 169–181. doi: 10.1080/1040841x.2020.1736986

Suff, N., Story, L., and Shennan, A. (2019). The prediction of preterm delivery: what is new? *Semin. Fetal Neonatal Med.* 24, 27–32. doi: 10.1016/j.siny.2018.09.006

Tabatabaei, N., Eren, A. M., Barreiro, L. B., Yotova, V., Dumaine, A., Allard, C., et al. (2019). Vaginal microbiome in early pregnancy and subsequent risk of spontaneous preterm birth: a case-control study. *BJOG* 126, 349–358. doi: 10.1111/1471-0528.15299

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58, 267–288.

Urushiyama, D., Ohnishi, E., Suda, W., Kurakazu, M., Kiyoshima, C., Hirakawa, T., et al. (2021). Vaginal microbiome as a tool for prediction of chorioamnionitis in preterm labor: a pilot study. *Sci. Rep.* 11, 18971. doi: 10.1038/s41598-021-98587-4

Yoo, J. Y., Rho, M., You, Y. A., Kwon, E. J., Kim, M. H., Kym, S., et al. (2016). 16S rRNA gene-based metagenomic analysis reveals differences in bacteria-derived extracellular vesicles in the urine of pregnant and non-pregnant women. *Exp. Mol. Med.* 48:e208. doi: 10.1038/emm.2015.110

Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H., et al. (2017). Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* 67, 1613–1617. doi: 10.1099/ijsem.0.001755

You, Y. A., Kwon, E. J., Choi, S. J., Hwang, H. S., Choi, S. K., Lee, S. M., et al. (2019). Vaginal microbiome profiles of pregnant women in Korea using a 16S metagenomics approach. *Am. J. Reprod. Immunol.* 82:e13124. doi: 10.1111/aji.13124