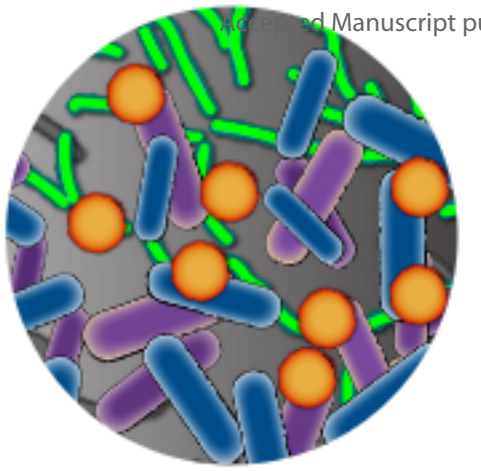
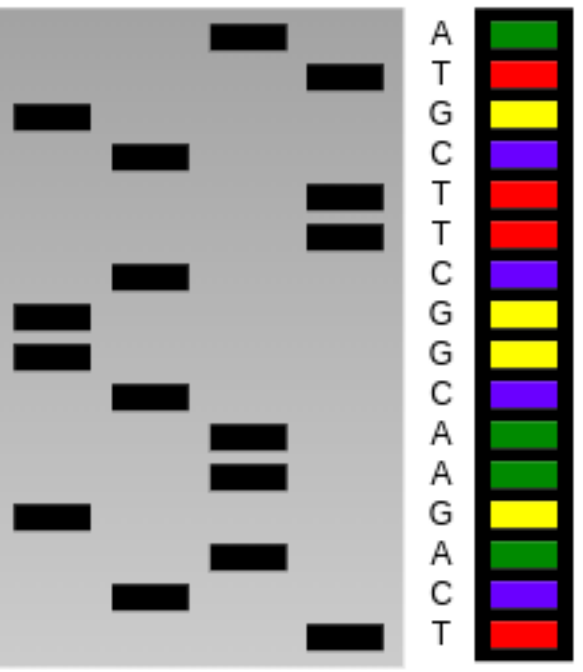
Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024



Vaginal Microbiota (Pregnancy Trimester 1 - Trimester 2)

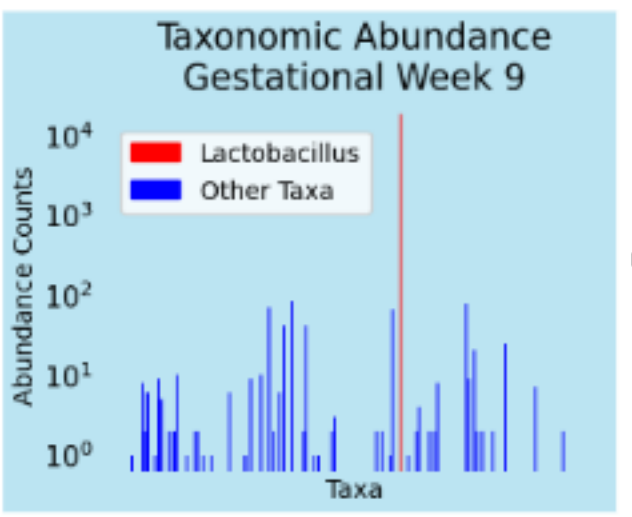
16S rRNA Gene

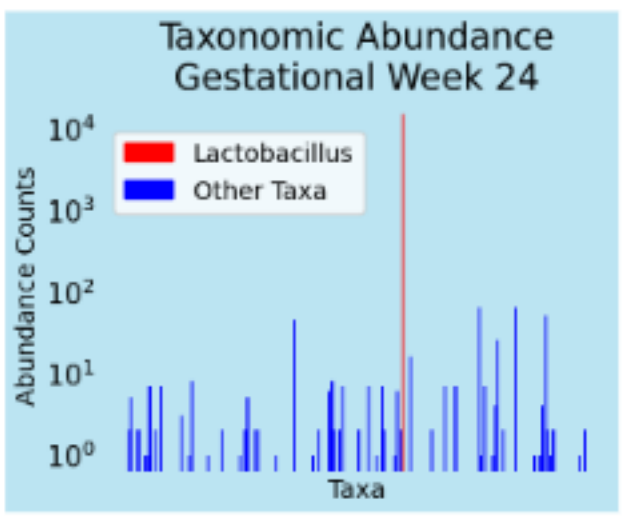
V4 Hypervariable Region



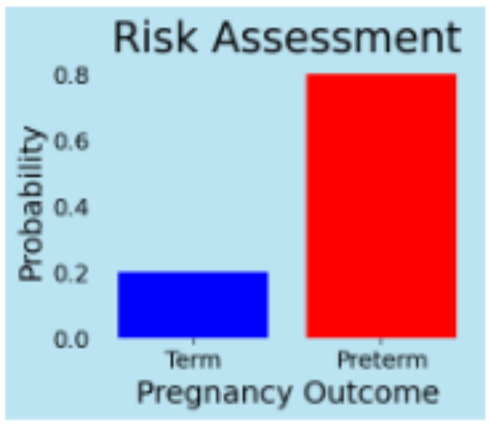
Sequencing

DADA2



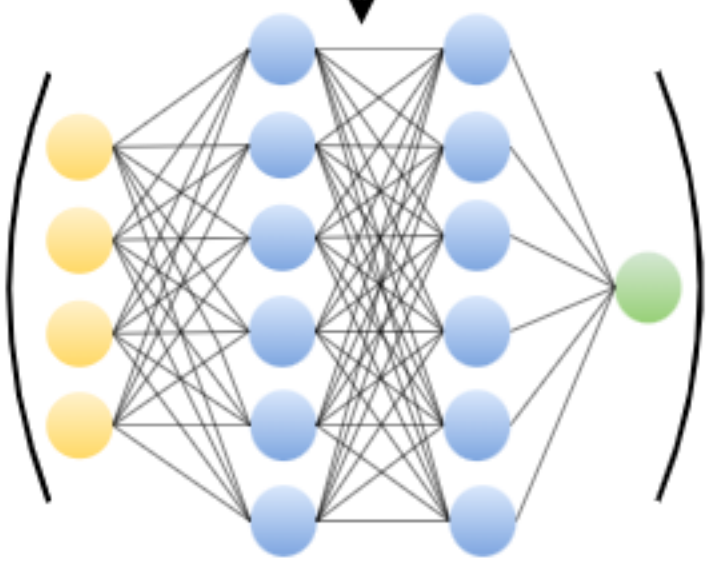


Week-Wise Taxonomic Abundance



Term v/s Preterm Prediction

$$\frac{d}{dt}$$



Neural Differential Equations

Copyright © 2023 the authors

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
via Open Access. This work is licensed under a Creative Commons Attribution
4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Deep Learning Enables Early-Stage Prediction of Preterm Birth Using Vaginal Microbiota

Kaushik Karambelkar[1, #] Mayank Baranwal[1 , 2, *]

[1]Data and Decision Sciences, Tata Consultancy Services Ltd., Thane, Maharashtra, India – 400607

[2]Systems and Control Engineering, Indian Institute of Technology, Bombay, Maharashtra, India – 400076

[#]Email (Institutional): kaushik.karambelkar@tcs.com

[*]Corresponding Author; Email (Institutional): baranwal.mayank@tcs.com, mbaranwal@iitb.ac.in

Time-Series Modeling of Vaginal Microbiota

Keywords: Preterm Birth, Deep Learning, Vaginal Microbiome, Time-Series Modeling

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Copyright © 2023 the authors
This Open Access

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

Abstract

Objective: Preterm birth (PTB) is one of the leading issues concerning infant health and is a problem that plagues all parts of the world. Vaginal microbial communities have recently garnered attention in the context of PTB, however, the vaginal microbiome varies greatly from individual to individual, and this variation is more pronounced in racially, ethnically and geographically diverse populations. Additionally, microbial communities have been reported to evolve during the duration of the pregnancy, and capturing such a signature may require higher, more complex modeling paradigms. In this study, we develop a neural controlled differential equations (CDEs) based framework for identifying early PTBs in racially diverse cohorts from irregularly sampled vaginal microbial abundance data.

Methods: We obtained relative abundances of microbial species within vaginal microbiota using 16S rRNA sequences obtained from vaginal swabs at various stages of pregnancy. We employed a recently introduced deep learning paradigm known as "Neural CDEs" to predict PTBs. This method, previously unexplored, analyzes irregularly sampled microbial abundance profiles in a time-series format.

Results: Our framework is able to identify signatures in the temporally evolving vaginal microbiome during trimester 2 and can predict incidences of PTB (mean test set ROC-AUC = 0.81, accuracy = 0.75, f1 score = 0.71) significantly better than traditional ML classifiers, thus enabling effective early-stage PTB risk assessment.

Conclusion and Significance: Our method is able to differentiate between term and preterm outcomes with a substantial accuracy, despite being trained using irregularly sampled microbial abundance profiles, thus overcoming the limitations of traditional time-series modeling methods.

1. Introduction

Preterm births (PTBs) are live births that occur before 37 weeks of pregnancy and are a major public health concern worldwide. It is estimated that about 15 million babies are born pre-term globally each year, putting the global PTB rate at about 11% (Blencowe et al., 2012). PTB is among the leading causes of neonatal mortality and morbidity, especially in low- and middle-income economies (Perin et al., 2022). The global PTB rate is also on a steady rise, thus making it a significant burden (Chawanpaiboon et al., 2019). Approximately 18% of the deaths among children under the age of 5 years happen within the first 28 days of life and can be attributed to complications arising from PTB (Walani, 2020). Additionally, it can lead to long-term health complications such as respiratory illnesses, neurodevelopmental disorders and learning disabilities, arising from the developmental issues associated with PTB (Townsi et al., 2018; Chung et al., 2020). PTB is also not a problem that is specific to underdeveloped or developing countries, although the ill-effects of it may be more pronounced in low-income countries. Incidences of PTB are also found in high-income parts of the world, albeit with a lower frequency, and the rates of PTB have not been on the decline either in most parts of the world (Chawanpaiboon et al., 2019).

The pathophysiology behind PTB is not completely understood yet, although certain risk factors, including but not limited to smoking habits, alcohol intake and reproductive history, have been identified to be associated with increased pre-term delivery risk (Blencowe et al., 2012; Pfinder et al., 2013; Stock et al., 2020). The ill-effects of PTB can be mitigated, and a healthy, full-term gestation outcome may also be achieved if appropriate interventions are administered (Newnham et al., 2014). Their success, however, depends on identifying at-risk subjects at earlier stages of their pregnancy, as these approaches are effective when administered during the earlier stages (Blencowe et al., 2012; Newnham et al., 2014). Current methods for assessing pre-term pregnancy outcomes involve the use of physical and biochemical markers, which are not accurately determinative of a potential incidence of premature culmination of pregnancy in the future (Georgiou et al., 2015).

Many non-pathogenic bacteria, viruses and fungi inhabit various areas of the human body, such as the gut, mouth, and reproductive tracts (Sender et al., 2016), and are collectively referred to as the "microbiome". Microbiomes are essential for normal functioning of the respective organs, and maintain a symbiotic relationship with the human body and drive key biochemical reactions, (Gordon et al., 1971) and dysregulated microbiomes are often implicated in various diseases. These microbial communities are also present in the reproductive tracts, and have been reported to influence the pregnancy outcome (MacIntyre et al., 2015). There is evidence linking the composition of vaginal microbiomes to risk of PTB, and the abundance levels of

Copyright © 2023 the authors
This is an Open Access work licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

specific microbiota, such as various species of the Lactobacillus genus, have the potential to be indicative of PTB even at earlier stages of pregnancy (Brown et al., 2019; Romero et al., 2014). Vaginal microbial communities can be categorized into specific Community State Types (CSTs), which are typically characterized by abundances of various Lactobacillus species (Romero et al., 2014). CSTs are associated with increased or decreased risk of abnormalities such as Bacterial Vaginosis (BV), Urinary Tract Infections (UTIs) and even PTB (Gudnadottir et al., 2022). Moreover, alpha-diversity indices, such as Shannon and Simpson diversity, which can quantify the diversity of vaginal microbiota, have been harnessed for predicting PTB (DiGiulio et al., 2015; Haque et al., 2017). However, the vaginal microbiome differs considerably from individual to individual, especially across races (Sun et al., 2022; Gupta et al., 2017). Additionally, the microbial abundance may further vary depending on the sequence processing methods used on 16S ribosomal RNA (rRNA) data, which is typically used to estimate taxonomic abundance at various levels of classification (Bharti et al., 2019). Consequently, the success of diversity indices for estimating PTB risk may be specific to certain cohorts, or be influenced by the sequence processing pipeline and consequently, may not translate across cohorts, as is our observation in this study. Machine Learning (ML)-based approaches have also been explored in this context, which leverage features such as abundance of various taxa, phylotype counts, CST of the vaginal microbiome, age, race and more, for PTB risk assessment.

The vaginal microbiome evolves as the pregnancy progresses (MacIntyre et al., 2015; Romero et al., 2014), and the numerous changes that it undergoes may contain a signature for identifying PTB risk. Currently, there is a severe lack of approaches that exploit the temporal dynamics of vaginal microbiomes, by looking at it as a time-series problem, for PTB risk assessment, and most current predictive methodologies use static data. Deep learning approaches such as recurrent neural networks (RNNs), have previously been used for modeling the dynamics of gut and other microbiota in various contexts (Baranwal et al., 2022; Fung et al., 2023; Medina et al., 2022) and have found success. To the best of our knowledge, such methods have not been applied to vaginal microbiota, especially in the context of PTB risk assessment, so far. This may be partly attributed to the fact that RNN-based approaches demand data sampled at regular intervals, which is challenging to collect as study subjects are often irregular in clinical visits. With this study, we present a deep learning-based approach, "neural controlled differential equations (CDEs)" that is capable of differentiating between term and preterm births using time-series vaginal microbiome data, which overcomes the dependence on regularly sampled microbial data. We also highlight the limitations of alpha diversity indices and traditional ML methods for PTB prediction in racially and ethnically diverse patient cohorts. We show that modeling the temporal dynamics of microbiota using deep learning methods results in more reliable PTB risk scoring than simple ML-based methods. Our best model, utilizing neural CDEs, outperforms any ML-based PTB prediction approaches so far. On the basis of this study, we show the potential of vaginal microbiota for PTB prediction, and that such approaches can be pushed towards complete clinical viability with further efforts.

2. Materials and Methods

2.1. Dataset

We obtained 16S rRNA sequences collected from human patient samples. The data was sourced from a previously published study by Callahan et al. on refinement of a vaginal microbiome signature of preterm birth (Callahan et al., 2017). The dataset is publicly available under the open access category in the Sequence Read Archive (SRA), BioProject ID PRJNA393472. It consists of 16S rRNA sequence samples, spread across 133 racially and ethnically diverse subjects, and sampled at different points of time during the pregnancy.

2.2. 16S rRNA Sequence Processing

It has been widely established that the hypervariable regions (V1-V9) within 16S rRNA gene can be used for phylogenetic studies and genus or species-level classification in diverse microbial populations (Weisburg et al, 1991). Furthermore, certain hypervariable regions (such as the V4) are semi-conserved and can reliably predict specific taxonomic levels (Yang et al., 2016). The procedure to convert the 16S rRNA sequence data to microbial abundance involves various stages of processing. In the first step, quality control checks are performed and sequencing artifacts, low quality reads, etc. are removed from the read sequences. Secondly, the preprocessed sequences are aligned against a chosen reference database, and a taxonomic class is assigned

Copyright © 2023 the authors
This is an Open Access work licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

to each sequence. The sequences are then grouped into clusters, which represent Operational Taxonomic Units (OTUs), based on sequence similarities. Lastly, the abundances of each OTU in samples are estimated (Schloss et al., 2009; Edgar, 2013; Estaki et al., 2020; Callahan et al., 2016). The microbial abundance obtained depends on the specific processing steps, and variations in processing steps can result in different abundance values (Schloss et al., 2009; Edgar, 2013; Estaki et al., 2020; Callahan et al., 2016).

The dataset (PRJNA393472) derived from SRA contains sequences generated after amplifying and sequencing the V4 hypervariable region of the 16S rRNA gene. We used the DADA2 processing pipeline (Callahan et al., 2016) to derive microbial abundance data from the sequence reads. The metadata and taxonomic abundance tables were generated using the SRA cloud (Katz et al., 2021) and abundances were obtained at various levels of taxonomic classification. We retained genus-level abundances for all our analyses since abundances at further levels were captured at a much lower resolution.

2.3. Processing Taxonomic Abundance Data

We eliminated the samples for which metadata information for certain key fields, viz., gestational age at the time of sample collection, gestational age at delivery, etc., was missing. We eliminated the genera abundance for samples collected during trimester 3 (gestational age > 24 weeks) for our analyses, with the intention of being able to predict instances of preterm delivery sufficiently early. Furthermore, we removed samples collected during or before the 8^{th} week of gestation, as they were present for very few (6 out of 133) subjects. Furthermore, for some of the analyses, we transformed the genera abundance data to sample-wise relative abundance, and filtered out genera with high skewness and high kurtosis, thus removing some of the genera whose abundances contributed to noise. We retained 70% of the subjects (90 out of 133) as the training dataset and the rest were used for validating the approaches. The training and test datasets were kept consistent across all the analyses. The processed taxonomic abundance data and the corresponding metadata are made available in the code repository (see "Code Availability", Section 5).

2.4. Diversity Metrics

Alpha diversity metrics have been reported to be potentially indicative of preterm birth, and a highly-diverse vaginal microbiome is correlated with increased risk of preterm delivery (DiGiulio et al., 2015; Hyman et al., 2014; Haque et al., 2017). We computed Shannon, Simpson, Chao1 and Gini alpha diversity indices, as well as Taxonomic Composition Skew (TCS) (Haque et al., 2017), a diversity index specifically tailored for the vaginal microbiome. Unlike other diversity indices, TCS takes into account that vaginal microbiomes are usually dominated by the Lactobacillus species and other genera are in the minority. TCS responds in a different manner, to changes in abundances of sparse and dominant taxa, and thus is possibly more suitable for quantifying the diversity in vaginal microbiomes. We checked for statistically significant differences in alpha-diversity index values between the term and preterm classes during various gestational periods using a two-sided, independent *t*-test. The standard diversity metrics were computed using the scikit-bio python library (version 0.5.8).

2.5. Traditional ML Approaches

We used two ML classifiers: Decision Tree (DT) and Random Forest (RF), to predict term/preterm outcomes. This constituted a secondary baseline for benchmarking the performance of higher, more complex deep learning-based prediction approaches. For each patient subject, the microbial abundance profile closest to the

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This is an Open Access work. This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Copyright © 2023 the authors

week of delivery and obtained during the period between the 9th and the 24th week of gestation, following the hypothesis that composition of vaginal microbial communities closer to the period of delivery are better indicative of preterm delivery risk. The resultant training and test sets contained 93 and 40 samples respectively.

2.6. Deep Learning Approaches

Machine learning classifiers, such as Support Vector Classifiers (SVCs), as well as tree-based classifiers such as DT and RF have been explored extensively for preterm birth prediction using vaginal microbiota, most often in tandem with other features such as physical markers and patient history. However, these classifiers have largely lacked the capability of making reliable predictions. Surprisingly, deep learning models have hardly been explored for this particular problem. Given the time-series nature of the data, we focused on deep learning algorithms for sequential data in this study.

2.6.1. Recurrent Neural Networks

Recurrent Neural Network (RNN) is a type of neural network designed to handle sequential or time-series data. The issue with standard RNNs however, is that they have difficulty in learning long-term dependencies in long sequences, due to the issue of vanishing/exploding gradients (Pascanu et al., 2012). Long Short-Term Memory (LSTM) is a type of RNN that is capable of learning long sequences, and are possibly more appropriate for the week-wise taxonomic abundance dataset. LSTM maintains a hidden state, which stores short-term information, and a cell state, which stores long-term information. The initial hidden and cell states are generally set to zero vectors. A LSTM cell at each time step updates the hidden and cell states based on the states at the previous time step and the input data at the current time step.

$$(h_{t_0}, c_{t_0}) = (\vec{0}, \vec{0})$$

$$z_{t_i} = \begin{cases} x_{t_i}, & \text{if } x_{t_i} \text{ is available} \\ \dot{x}_{t_i}, & \text{if } x_{t_i} \text{ is unavailable} \end{cases}$$

$$(h_{t_i}, c_{t_i}) = \text{LSTM}\left(z_{t_i}, [h_{t_{i-1}}, c_{t_{i-1}}]\right), \text{ for all } i \in \{1, 2, \ldots, T\}$$

$$y = \sigma(W \cdot h_{t_T} + b)$$

However, the conventional LSTM system demands a continuous and uniform time-series dataset, i.e., the time-steps must represent uniform intervals and the input data should be available for each step. While the taxonomic abundance data is uniformly sampled (week-wise), data for some subjects is not present for some weeks. Additionally, the first week for which data is available is different for each subject, and thus, a zero-vector initialization will not be appropriate for the initial hidden state. To address these issues, we modified the LSTM network accordingly. Firstly, we initialized the hidden state, h_{t_0}, with a trainable embedding layer, and the cell state, c_{t_0}, was initialized as a zero vector. Secondly, at each time step, we had the LSTM cell generate the taxonomic abundance forecast, x_{t_i}, at each time step, and used the forecast whenever the input data x_{t_i} was not available. The final hidden state, h_{t_T}, was fed to a linear layer with parameters W and b followed by a sigmoid activation, $\sigma(\cdot)$, to predict the term/pre-term outcome (y). The entire network was trained end-to-end.

Figure 1 outlines the described LSTM network. The LSTM implementation assumes that the input data is continuously and regularly sampled. However, in our case, data for some intervals may be missing. To overcome this, we masked the data for missing time steps by using zero vectors to substitute the missing data points. In parallel, we also used a vector indicating the coordinates of the masked time intervals for each sample, for which the model used the forecast, i.e., the model-predicted microbial abundance instead of the ground truth values. Additional method details are provided in the supplementary material (Section 2.1) and the hyperparameter values are listed in supplementary Table 3. The LSTM model was implemented using the pytorch library (Paszke et al., 2019) (version 2.0.1, cuda version 11.8).

2.6.2. Neural Differential Equations

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This is an Open Access work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Copyright © 2023 the authors

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

A significant limitation associated with clinical data pertains to its irregular sampling of data points, which presents challenges in constructing effective machine learning models that can effectively harness the inherent time-series information. The irregularity in data sampling introduces two notable drawbacks: firstly, the size of input data, contingent upon the number of sampling instances, differs among various subjects; secondly, the timing of sampling instances is not strictly discrete, thereby restricting the applicability of commonly employed RNN models that assume uniform intervals between sampled data points. To overcome this, we leverage a recently introduced class of deep learning models - Neural Ordinary Differential Equations (ODEs) that combine a neural network with ODEs and allow for continuous interpolation between two randomly spaced sampling instants.

Notably, Neural ODEs exclusively consider the evolution of time-series data commencing at a fixed time point, denoted as t_0, which is accompanied by an initial condition represented as x_{t_0}. In the context of our research, x_{t_0} signifies the initial abundance of genera at this specific time t_0, which might correspond to the onset of gestation week 9. Regrettably, the trajectory of microbial abundance varies from subject to subject, and the initial abundance data for all subjects may not be accessible for subsequent analysis, as the microbial profiles of each subject were not uniformly sampled at the same time point, namely t_0. Consequently, while Neural ODEs excel in interpolation tasks, they cannot be seamlessly integrated into our framework due to the absence of consistent initial abundance data. Nevertheless, it transpires that addressing this problem, specifically how to integrate incoming information, has already been thoroughly explored within the realm of mathematics, particularly in the field of rough analysis, which is dedicated to the examination of CDEs (Lyons, 1994; Lyons et al., 2007). Kidger et al. (Kidger et al., 2020) have introduced a novel framework known as Neural CDEs, which extends CDEs to Neural ODE models. To put it simply, Neural CDEs can be seen as continuous-time counterparts of Recurrent Neural Network (RNN) models. These models can be trained efficiently using a method called "adjoint backpropagation", which is elaborated on briefly in the supplementary materials (Section 3.1), and detailed mathematical representation of it can be found in (Chen et al., 2018). In brief, the Neural CDE model can be summarized through the following sequence of operations:

$$\text{(Initialization)} \quad h_{t_0} = \psi_\theta(t_0, x_{t_o})$$

$$\text{(CDE)} \quad \frac{dh}{dt}(t) = f_\theta(h_t)\frac{dX}{dt}(t)$$

$$\text{(Result)} \quad y = l_\theta(h_{t_T})$$

Here, ψ_θ and l_θ correspond to linear models responsible for transforming the initial taxa abundance (along with the time-stamp t_0) into the initial hidden state, h_{t_0}, and the final hidden state, h_{t_T}, into the output label, respectively. The map, ψ_θ is used to avoid translational invariance to first sampled time instant. X is the natural cubic spline with knots at $t_0,\ldots,t_T$ such that $X_{t_i} = (x_i, t_i)$. Natural cubic splines allow for smooth interpolation and minimum regularity for handling certain edge cases. f_θ is neural network model depending on parameters, θ. Due to its dependence on cubic splines, Neural CDEs (Kidger et al., 2020) can be applied to irregularly sampled time series, even with temporally-scattered initial conditions. Thus, we chose to apply Neural CDEs to predicting preterm birth using the irregularly-sampled microbial abundance dataset. A comprehensive mathematical introduction to Neural CDEs is outside the scope of this paper. For those interested in delving deeper into the mathematical details, we recommend consulting (Kidger et al., 2020) for a more thorough explanation. The torchcde library (version 0.2.5) was used to implement the Neural CDE model in python. Hyperparameter values for the model are listed in supplementary Table 4.

3. Results

3.1. Microbial abundance dataset contains racially and ethnically diverse subjects

The 16S rRNA sequence data was converted to taxonomic abundance (Methods, Section 2.2), which led to approximately 290,000 abundance counts, spanning taxonomic counts at various levels of classification, out

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This is an Open Access work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Copyright © 2023 the authors

of which approximately 65,000 corresponded to genus-level, belonging to 2,326 unique samples which collected at various weeks of gestation throughout the pregnancy, spread across 133 subjects of diverse race and ethnicities (Figure 2a), out of which 85 subjects delivered at term and 48 subjects delivered preterm (Figure 2 b, d). We aligned the abundance counts subject- and gestational week-wise for ease of interpretation **(**Figure 2c). Abundance counts for multiple samples derived from the same subject collected during the same week of gestation, if any, were replaced by the mean of those counts to ensure consistency, as future analyses were carried out on week-wise data. The distribution of number of genera present in each sample (i.e., number of genera with non-zero abundance in each sample) is visualized in Figure 2e. Microbial abundance profiles corresponding to 43 out of the 133 subjects were reserved as the test set. (refer methods Section 2.3).

3.2. Diversity metrics do not reliably identify at-risk PTB subjects

Alpha-diversity indices were computed on samples collected during trimester 1 and trimester 2 (gestational weeks 9 to 24, see methods Section 2.3). The visualization of Chao1, Gini, Shannon, Simpson and TCS alpha-diversity indices computed at various weeks of gestation for subjects who delivered at term and preterm is presented in Figure 3 a-e, respectively. The results of two-sided, independent *t*-tests to examine differences in diversity index values across term and preterm groups are presented in Table 1. Although *t*-test reveals a statistically significant difference in the Chao1 diversity index between term and preterm groups during gestational weeks 9-12 ($p = 0.005$) and 13-17 ($p = 0.02$), there are too few samples in the preterm group during these gestational periods to draw reliable conclusions (see Figure 3). No signature that can distinguish term/preterm birth is observable from these metrics. As can be seen from the plot, the alpha-diversity indices are not predictive of a preterm delivery outcome, and perform worse than random classifiers (prediction accuracy on test set < 50%, refer methods Section 2.4).

3.3. Statistical analyses indicate presence of signatures for PTB prediction in evolving microbiomes

To reduce the set of features used for the classification task, we removed the genera with high skewness and kurtosis. Skewness and kurtosis were computed on relative abundance of microbial genera during the period of trimester 1 - trimester 2 (gestational weeks 9 to 24). The abundance distributions of a large number of genera have a high positive skewness and kurtosis (see Figure 4 a, b), and may contribute to noise. Thus, we excluded the genera which had highly skewed relative abundances (i.e., skewness > 10) as well as genera with high kurtosis (kurtosis > 10). As a result, only 6 genera were retained, namely, Lactobacillus, Anaerococcus, Gardnerella, Peptoniphilus, Finegoldia and Prevotella. Except for Finegoldia, these genera have been identified to be linked to PTB risk previously. Lactobacillus is the most dominant genus within the vaginal microbiome, and low counts of Lactobacillus have previously been stated to be indicative of increased PTB risk (Bayar et al., 2020; Gudnadottir et al., 2022). Certain species of Anaerococcus are found to be associated with increased PTB risk (Ansari et al., 2021), however, there are also reports that Anaerococcus species may be protective in nature (dos Anjos Borges et al., 2023). There is strong evidence linking high Gardnerella vaginalis presence with PTB and bacterial vaginosis, which also increases PTB risk (Nelson et al., 2009; Ng et al., 2023). In some populations, increased counts of certain Peptoniphilus species were also found to be associated with high PTB risk (Park et al., 2022a). Similar evidence exists associating high Prevotella abundances with increased PTB risk (Freitas et al., 2018; Fettweis et al., 2019; Park et al., 2022b). However, there is insufficient evidence to establish a link between these observations and the changes that vaginal microbiota undergo throughout the duration of the pregnancy.

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
Copyright © 2023 the authors This is an Open Access work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

We further re-computed the relative abundance based on the abundance counts of these 6 genera only, and visualized the gestational week-wise abundances and corresponding term/preterm delivery outcomes. The results are presented in Figure 5 a-f. This analysis highlights that the composition of the 6 genera mentioned above, changes significantly during the period from end of trimester 1 to trimester 2 of pregnancy. The trends indicate that low abundance of the Lactobacillus genus during trimester 2 (gestational
Gardnerella counts with PTB risk (Figure 5c, $p = 0.0073$), and the effect is more pronounced during gestational weeks 16-18. Most of the samples with increased Prevotella counts during weeks 19-21 belonged to subjects who went on to deliver preterm (Figure 5e, $p \approx 10^{-6}$).

3.4. ML classifiers do not make adequate PTB risk assessment

We tested the performance of two ML classifiers, viz., RF and DT towards prediction of preterm birth. For this, we isolated the latest available microbial abundance profiles of training set subjects as well as the test set subjects, collected during the period between the 9th and the 24th week of gestation, with the rationale that the state of the microbiome closer to the period of delivery should be better predictive of term/preterm delivery. Optimal hyperparameters for both RF and DT were identified by performing a grid search on pre-defined parameter search spaces using 3-fold cross-validation on the training set, and are listed in supplementary Table 2. The resultant models were validated on the test set by computing ROC-AUC, accuracy, precision-recall and f1 score. The RF model performed significantly better (Figure 6 b, d) compared to DT (Figure 6 a, c) which performs worse than a random predictor. The detailed results of both classifiers are presented in Table 2 Neither of the models, however, make adequately reliable predictions on the test set.

3.5. LSTMs do not decode the temporal dynamics of vaginal microbiomes

LSTM was trained on the week-wise genera abundance data sampled during gestational weeks 9 to 24, after making appropriate adjustments to account for the irregularity in sampling (see Methods, Section 2.6.1, Figure 1). When trained on the entire set of genera with non-zero variance in the training set, the model overfits and does not generalize well to making predictions on the test set (training set accuracy = 100%, test set accuracy < 60%). Even when trained on the set of genera with low skewness and kurtosis, the model fails to make sufficiently accurate predictions on the test set (accuracy = 63%), and is outperformed by the RF model described above. We attribute this lack of predictivity to the increased estimations that the model makes to fill the temporal gaps in the data, and it may necessitate availability of additional data samples to be able to make these estimations more accurately, either in terms of more patient subjects or increased density of samples per subject.

3.6. Neural CDEs are capable of achieving PTB prediction with a substantial accuracy

The neural CDE model trained on relative abundances of genera selected by the skewness and kurtosis filtering outperforms all other models described above. The resultant model performs reasonably well on the test set (mean test set ROC-AUC = 0.82, accuracy = 74.5% precision = 0.65, recall = 0.71, f1 score = 0.71). The results are presented in Figure 7 a, b. We then trained the model after shuffling the term/preterm labels for subjects in the training set. As expected, the model behaved similar to a random predictor on the test set (ROC-AUC < 0.5). Albeit not on the same validation dataset, our approach describes better results than the best submission in the DREAM challenge for term vs preterm prediction (ROC-AUC = 0.68, accuracy = 67%, sensitivity = 0.48, specificity = 0.79) (Golob et al., 2023), despite using microbial abundances only up to the end of the 2nd trimester, i.e., the 24th week of gestation, as opposed to the 32nd week of gestation in the DREAM challenge.

Copyright © 2023 the authors
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
Open Access

4. Discussion

Current in vitro or in vivo approaches lack the ability to detect PTB incidences at earlier stages reduces the effectiveness of prophylactic or therapeutic interventions that can be administered to mitigate neonatal health concerns associated with it. Current risk assessment approaches involve physical examinations, and factors such as cervical length may be used to estimate the risk. Additionally, abnormality in levels of biochemical markers such as Pregnancy-Associated Protein A (PAPP-A) (Smith et al., 2002; Gundu et al., 2016), Cervicovaginal Interleukins (Manning et al., 2019; Park et al., 2020), etc., may help detect PTB as well. However, there is a lack of definitive confidence intervals for these physical or biochemical tests. Recent studies have highlighted the utility of diversity of vaginal microbial communities towards PTB prediction, by establishing correlations between alpha-diversity indices associated with abundances of various microbial species (or genera) and incidences of PTB (DiGiulio et al., 2015; Hyman et al., 2014; Haque et al., 2017). However, microbial communities are highly diverse across various individuals, and more so for individuals belonging to different ethnic populations (Sun et al., 2022; Gupta et al., 2017). We have demonstrated above, that in a heterogenous dataset, with microbial profiles derived from ethnically and racially diverse subjects, diversity metrics could not accurately estimate PTB risk. This indicates that previously reported success of diversity indices in identifying subjects at high risk of PTB may be dataset dependent, either with respect to the subject cohort or to the pipeline used in computation of microbial abundance from 16S rRNA sequences, based on our observations on a mixed-race dataset with 16S rRNA sequences transformed using a standardized method.

Traditional ML methods, which have been explored previously in the context of predicting PTB using vaginal microbial species abundance, fail to learn an abundance signature associated with PTB in our dataset. While others (Park et al., 2022a) report success with machine learning methods in ethnically homogenous cohorts, we found that the predictive performance did not translate to mixed-race cohorts. As we have demonstrated above, vaginal microbial communities evolve throughout the duration of pregnancy, and abundance levels of certain species or genera may change significantly as the pregnancy progresses. Learning a PTB-associated signature in an evolving microbiome may be out of scope of such models as they are not designed to handle time-series datasets. We explored the utility of LSTM, a type of RNN, which is able to work with sequential datasets. The architecture of RNN-based approaches requires input datasets to be regularly and continuously sampled. As far as human patient subject data is concerned, obtaining such a dataset is a challenge, as study subjects may not be regular or consistent in clinical visits. For this purpose, we filtered the dataset such that a single sample was present across each of the gestational weeks, which constituted the time intervals for LSTM. We suitably modified the LSTM workflow to overcome missing time intervals, however it proved to be incapable of learning any signature associated with PTB.

Neural differential equations have recently gained traction with regards to analyzing sequential data. Since it uses differential equations to model the temporal dynamics, it can handle irregularly and/or inconsistently sampled data. Neural CDEs are more efficient than neural ODEs (Kidger et al., 2020), and are even capable of working with partially sampled datasets, although we have not harnessed that in this study. We found considerable success in using Neural CDEs to predict PTB, in spite of working with a dataset sourced from an ethnically varied population (refer Figure 2), and outperformed all other approaches that we tested. To the best of our knowledge, this is the first effort towards modeling the temporal dynamics of vaginal microbial communities using deep learning, and the first instance of applying neural differential equations for a problem of this kind.

The DREAM challenge for PTB prediction (Golob et al., 2023), was issued in 2019 with the goal of driving efforts for PTB prediction using the vaginal microbiome. One of the sub-problems for the DREAM challenge consisted of predicting term births (>= 37 weeks of gestation) and preterm births (< 37 weeks of gestation) using vaginal microbiomes. The dataset for this challenge was derived from 9 different studies, and amounted to 3578 samples collected from 1268 individuals (Golob et al., 2023). The dataset used in this study was also part of the DREAM challenge. We also considered using some of the other datasets in the DREAM challenge while outlining this study, but dropped either due to not being labelled week-wise or due to insufficient week-wise samples per patient, for modeling the temporal dynamics. Most of the top submissions in the challenge used tree-based classifiers. On our test dataset, neural CDEs show better predictivity (mean test set ROC-AUC = 0.82, accuracy = 75%, sensitivity = 0.71, specificity = 0.85) than the best submission in the DREAM challenge on their validation dataset (ROC-AUC = 0.69, accuracy = 67%, sensitivity = 0.48, specificity = 0.79) (Golob et al., 2023), despite using microbial abundances only up to the end of the 2nd

Copyright © 2023 the authors
This is an Open Access
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

trimester, i.e., the 24th week of gestation. The DREAM challenge for PTB prediction used taxonomic abundance data upto the 32nd week of gestation. Our emphasis was on early-stage PTB prediction, and we were able to achieve better predictive performance in spite of restricting the input data till the 2nd trimester.

Predictive approaches using data other than microbial communities also exist. For instance, Tarca et al. (Tarca et al., 2021) report the results of the DREAM challenge for PTB prediction using the maternal blood transcriptome and the proteome. The top performing models report better results than what was reported on microbial communities, with a ROC-AUC of 0.76 when proteomics data from weeks 27-33 was used. However, in another sub-challenge where early-stage data (weeks 17-22) was used, the top performing model had a ROC-AUC of 0.62. Obtaining blood transcriptomic or proteomic data may pose difficulties due to the involvement of invasive procedures requiring clinical expertise to perform. On the other hand, microbial abundance data is sourced from vaginal swabs, which can be obtained without invasive procedures, by patient subjects themselves. Several attempts have been made at predicting PTB using biochemical marker (Aung et al., 2019; Leow et al., 2020), however, obtaining such data may require regular clinical visits, and their viability in racially-diverse populations is unknown.

Poorer and remote parts of the world may even lack the medical infrastructure or presence of adequate facilities that are required for PTB assessment and prevention. For example, in remote areas in India, there are clinics called Anganwadis, which roughly translates to courtyard shelter. As of 2018, the Ministry of Women and Child Development reports the existence of 1.4 million Anganwadi centres spread out across the country. Anganwadis provide limited healthcare facilities for maternal and infant health and lack the funding and facilities, or even trained medical personnel required to mitigate PTB and its ill-effects. Given the simplicity of obtaining samples from which microbial abundance is derived, reliable approaches for PTB risk assessments developed on microbiota will greatly help such remote clinics.

While we have demonstrated the capability of Neural CDEs towards PTB prediction using the vaginal microbiome, further effort can be made for increasing its clinical viability. Firstly, our dataset is limited in size (133 patient subjects), and we believe that larger datasets with better racial and ethnic representation may help learn signatures which take into account the diversity of vaginal microbiomes across individuals/races. Secondly, predicting the extent of preterm birth (extremely preterm, very preterm, moderate to late preterm) is also important as far as administering interventions is concerned, as they may have varying impact on maternal and infant health and may require different strategies. This may be achieved by predicting the gestational week of delivery, or by treating PTB as a multi-class problem with different extents of PTB as the classes, on more high-quality datasets. We strongly believe that vaginal microbial communities may be the key to achieving early-stage PTB prediction, and our findings strongly encourage future efforts for pregnancy microbiome data generation and further refinements in modeling procedures, which may take us closer to achieving full clinical viability.

5. Data & Code Availability

The code and data files for this study have been made available on https://tinyurl.com/3427p6y4. The repository contains a readme file which describes the contents of the individual data files along with the demographics of the train and test data, as well as a brief description of the code files.

6. Declarations

6.1. Competing Interests

The authors declare no competing interests

6.2. Funding

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

6.3. Ethical Approval

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
Copyright © 2023 the authors This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

The data used in this study was derived from a public, non-controlled access dataset from within the Sequence Read Archive (SRA), and hence, ethical approval was not required.

6.4. Author Contributions

M.B. and K.K. designed the study. K.K. performed the computational analyses along with assistance from M.B., and M.B. and K.K drafted the final manuscript.

6.5. Acknowledgements

The authors would like to thank Dr. Mohammed Haque, Dr. Anirban Dutta, and Mr. Nishal Kumar Pinna for their help in processing the 16S rRNA sequences from the SRA cloud, and Mr. Sunil Nagpal and Dr. Anirban Dutta for outlining the state of the art for PTB prediction using diversity metrics. We would also like to thank Dr. Rajgopal Srinivasan for proofreading and helping improve the manuscript.

Figure 1: Visualization of the LSTM Network. Week 0 represents the earliest week of gestation for which microbial data is available. This is passed through an embedding layer, which generates the initial hidden state. Subsequent hidden states are determined by the previous hidden and cell states, and the microbial data at the respective time step. If the microbial data is not available, the output of the previous hidden state is used instead. The final hidden state is passed through a linear layer with a single output neuron to predict the term/preterm outcome, and the entire network is trained end-to-end.

Figure 2: Distribution of (a) race, (b) number of subjects who delivered at term/pre-term, (c) sample gestational age at the time of collection, (d) gestational age of subject at the time of delivery, and (e) number of unique genera detected in each sample in the microbial abundance dataset

Figure 3: (a) Chao1, (b) Gini, (c) Shannon, (d) Simpson and (e) TCS diversity metrics computed on microbial abundance data collected during trimester 1 and trimester 2 of pregnancy. Blue and red points represent samples derived from subjects who delivered at term and pre-term, respectively.

Figure 4: Distribution of (a) skewness and (b) kurtosis: computed on relative abundances of genera.

Figure 5: (a)-(f): Variation in relative abundances for Anaerococcus, Finegoldia, Gardnerella, Lactobacillus, Peptoniphilus and Prevotella, respectively, during gestational weeks 9-24.

Figure 6: Receiver operating characteristic (ROC) curve for the (a) decision tree (DT) and (b) random forest (RF) classifiers on the training and test datasets, and validation metrics (AUC, accuracy, precision, recall and f1 score) computed for the (c) DT and (d) RF classifiers.

Figure 7: (a) ROC curves and (b) classification metrics on the training and test datasets for the Neural CDE model.

Table 1: *p*-values representing significance of differences between diversity metric values across the term and preterm groups for gestational weeks (a) 9-12, (b) 13-16, (c) 17-20 and (d) 21-24.

Table 2: Comparative performance of various machine learning methods

References

Ansari, A. et al. (2021), 'Molecular mechanism of microbiota metabolites in preterm birth: Pathological and therapeutic insights', *International Journal of Molecular Sciences* **22**(15), 8145.
URL: *https://doi.org/10.3390/ijms22158145*

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This is an Open Access work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Copyright © 2023 the authors

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

Aung, M. T. et al. (2019), 'Prediction and associations of preterm birth and its subtypes with eicosanoid enzymatic pathways and inflammatory markers', *Scientific Reports* **9**(1).
URL: *https://doi.org/10.1038/s41598-019-53448-z*

Baranwal, M. et al. (2022), 'Recurrent neural networks enable design of multifunctional synthetic human gut microbiome dynamics', *eLife* **11**.
URL: *https://doi.org/10.7554/elife.73870*

Bayar, E. et al. (2020), 'The pregnancy microbiome and preterm birth', *Seminars in Immunopathology* **42**(4), 487–499.
URL: *https://doi.org/10.1007/s00281-020-00817-w*

Bharti, R. et al. (2019), 'Current challenges and best-practice protocols for microbiome analysis', *Briefings in Bioinformatics* **22**(1), 178–193.
URL: *https://doi.org/10.1093/bib/bbz155*

Blencowe, H. et al. (2012), 'National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications', *The Lancet* **379**(9832), 2162–2172.
URL: *https://doi.org/10.1016/s0140-6736(12)60820-4*

Brown, R. G. et al. (2019), 'Establishment of vaginal microbiota composition in early pregnancy and its association with subsequent preterm prelabor rupture of the fetal membranes', *Translational Research* **207**, 30–43.
URL: *https://doi.org/10.1016/j.trsl.2018.12.005*

Callahan, B. J. et al. (2016), 'DADA2: High-resolution sample inference from illumina amplicon data', *Nature Methods* **13**(7), 581–583.
URL: *https://doi.org/10.1038/nmeth.3869*

Callahan, B. J. et al. (2017), 'Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women', *Proceedings of the National Academy of Sciences* **114**(37), 9966– 9971.
URL: *https://doi.org/10.1073/pnas.1705899114*

Chawanpaiboon, S. et al. (2019), 'Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis', *The Lancet Global Health* **7**(1), e37–e46.
URL: *https://doi.org/10.1016/s2214-109x(18)30451-0*

Chen et al. (2018), 'Neural ordinary differential equations', *arXiv* .
URL: *https://arxiv.org/abs/1806.07366*

Chung, E. H. et al. (2020), 'Neurodevelopmental outcomes of preterm infants: a recent literature review', *Translational Pediatrics* **9**(S1), S3–S8.
URL: *https://doi.org/10.21037/tp.2019.09.10*

DiGiulio, D. B. et al. (2015), 'Temporal and spatial variation of the human microbiota during pregnancy', *Proceedings of the National Academy of Sciences* **112**(35), 11060–11065.
URL: *https://doi.org/10.1073/pnas.1502875112*

dos Anjos Borges, L. G. et al. (2023), 'Vaginal and neonatal microbiota in pregnant women with preterm premature rupture of membranes and consecutive early onset neonatal sepsis', *BMC Medicine* **21**(1).
URL: *https://doi.org/10.1186/s12916-023-02805-x*

Edgar, R. C. (2013), 'UPARSE: highly accurate OTU sequences from microbial amplicon reads', *Nature Methods* **10**(10), 996–998.
URL: *https://doi.org/10.1038/nmeth.2604*

Estaki, M. et al. (2020), 'QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data', *Current Protocols in Bioinformatics* **70**(1).
URL: *https://doi.org/10.1002/cpbi.100*

Fettweis, J. M. et al. (2019), 'The vaginal microbiome and preterm birth', *Nature Medicine* **25**(6), 1012–1021.
URL: *https://doi.org/10.1038/s41591-019-0450-2*

Freitas, A. C. et al. (2018), 'Increased richness and diversity of the vaginal microbiota and spontaneous preterm birth', *Microbiome* **6**(1).
URL: *https://doi.org/10.1186/s40168-018-0502-8*

Fung, D. L. X. et al. (2023), 'A self-knowledge distillation-driven CNN-LSTM model for predicting disease outcomes using longitudinal microbiome data', *Bioinformatics Advances* **3**(1).

Copyright © 2023 the authors
This is an Open Access work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

URL: https://doi.org/10.1093/bioadv/vbad059
Georgiou, H. M. et al. (2015), 'Predicting preterm labour: Current status and future prospects', *Disease Markers* 2015, 1–9.
URL: *https://doi.org/10.1155/2015/435014*
Golob, J. L. et al. (2023), 'Microbiome preterm birth DREAM challenge: Crowdsourcing machine learning approaches to advance preterm birth research', *Preprint* .
URL: *https://doi.org/10.1101/2023.03.07.23286920*
Gordon, H. A. et al. (1971), 'The gnotobiotic animal as a tool in the study of host microbial relationships', *Bacteriological Reviews* **35**(4), 390–429.
URL: *https://doi.org/10.1128/br.35.4.390-429.1971*
Gudnadottir, U. et al. (2022), 'The vaginal microbiome and the risk of preterm birth: a systematic review and network meta-analysis', *Scientific Reports* **12**(1).
URL: *https://doi.org/10.1038/s41598-022-12007-9*
Gundu, S. et al. (2016), 'Correlation of first-trimester serum levels of pregnancy-associated plasma protein a with small-for-gestational-age neonates and preterm births', *International Journal of Gynecology & Obstetrics* **133**(2), 159–163.
URL: *https://doi.org/10.1016/j.ijgo.2015.09.022*
Gupta, V. K. et al. (2017), 'Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity', *Frontiers in Microbiology* **8**.
URL: *https://doi.org/10.3389/fmicb.2017.01162*
Haque, M. M. et al. (2017), 'First-trimester vaginal microbiome diversity: A potential indicator of preterm delivery risk', *Scientific Reports* **7**(1).
URL: *https://doi.org/10.1038/s41598-017-16352-y*
Hyman, R. W. et al. (2014), 'Diversity of the vaginal microbiome correlates with preterm birth', *Reproductive Sciences* **21**(1), 32–40.
URL: *https://doi.org/10.1177/1933719113488838*
Katz, K. et al. (2021), 'The sequence read archive: a decade more of explosive growth', *Nucleic Acids Research* **50**(D1), D387–D390.
URL: *https://doi.org/10.1093/nar/gkab1053*
Kidger, P. et al. (2020), 'Neural controlled differential equations for irregular time series'.
URL: *https://arxiv.org/abs/2005.08926*
Leow, S. M. et al. (2020), 'Preterm birth prediction in asymptomatic women at mid-gestation using a panel of novel protein biomarkers: the prediction of PreTerm labor (PPeTaL) study', *American Journal of Obstetrics & Gynecology* MFM **2**(2), 100084.
URL: *https://doi.org/10.1016/j.ajogmf.2019.100084*
Lyons, T. (1994), 'Differential equations driven by rough signals (i): An extension of an inequality of lc young', *Mathematical Research Letters* **1**(4), 451–464.
Lyons, T. J. et al. (2007), Differential equations driven by rough paths, *Springer*.
MacIntyre, D. A. et al. (2015), 'The vaginal microbiome during pregnancy and the postpartum period in a european population', *Scientific Report*s **5**(1).
URL: *https://doi.org/10.1038/srep08988*
Manning, R. et al. (2019), 'Predictive value of cervical cytokine, antimicrobial and microflora levels for pre-term birth in high-risk women', *Scientific Reports* **9**(1).
URL: *https://doi.org/10.1038/s41598-019-47756-7*
Medina, R. H. et al. (2022), 'Machine learning and deep learning applications in microbiome research', *ISME Communications* **2**(1).
URL: *https://doi.org/10.1038/s43705-022-00182-9*
Nelson, D. B. et al. (2009), 'Preterm labor and bacterial vaginosis-associated bacteria among urban women', *Journal of Perinatal Medicine* **37**(2).
URL: *https://doi.org/10.1515/jpm.2009.026*
Newnham, J. P. et al. (2014), 'Strategies to prevent preterm birth', *Frontiers in Immunology* **5**.
URL: *https://doi.org/10.3389/fimmu.2014.00584*
Ng, B. K. et al. (2023), 'Maternal and fetal outcomes of pregnant women with bacterial vaginosis', *Frontiers in Surgery* **10**.
URL: *https://doi.org/10.3389/fsurg.2023.1084867*

Copyright © 2023 the authors
This is an Open Access
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

Park, S. et al. (2020), 'Cervicovaginal fluid cytokines as predictive markers of preterm birth in symptomatic women', *Obstetrics & Gynecology Science* **63**(4), 455–463.
URL: *https://doi.org/10.5468/ogs.19131*

Park, S. et al. (2022a), 'Predicting preterm birth through vaginal microbiota, cervical length, and WBC using a machine learning model', *Frontiers in Microbiology* **13**.
URL: *https://doi.org/10.3389/fmicb.2022.912853*

Park, S. et al. (2022b), 'Ureaplasma and prevotella colonization with lactobacillus abundance during pregnancy facilitates term birth', *Scientific Reports* **12**(1).
URL: *https://doi.org/10.1038/s41598-022-13871-1*

Pascanu et al. (2012), 'On the difficulty of training recurrent neural networks', *arXiv* .
URL: *https://arxiv.org/abs/1211.5063*

Paszke, A. et al. (2019), 'Pytorch: An imperative style, high-performance deep learning library', *arXiv* .
URL: *https://arxiv.org/abs/1912.01703*

Perin, J. et al. (2022), 'Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the sustainable development goals', *The Lancet Child & Adolescent Health* **6**(2), 106–115.
URL: *https://doi.org/10.1016/s2352-4642(21)00311-4*

Pfinder, M. et al. (2013), 'Preterm birth and small for gestational age in relation to alcohol consumption during pregnancy: stronger associations among vulnerable women? results from two large western-european studies', *BMC Pregnancy and Childbirth* **13**(1).
URL: *https://doi.org/10.1186/1471-2393-13-49*

Romero, R. et al. (2014), 'Erratum to: The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women', *Microbiome* 2(1).
URL: *https://doi.org/10.1186/2049-2618-2-10*

Schloss, P. D. et al. (2009), 'Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities', *Applied and Environmental Microbiology* **75**(23), 7537–7541.
URL: *https://doi.org/10.1128/aem.01541-09*

Sender, R. et al. (2016), 'Revised estimates for the number of human and bacteria cells in the body', *PLOS Biology* **14**(8), e1002533.
URL: *https://doi.org/10.1371/journal.pbio.1002533*

Smith, G. C. S. et al. (2002), 'Early pregnancy levels of pregnancy-associated plasma protein a and the risk of intrauterine growth restriction, premature birth, preeclampsia, and stillbirth', *The Journal of Clinical Endocrinology & Metabolism* **87**(4), 1762–1767.
URL: *https://doi.org/10.1210/jcem.87.4.8430*

Stock, S. J. et al. (2020), 'Maternal smoking and preterm birth: An unresolved health challenge', *PLOS Medicine* **17**(9), e1003386.
URL: *https://doi.org/10.1371/journal.pmed.1003386*

Sun, S. et al. (2022), 'Race, the vaginal microbiome, and spontaneous preterm birth', *mSystems* **7**(3).
URL: *https://doi.org/10.1128/msystems.00017-22*

Tarca, A. L. et al. (2021), 'Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth', *Cell Reports Medicine* **2**(6), 100323.
URL: *https://doi.org/10.1016/j.xcrm.2021.100323*

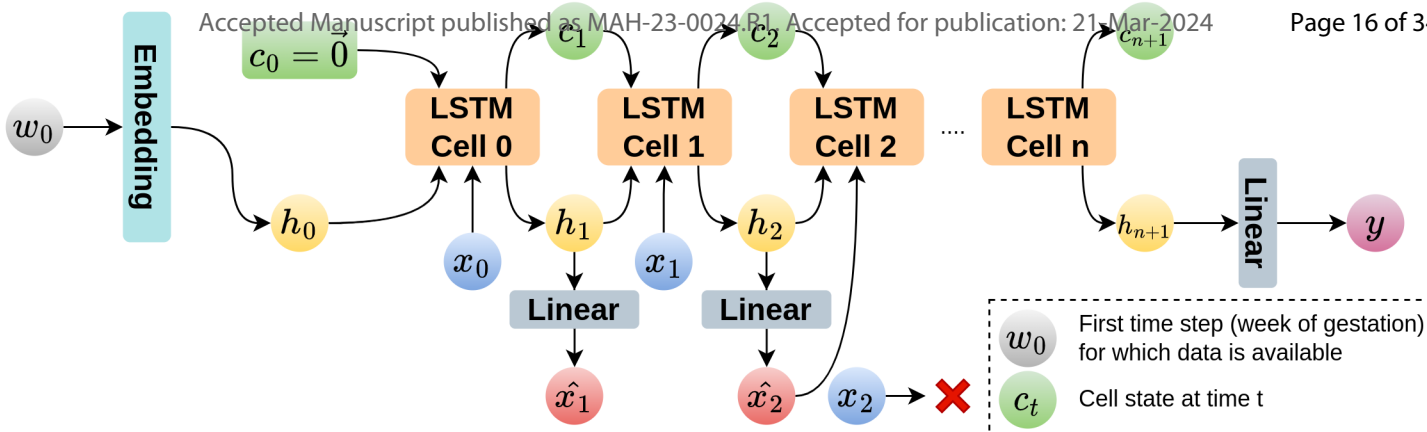
Townsi, N. et al. (2018), 'The impact of respiratory viruses on lung health after preterm birth', European *Clinical Respiratory Journal* **5**(1), 1487214.
URL: *https://doi.org/10.1080/20018525.2018.1487214*

Walani, S. R. (2020), 'Global burden of preterm birth', *International Journal of Gynecology & Obstetrics* **150**(1), 31–33.
URL: *https://doi.org/10.1002/ijgo.13195*

Weisburg, W. G. et al. (1991), '16s ribosomal DNA amplification for phylogenetic study', *Journal of Bacteriology* **173**(2), 697–703.
URL: *https://doi.org/10.1128/jb.173.2.697-703.1991*

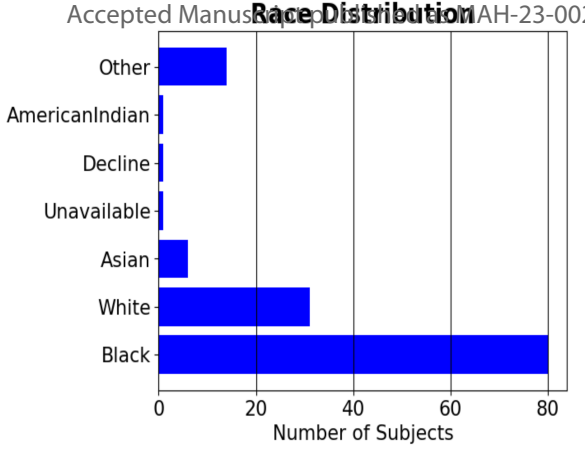
Yang, B. et al. (2016), 'Sensitivity and correlation of hypervariable regions in 16s rRNA genes in phylogenetic analysis', *BMC Bioinformatics* **17**(1).
URL: *https://doi.org/10.1186/s12859-016-0992-y*

Copyright © 2023 the authors
This is an Open Access
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

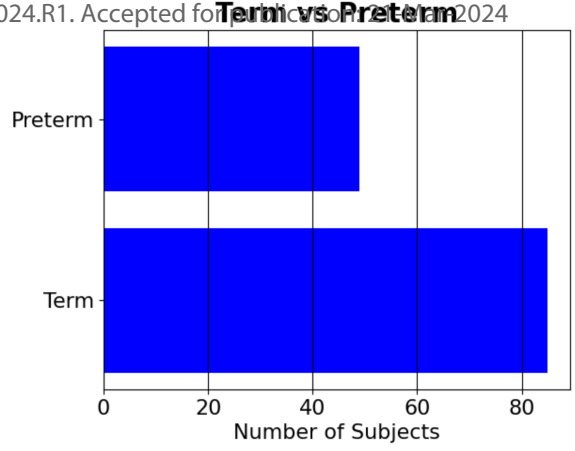


Copyright © 2023 the authors
Downloaded from Bioscientifica.com at 03/29/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License

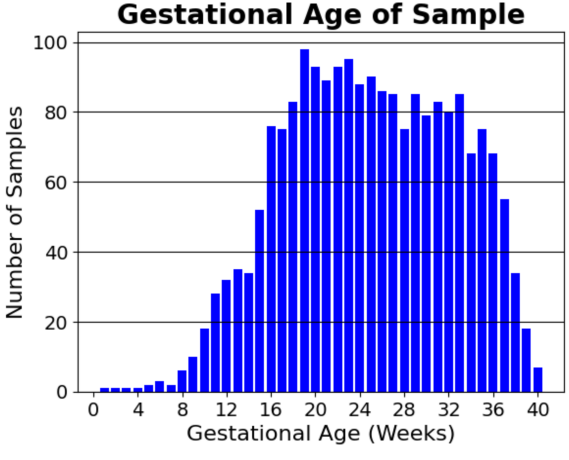
Accepted Manuscript published MAH-23-0024.R1. Accepted for publication 12 February 2024



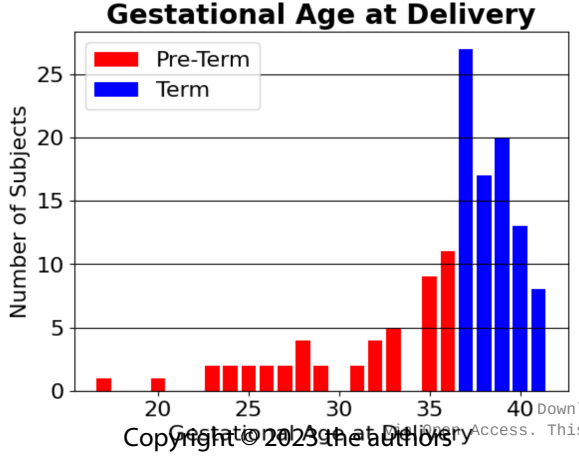
(a)



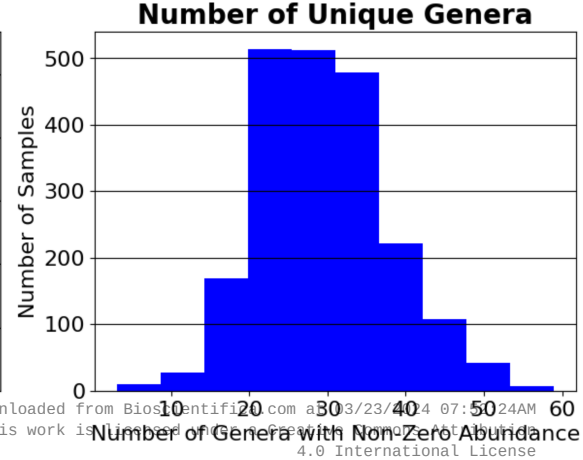
(b)



(c)

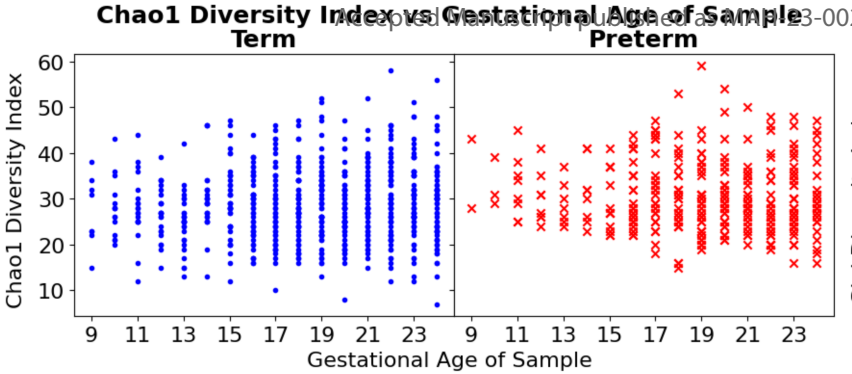


(d)

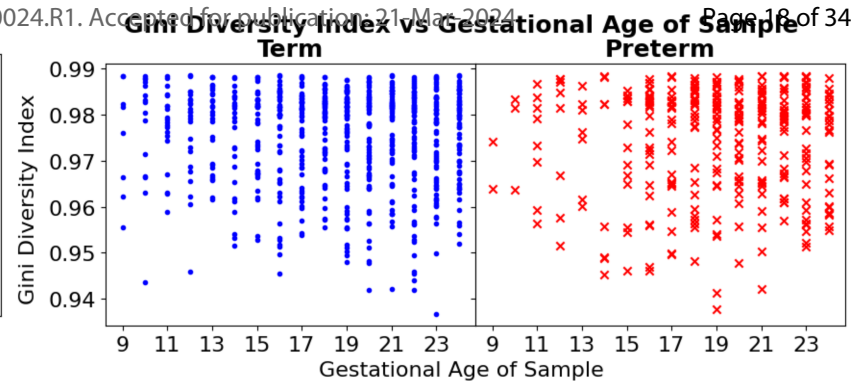


(e)

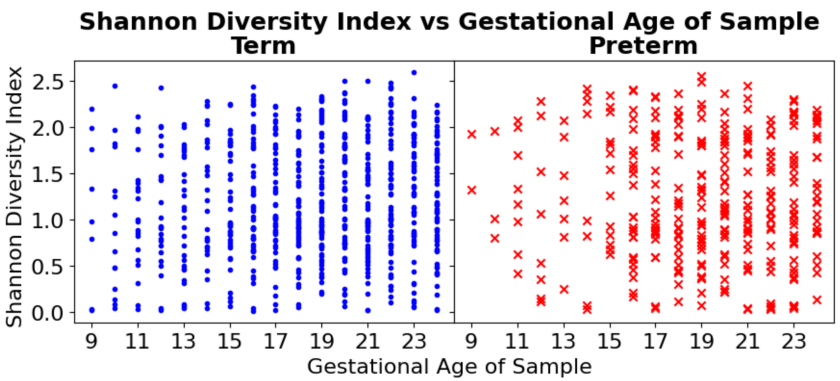
Copyright © 2023 the authors
Downloaded from BioScientifica.com at 03/23/2024 07:52:24AM
Open Access. This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB



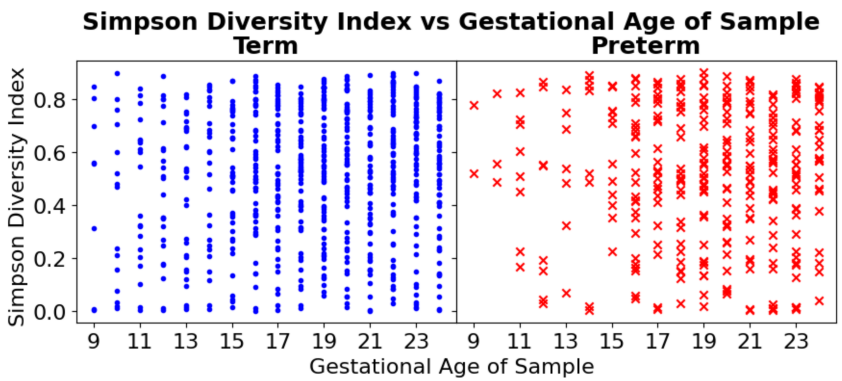
(a)



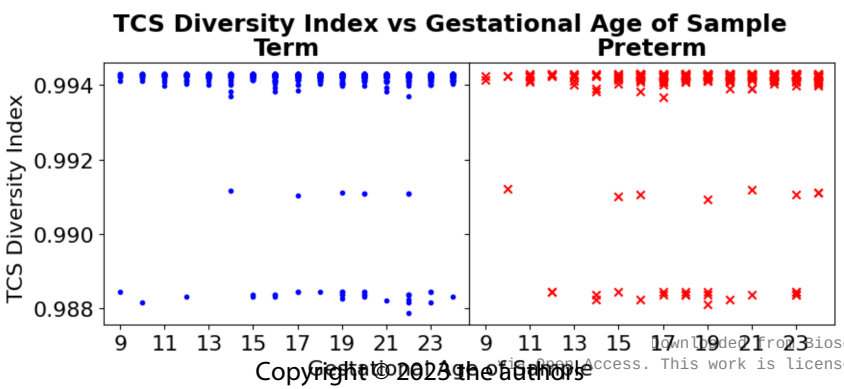
(b)



(c)



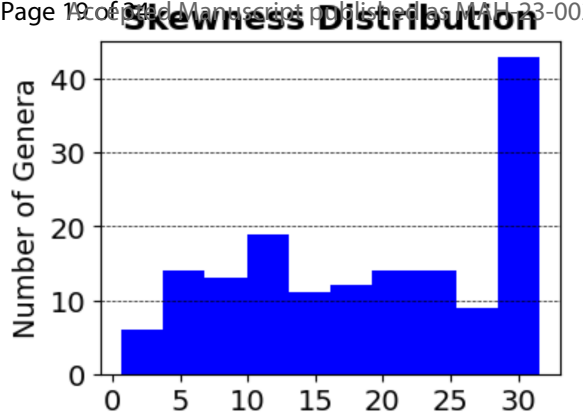
(d)



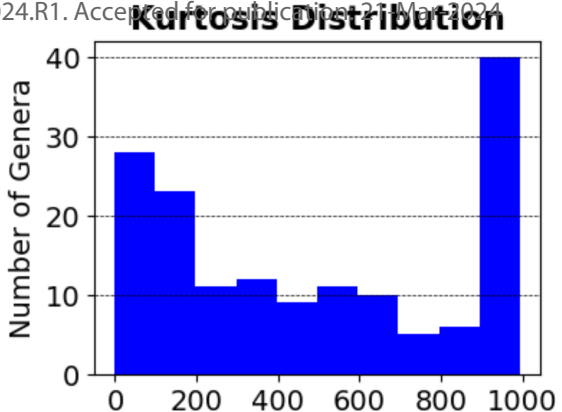
(e)

Copyright © 2023 the authors
Unauthenticated | Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
via Open Access. This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Accepted Manuscript published as MAU 23-0024.R1. Accepted for publication 21 Mar 2024

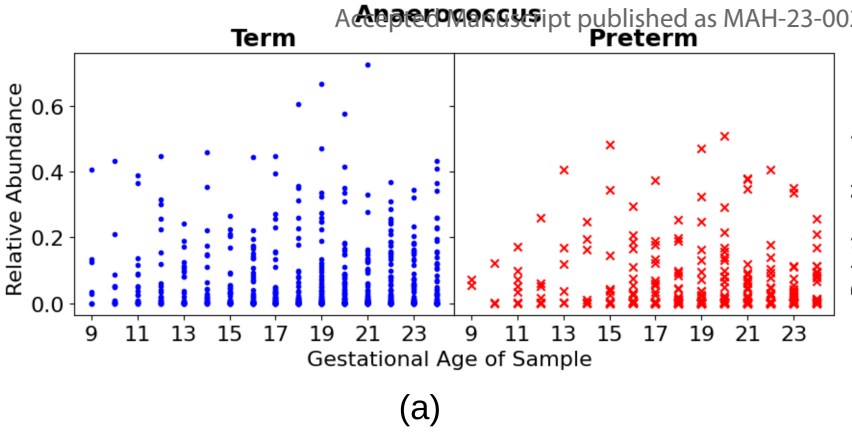


(a)

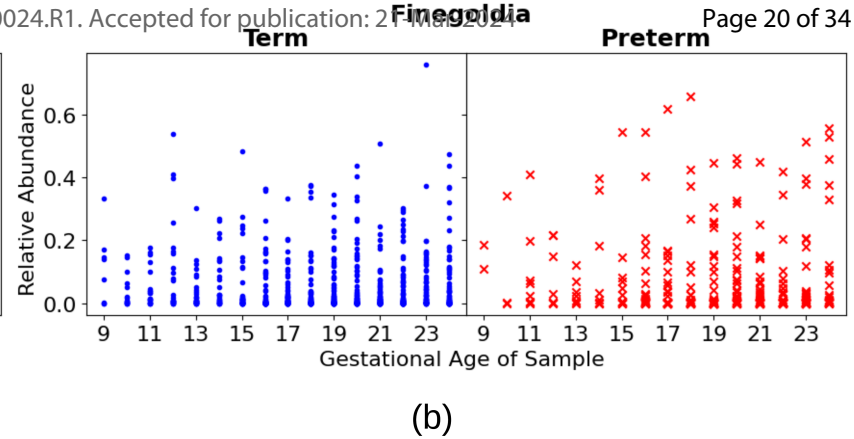


(b)

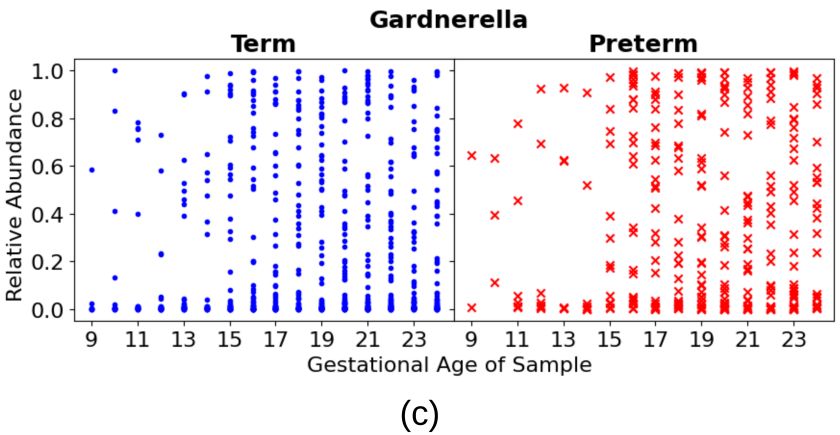
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
Copyright © 2023 the authors
Open Access. This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB



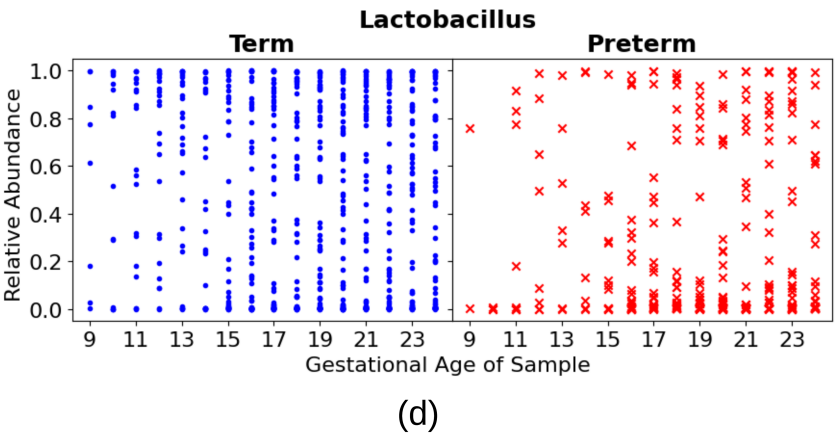
(a)



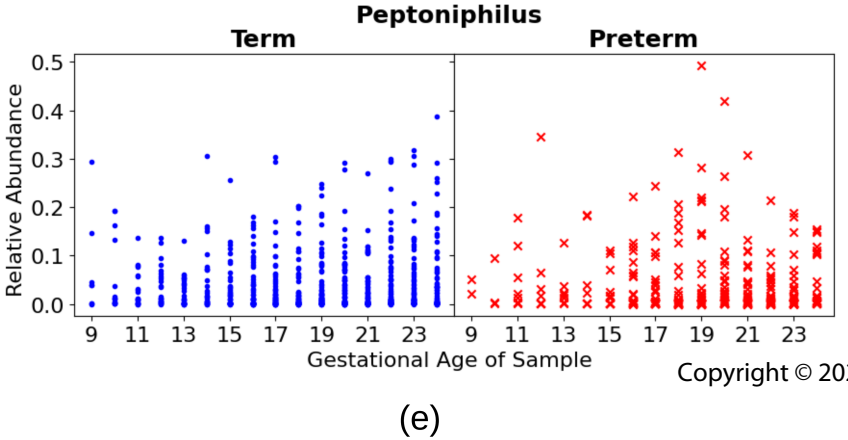
(b)



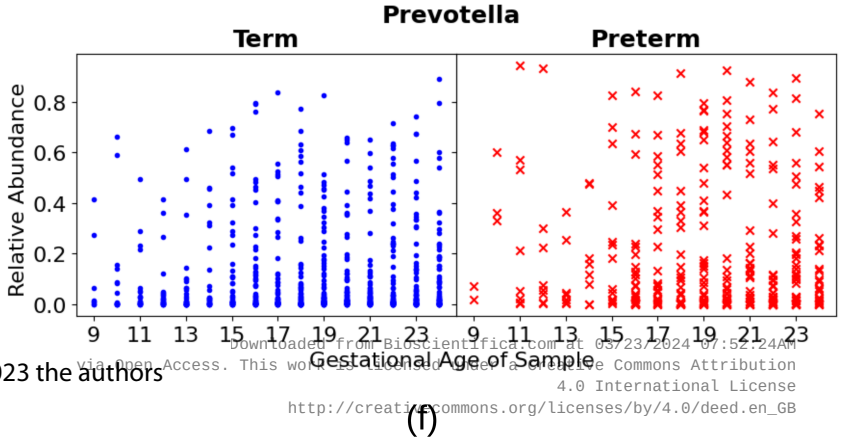
(c)



(d)



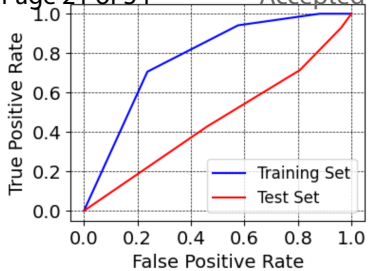
(e)



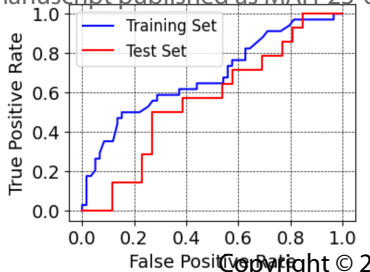
(f)

Copyright © 2023 the authors
Downloaded from Bioscientifica.com at 05/23/2024 07:52:24AM
via Open Access. This work is licensed under a Creative Commons Attribution
4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

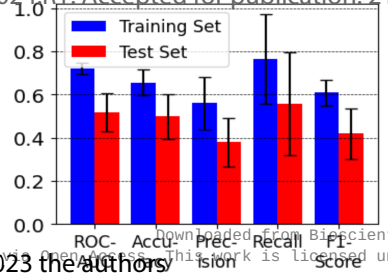
Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024



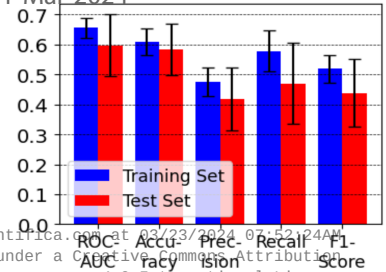
(a)



(b)



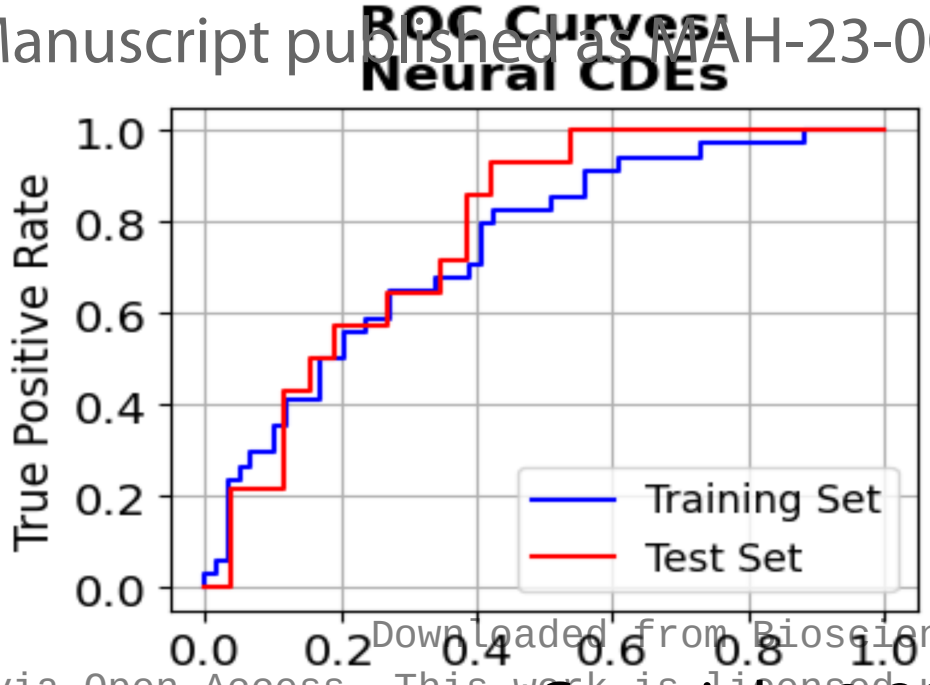
(c)



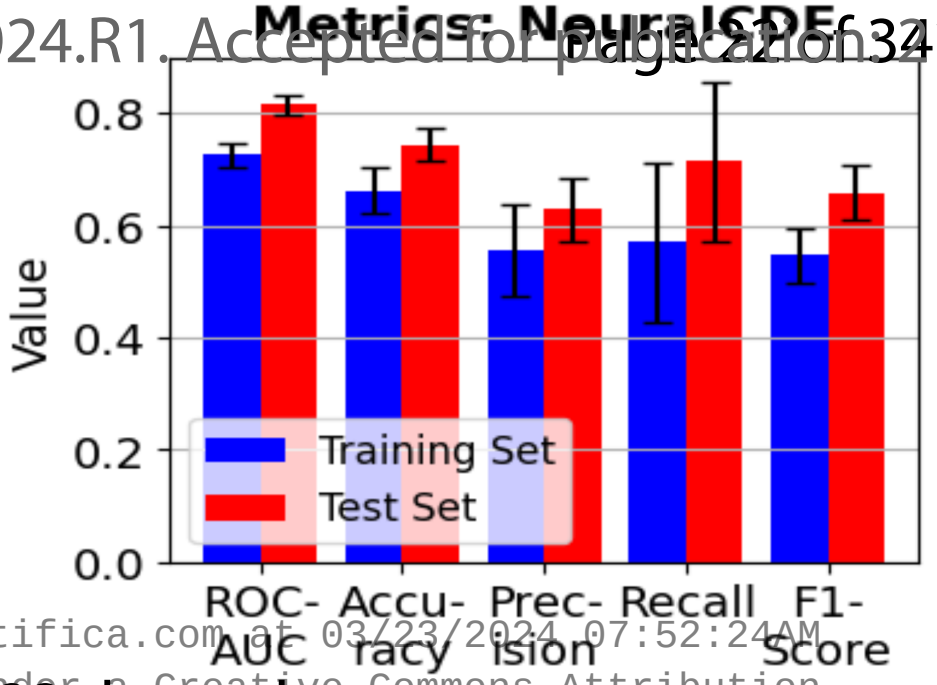
(d)

Copyright © 2023 the authors
Downloaded from Bioscientifica.com at 05/23/2024 03:52:24AM
via Open Access. This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Manuscript published as MAH-23-0024.R1. Accepted for publication 21-



(a)



(b)

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
via Open Access. This work is licensed under a Creative Commons Attribution
4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Copyright © 2023 the authors

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

Table 1: *P*-values representing significance of differences between diversity metric values across the term and preterm groups for gestational weeks 9-12, 13-16, 17-20 and 21-24.

Gestational week/Metric		***P*-value**
Gestation weeks 9 – 12		
	Shannon	0.8034
	Simpson	0.9271
	Gini	0.4074
	Chao1	0.0049
	TCS	0.0931
Gestation weeks 13 – 16		
	Shannon	0.7225
	Simpson	0.4616
	Gini	0.9042
	Chao1	0.0190
	TCS	0.5035
Gestation weeks 17 – 20		
	Shannon	0.6221
	Simpson	0.4852
	Gini	0.7739
	Chao1	0.3352
	TCS	0.8353
Gestation weeks 21 – 24		
	Shannon	0.3058
	Simpson	0.2270
	Gini	0.2935
	Chao1	0.5482
	TCS	0.3751

Copyright © 2023 the authors

This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

Method	Metrics				
	ROC-AUC	Accuracy	Precision	Recall	F1-score
Decision Tree	0.46 ± 0.09	0.45 ± 0.12	0.39 ± 0.11	0.58 ± 0.22	0.41 ± 0.14
Random Forest	0.60 ± 0.10	0.59 ± 0.09	0.42 ± 0.10	0.48 ± 0.13	0.44 ± 0.12
Ours (Neural-CDE)	0.82 ± 0.02	0.74 ± 0.04	0.65 ± 0.05	0.71 ± 0.12	0.71 ± 0.03

Table 2: Comparative performance of various machine learning methods

Copyright © 2023 the authors
This is an Open Access work licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

Deep Learning Enables Early-Stage Prediction of Preterm Birth Using Vaginal Microbiota

Supplementary Material

1 Machine Learning Classifiers: Hyperparameters

1.1 Hyperparameter search spaces

Hyperparameters for both, the Random Forest (RF) and Decision Tree (DT) models were tuned by carrying out 3-fold cross-validation on the training set. For each set of hyperparameters, the training set was divided into 3 folds, and classification performance was evaluated on each fold after training on the 2 remaining folds. Accuracy, ROC-AUC, precision, recall and F1-score was computed for each fold. Finally, the optimal hyperparameters were selected based on mean F1-score (i.e., the hyperparameter set with the highest mean F1-score was selected as the optimal set). The search spaces of the hyperparameters are presented in supplementary table 1. The explanation and significance of parameters can be found in the scikit-learn documentation.

1.2 Optimal Hyperparameter Values

The optimal values of the hyperparameters, as identified by grid search, are listed in supplementary table 2.

2. Long Short-Term Memory (LSTM):

2.1 Implementation details

To facilitate the LSTM model to make predictions on a non-continuous and irregularly sampled dataset, we made two modifications to it. Let w_0 be the first time step for which data is available. Firstly, the hidden state is initialized by an embedding layer, which takes w_0 as input, and the cell state is initialized as a zero vector. Since week 9 is the earliest possible time point for which data is used, we denote it as t = 0. The LSTM model is run from weeks 9 to 24 (t = 0 to t = 15). From t = 0 to t = w_0 - 1, the LSTM cells at each time step do not update the hidden state, and simply act as dummy cells. At t = w_0, the hidden state is still the same as what was initialized with the embedding, and the LSTM cell at t = w_0 updates the hidden state based on the input data.

The LSTM cell at each time step is followed by a linear layer which generates the “forecast”, or the predicted genera abundance at that time step. As part of our second modification, for the time steps beyond t = w_0 for which input data is missing, we use this forecast to update the hidden state. Since the missing sampling intervals are different for each subject, we have to process each subject with its own set of rules to update the model weights. Thus, the batch size is restricted to 1.

2.2 Model Parameters

We used the Adam optimizer implemented within pytorch for training the model. The model parameters for LSTM are listed in supplementary table 3.

3. Neural Controlled Differential Equations (CDEs)

3.1 Training using adjoint backpropagation

Neural differential equation models are solutions to a time-dependent system whose function is a neural network that can be learned. This neural network is typically a linear fully connected network, however, a non-linearity can be introduced by using a tanh or a relu activation. In case of CDEs, there exists a “hidden state” (h), essentially a parametrized version of the input data, and the ODE can be represented as:

Copyright © 2023 the authors
This is an Open Access work licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

$$\frac{dh}{dt} = f_\theta(h)\frac{dx}{dt}$$

Where f_θ is the neural network with parameters θ. Training the model essentially refers to update these parameters based on the gradient of the cost function (J) with respect to the parameter vector, scaled by a learning rate (η).

$$\theta_{t+1} = \theta_t + \eta\frac{dJ}{d\theta}$$

And the gradient of the cost function can be represented as:

$$\frac{dJ}{d\theta} = \int_0^t g(h, \theta)\, dt$$

Where g is the loss function (binary cross-entropy in our case). The above expression essentially maps this gradient to a scalar, using which the parameters are updated. This can be further represented as:

$$\frac{dJ}{d\theta} = \int_0^t \frac{dJ}{dh}\frac{\partial h}{\partial h}\frac{\partial h}{\partial \theta}dt$$

Adjoint sensitivity analysis or the adjoint state method essentially refers to computing the cost function gradient using the above expression in an efficient manner. The exact mathematical derivation for this is comprehensively described in Chen et al. [1].

3.2 Model parameters

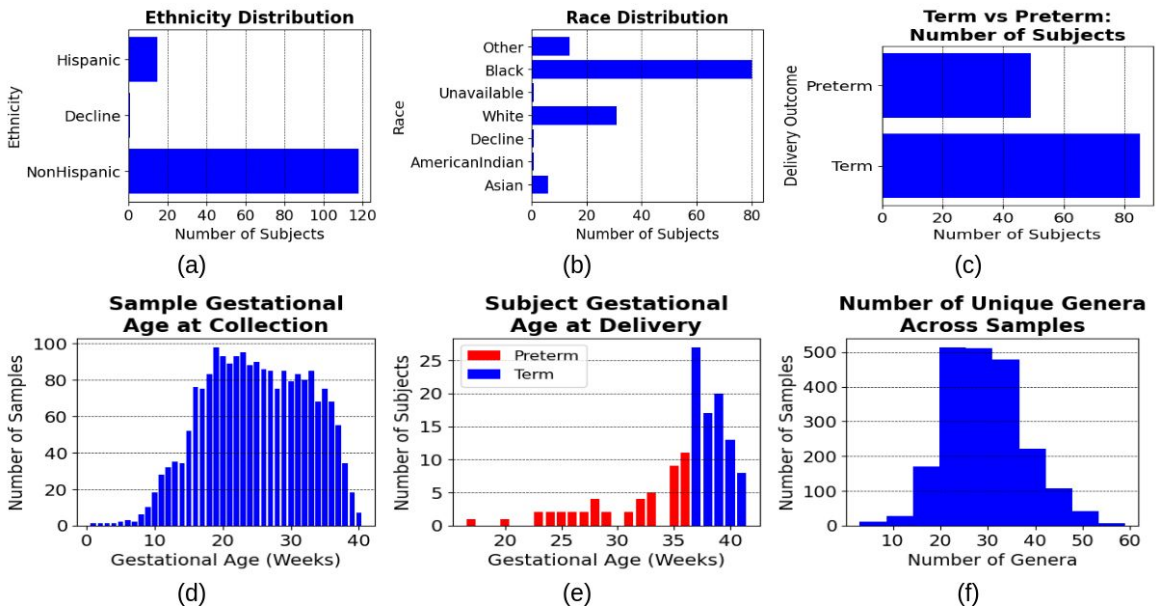
Neural CDE model parameters are listed in supplementary table 4.

Copyright © 2023 the authors This is an Open Access work is licensed under a Creative Commons Attribution 4.0 International License http://creativecommons.org/licenses/by/4.0/deed.en_GB

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

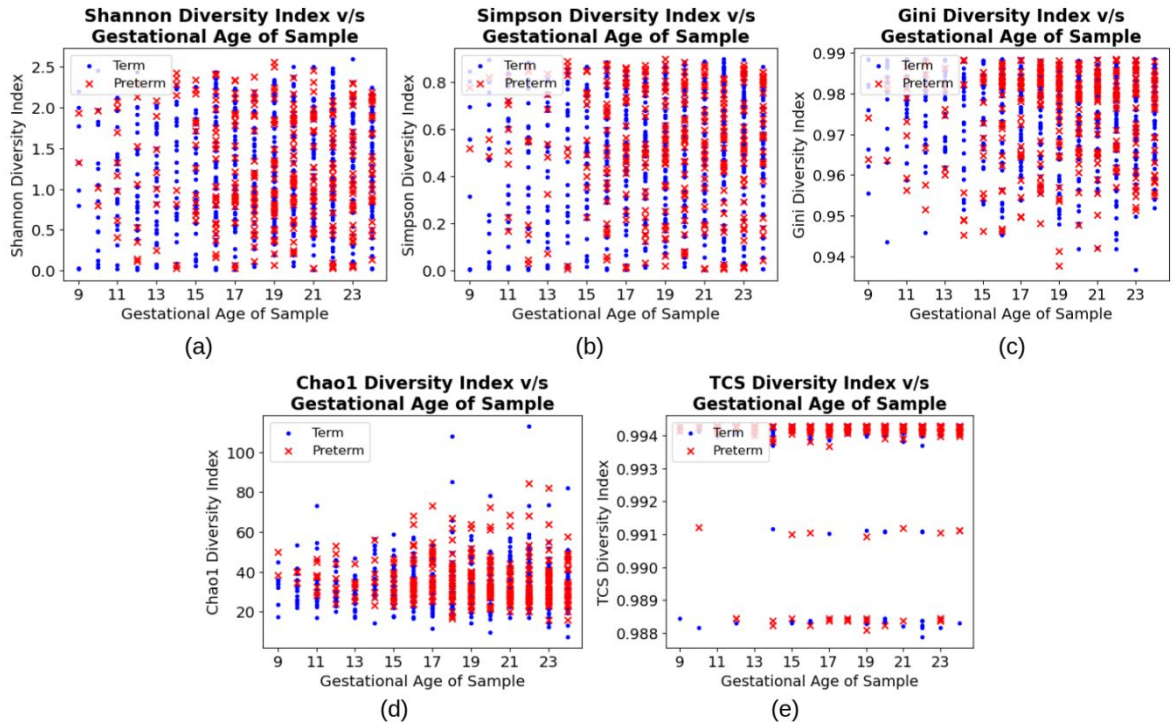
Supplementary Figures



Supplementary Fig. 1: Distribution of (a) Ethnicity, (b) Race, (c) Number of subjects who delivered at term/pre-term, (d) Sample gestational age at the time of collection, (e) Gestational age of subject at the time of delivery (f) Number of unique genera detected in each sample, in the microbial abundance dataset

Copyright © 2023 the authors

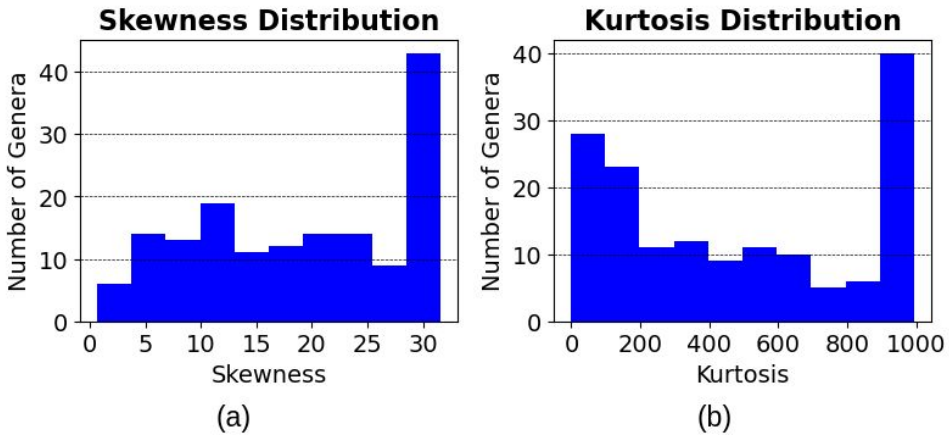
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This is Open Access. This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB



Supplementary Fig. 2: (a) Shannon, (b) Simpson, (c) Gini, (d) Chao1 and (e) TCS diversity metrics computed on microbial abundance data collected during trimester 1 and trimester 2 of pregnancy. Blue and red points represent samples derived from subjects who delivered at term and pre-term, respectively.

Copyright © 2023 the authors
This is an Open Access
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024



Supplementary Fig. 3: (a) Skewness and (b) Kurtosis distribution computed on the relative abundance of each microbial genus during the period from gestational weeks 9 to 24.

Copyright © 2023 the authors
This Open Access
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Supplementary Tables

Parameter	Search Spaces
n_estimators (RF only)	50, 100, 150
max_depth	None, 2, 3, 4
min_samples_split	0.25, 0.33, 0.5, 2, 4
max_features	sqrt, log2, 0.25, 0.33, 0.5, 2, 4
max_leaf_nodes	None, 2, 4
min_samples_leaf	0.25, 0.33, 0.5, 2, 4
criterion	entropy, gini, log_loss
splitter (DT only)	best, random
class_weight	balanced, balanced_subsample

Supplementary Table 1: Hyperparameter search spaces for the Random Forest and Decision Tree classifiers. The search spaces for both models are common unless specified otherwise.

Copyright © 2023 the authors
This Open Access
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

Parameter	Optimal Value	
	Random Forest	**Decision Tree**
n_estimators	50	-
max_depth	4	3
min_samples_split	0.33	0.5
max_features	0.25	0.25
max_leaf_nodes	2	4
min_samples_leaf	0.33	0.2
criterion	entropy	gini
splitter	-	best
class_weight	balanced_subsample	balanced

Supplementary Table 2: Optimal hyperparameter values for the Random Forest and Decision Tree classifiers.

Copyright © 2023 the authors
This Open Access
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM
This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Parameter	Value	Explanation
Input channels	30	Dimension of the input (number of genera in the input dataset)
Hidden channels	128	Dimension of the hidden state
Embedding input dimensions	14	Number of possible inputs to embedding layer (i.e., number of unique weeks between 9 and 24 which serve as first available data sample)
Learning rate	1e-4	Learning rate for updating weights
Learning rate decay	1 (No decay)	Reduction in learning rate across epochs
Loss function	Binary cross-entropy	The function to minimize while optimizing the LSTM model
Number of epochs	35	Number of epochs (iterations) for which the model is trained
Batch size	1 (Fixed)	Number of samples per training batch. Batch size is restricted to 1 due to different time steps for which data is missing, which forces handling each sample individually

Supplementary Table 3: Parameters for the long short-term memory model

Copyright © 2023 the authors

This is an Open Access work. This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB

Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

 Accepted Manuscript published as MAH-23-0024.R1. Accepted for publication: 21-Mar-2024

Parameter	Value	Explanation
Input channels	6	Dimension of the input (number of genera in the input dataset)
Hidden channels	64	Dimension of the hidden state
Output channels	1 (Fixed)	Dimension of the output (fixed to 1, since we are only outputting a single probability)
Interpolation type	Cubic	Method of interpolation of missing data from nearest observations
Learning rate	1e-4	Learning rate for updating weights
Learning rate decay	0.5	Reduction in learning rate across epochs
Learning rate decay step size	15	Number of epochs after which learning rate decay is applied
Loss function	Binary cross-entropy	The function to minimize while optimizing the LSTM model
Positive class weight	1.5	Weighted penalty to apply to negative class predictions (i.e., encourage predictions of positive class)
Batch size	10	Number of samples per training batch

Supplementary Table 4: Parameters for the Neural CDE model.

Copyright © 2023 the authors
This Open Access work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM

References

[1]. R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. ArXiV, 2018.

Copyright © 2023 the authors
This is Open Access - This work is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/deed.en_GB
Downloaded from Bioscientifica.com at 03/23/2024 07:52:24AM