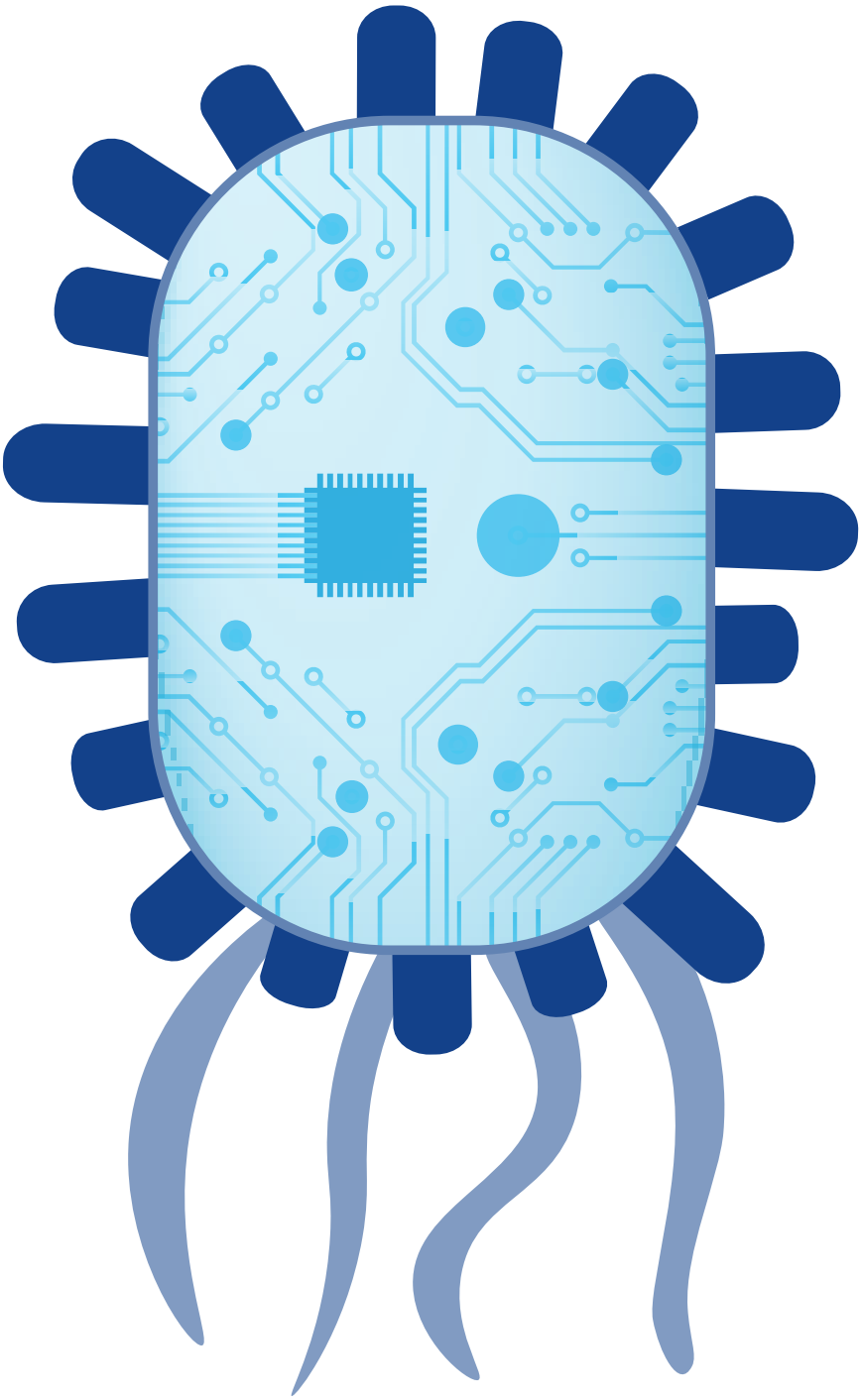
An evaluation of machine learning models for atopic health outcome prediction using infant gut microbiome data.



By Jip de Kok

Jip de Kok

j.dekok@student.maastrichtuniversity.nl

i6119367

Master Thesis

Maastricht University

System Biology

Faculty of Science and Engineering

Maastricht Centre for Systems Biology

First supervisor:

Prof. Ilja Arts

Second supervisor:

Assoc. Prof. John Penders

Practical supervisor:

David Barnett

Submission date: 28/05/2021

Word count: 14,283

Table of Contents

1. Abstract

Background: The infant gut microbiome is complex, highly dynamic, and influenced by many environmental aspects including the type of feeding, birth mode and antibiotics. Research shows that differences in the infant gut microbiome composition associate with development of atopic conditions such as asthma and eczema. Also, research is being done into the use of probiotics to prevent or reduce the severity of these conditions. Although asthma and eczema are incurable, accurate and early prediction of these diseases could greatly benefit disease control through early intervention. However, such predictive models do not yet exist.

Objective: This thesis project aimed at evaluating the predictive capacity of multiple machine learning algorithms on clinical and gut microbiome data of one-month-old infants to predict development of asthma and eczema.

Methods: Logistic regression, random forest, extreme gradient boosting, support vector machine, and sparse partial least squares discriminant analysis were used for predicting asthma and eczema. Hyperparameters were optimised through repeated 5-fold cross validation, models were trained on a training set (80%) and evaluated on a test set (20%). This was done separately on clinical data, microbiome data, and the combination of the clinical and microbiome data.

Results & conclusion: No predictive capacity was identified for asthma and eczema on the complete study population. Some predictive capacity was found when predicting asthma for infants with atopic parents using microbiome data (AUROC = 0.67). Nevertheless, prediction of asthma and eczema requires further investigation by including genetic information and preferably more cases.

2. Introduction

2.1 Background information

The incidence rate of atopic conditions such as asthma and eczema has been increasing for multiple decades, especially in children living in Western countries it has grown particularly high[1]. As atopy has become a public health concern, a lot of research has been done into the aetiology of atopic conditions. However, the process of developing atopic conditions remains largely unknown. Although, genetics are known to play some role in atopic development, they have not been able to explain which individuals will, or will not, develop some atopic conditions[2]. Furthermore, it is highly unlikely that genetic mutations have occurred at a rate that reflects the increase in prevalence of atopy[3]. Instead, it has been hypothesised that it is a change in environmental exposures in early life, that triggered the increasing incidence of atopy. To this end, the hygiene hypothesis was formulated, which suggests that infections in early life, that are triggered by various unhygienic environmental exposures, could have a protective effect against atopic conditions including asthma and eczema[4]. This could mean that the incidence rate of atopy has been so rapidly increasing, because genetically susceptible individuals have been less exposed to protective factors. Alternatively, it is also possible that these high-risk individuals have been more exposed to atopy inducing factors. From the hygiene hypothesis, another theory emerged. Namely, that the shift towards the more western and hygienic lifestyle caused alterations in the microbiome composition of the gastrointestinal tract, which in turn affects various immunological regulatory processes[5–12], and as such, could play an important role in the observed increase in prevalence of atopic conditions in western society.

The gut microbiome is a complex system, with 34 trillion bacteria residing in the gut of an average adult male[13]. Although, the gut microbiome of infants is not as rich as those of adults, they are very heterogeneous due to their unstable and dynamic properties during

its development into a more mature and stable microbial ecosystem[14]. Many environmental factors are known to influence the infant gut microbiome e.g. birth mode, feeding type (breastfeeding vs. formula feeding), antibiotics, hospitalisation, gestational age, and more[5,15,16]. As the interactions between the infant gut microbiome and its human host are so heterogeneous, it is challenging to identify microbe-specific associations with human phenotypes, such as atopic conditions. This is confirmed by conflicting results in research regarding the association of the gut microbiome with eczema as well as asthma. Although the presence of some association is generally acknowledged, the direct involvement of specific microbes, or overall microbiome diversity, remains an open discussion. It should be mentioned however, that inconsistencies in literature can be introduced by methodological differences such as sample collection, DNA isolation and sequencing, taxon classification, and differences in age or geography of the studied subjects[17,18]. Nevertheless, multiple associations have been established between microbial colonisation patterns during neonatal life and atopic conditions after human maturation[5,6,8–10].

The first atopic condition that this thesis focusses on is asthma. Asthma is a complex and heterogeneous chronic lung disease, which occurs in different endotypes, symptoms, and levels of severity and is incurable[19]. The prevalence of asthma has been increasing since the 1950s in many countries, and particularly in regions undergoing urbanisation[20,21]. Asthma prevalence puts a heavy burden on public health care and has been shown to result in substantial economic costs as well. It has been estimated that the total of direct and indirect costs of asthma reached $81.9 billion in the United States during 2013[22]. For the United Kingdom, the total costs during 2011 were estimated at £1.1 billion[23]. It has been argued that these costs could be greatly reduced by improving disease control[24]. However, the ideal solution is prevention, which would not only reduce the economic cost, but could completely remove the disease burden of individuals for whom this disease could be prevented. Additionally, the disease burden would be drastically reduced for individuals for whom intervention results in a reduction of disease severity. Nevertheless,

for any form of prevention to be successful, the ability of early identification of high-risk individuals that require such an intervention, is a must. Unfortunately, early asthma diagnosis is challenging. Roughly, 20-70% of asthmatic adults are undiagnosed. Overdiagnosis, where non-asthmatic adults are diagnosed as asthmatics, also occurs. Such overdiagnosis can lead to unnecessary use of corticosteroids inhalation[25], the most effective and commonly used long-term treatment for asthma which can have severe adverse effects[26,27]. Moreover, clinical diagnosis of asthma is particularly challenging in young children, as children up to five years of age generally struggle with cooperating reliably in lung function measurements. This is of concern as early diagnosis allows for a timely asthma treatment, which has many benefits compared to treatment at a later age. Early treatment can reduce respiratory symptoms[28–31], health care costs[26,31], permanent damage to airway structures[29,31], and has further benefits described elsewhere[32]. Moreover, many potential preventative intervention treatments are being researched worldwide[33–35], which if successful, would increase the benefit of early treatment even more. Evidently, early identification of high-risk individuals, and subsequent diagnosis of asthma is critical for improving asthma disease control.

The second atopic condition explored in this thesis is eczema, a chronic inflammatory skin disease that often starts early in life. This condition is often referred to as atopic dermatitis, however, as this thesis project did not make use of immunoglobulin E in serum, allergen skin prick tests or detection of specific immunoglobulin E antibodies, the general term eczema is most appropriate[36]. Eczema is the most common chronic inflammatory skin disease[37,38] with a reported prevalence of 11.3-12.7% and 6.9-7.6% in American children and adults respectively[39]. To find a solution for the high prevalence of eczema, research is being done into the prevention of eczema through use of probiotics[40–45]. It would benefit such research and treatments if early life risk prediction of eczema were possible. Particularly, if such a prediction is based on the gut microbiome composition, as this could then also aid in directing these treatments and/or research. However, more research is required towards such predictive models as regardless of the fact that many

association have already been identified between the gut microbiome and eczema[5,7,9], none of the currently identified microbiome-eczema associations are strong enough to determine which individual from a general population will develop eczema.

There is already a prediction model for the development of asthma that can outperform physician's asthma diagnoses in young children (0 - 4 years). According to Caudri *et al.*, 2009, a doctor's diagnosis of asthma has a low sensitivity of 29%, and specificity of 88%, whereas they proposed a predictive model using clinical variables, which could achieve a sensitivity of 36% and specificity of 91%[46]. Many more predictive models for asthma have been developed, yet to the best of our knowledge, none are adequate for clinical use. This is due to various reasons. First, the models are not accurate enough, either not identifying enough true positives, or selecting too many false positives. Second, some models use variables that are collected at a time point that is too late for early intervention. Finally, another downside of some of the models is the use of invasively collected measurements, which would render the predictive model ineffective for use in clinical practice[32]. As it is known that for both asthma and eczema an accurate prediction method is sought to improve disease control, and that both conditions have been shown to be associated with the gut microbiome, this thesis explores a series of predictive machine learning (ML) methods and their predictive capacity for asthma and eczema using clinical and microbiome data. ML models have already been successfully deployed on microbiome data sets for the prediction of inflammatory bowel disease[47,48], irritable bowel syndrome[49,50] and other conditions[51]. However, for asthma and eczema such publications are generally hard to find. One promising study, however, was performed by Stokholm *et al.*, 2018[6], who were able to predict, to some extent, the development of asthma at 5 years using gut microbiome data of one year old infants with asthmatic mothers. However, they did not find predictive performance when also looking at infants without asthmatic mothers. More research and validation is required to determine if the gut microbiome truly

holds predictive capacity towards specific atopic conditions, and if so, how accurately these can be predicted on a general population. Therefore, this thesis explores a multitude of ML models to evaluate how well these can predict asthma and eczema using clinical and/or microbiome data.

2.2 Report overview

This thesis was written for an interdisciplinary audience, to illustrate the potential and the limitations of ML in combination with microbiota composition data for prediction of atopic conditions. A descriptive overview of all the abbreviations that are used throughout this thesis, has been provided in the appendix section 9.1. For less experienced readers on the topic of ML or data fusion, it is recommended to read the appendix methodology section 9.2 first, which describes some of the used methods more fundamentally. The exact implementation of all the methodologies is described in section 3. The results are shown in section 4, where section 4.1 shows some of the explorative analysis to gain insight into the properties of the clinical and microbiome data. The results of the ML models can be found in section 4.2. This thesis discusses the implications of the results in the discussion (section 5), which has been subdivided into the findings (section 5.1), limitations (section 5.2), and future directions (section 5.3). This thesis concludes with some final remarks in section 6.

3. Materials and methods

3.1 General setup

This section elaborates on the general setup of this thesis project. To this end, references will be made within this section to different steps indicated as green circles in Figure 1, which is a flowchart that summarises the general setup of the work that was performed for this thesis.

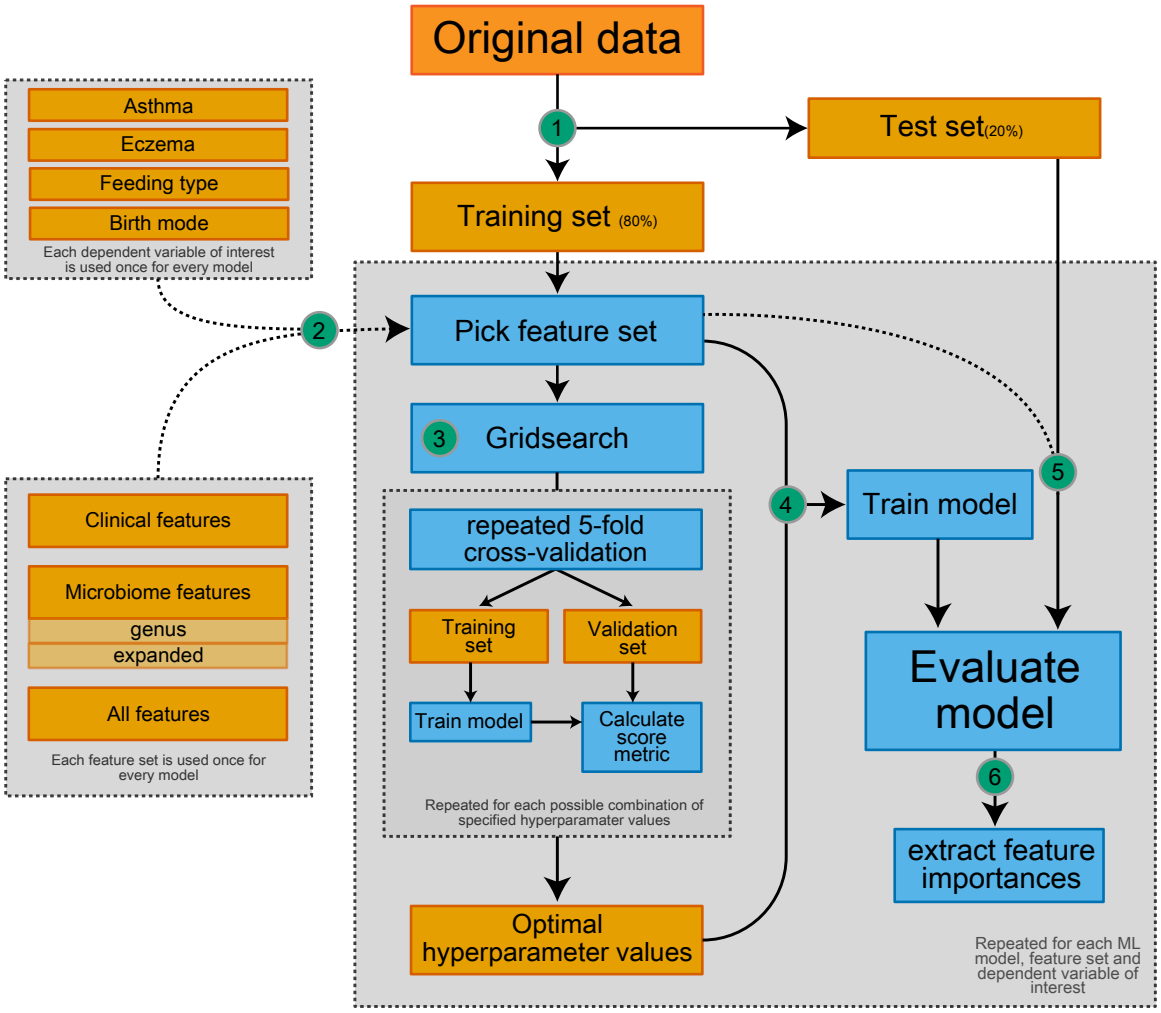


Figure 1: Flowchart of the implementation, optimisation, and evaluation of the machine learning algorithms. Orange boxes represent data/information, blue boxes represent procedures, and green circles indicate the different step numbers.

The first step, as indicated in Figure 1, was to split the data into a training and test set, containing 80% and 20% of the samples of the original data, respectively (described in section 3.4). This is an important step, that allows for a proper evaluation of the final model on data that the model has not seen before. This mimics the evaluation of the model on new data to some extent. Subsequently, the data was prepared for the ML methods, depicted by step 2. Accordingly, the data was prepared separately for the different dependent variables of interest, which had to be predicted by the ML models. For all these different dependent variables, specific feature sets were generated, each containing different combinations of variables (section 3.5). These feature sets contained clinical and/or microbial variables. All the resulting combinations of dependent variables and feature sets were fed to the different ML methods. Step 3 corresponds to the grid search, which is the form of hyperparameter optimisation that was performed to tune the models such that they could be optimally trained on the different feature sets, to predict the variable of interest. The grid search was performed through repeated 5-fold cross-validation as is described in section 3.8.1. The final model for each given feature set, dependent variable, and ML model, was trained using the optimal hyperparameter values that resulted from the grid search, as shown in step 4. In the fifth step, the resulting model was evaluated on the test set. Finally, in step 6 feature importance values were extracted from the model using the Shapley value method, to gain model interpretability.

3.2 General code overview

The code for this thesis was written in R[52] using Rstudio version 1.4.1106 64-bit[53]. The renv[54] package was used to create a snapshot of all the used packages, allowing for

seamless reproducibility through an identical R environment. All models were run on a desktop with a 12-core, 32-thread AMD Ryzen™ 9 3950x central processing unit, in combination with 32 gigabytes of random-access memory. Parallelised tasks were given access to all 32 threads. For reproducibility, the random seed was set to 8392 where possible. For parallelised tasks the random seed was registered using the doRNG[55] package. All visualisation were created with ggplot2[56].

3.3 The data

The data on which the models were built and evaluated comes from the prospective KOALA birth cohort study in the Netherlands. A study whose aim was to identify factors influencing atopic conditions. In this study, a total of 2343 pregnant women with "conventional" lifestyles were recruited from another study regarding pregnancy-related pelvic girdle pain. Also, 491 pregnant women with an 'alternative' lifestyle were selected through organic food shops, anthroposophical doctors and midwives, Steiner schools, and dedicated magazines. Here, 'alternative' lifestyle refers to child rearing practices, dietary habits (organic or vegetarian), vaccination schemes and/or antibiotics use. For 1176 infants, faecal samples were collected at the age of 1 month. The faeces were collected by the participants at home in a faeces tube and delivered by mail, accompanied by a faecal collection questionaire[15]. Additionally, multiple questionnaires were conducted at 14-18, 30 and 34 weeks of gestation, and 3-, 7-, 12-, and 24-months post-partum. At two years of age, a home visit was realised to identify individuals who had developed atopic dermatitis[57] (using diagnostic criteria of the United Kingdom Working Party[58–60] and severity scoring of atopic dermatitis index[61]). Multiple follow-up questionnaires were conducted up until the age of 11 years, including various atopy related questions. The collected faecal samples were diluted 10-fold in the laboratory in peptone water containing 20% v/v glycerol and kept at −20 °C until analysis. For this thesis, samples were excluded if they contained less than one gram of faeces, were not collected between

3 and 6 weeks of age, did not complete the faecal collection questionnaire, were born before 37 weeks of gestation, had twins, had growth-related congenital abnormalities (including Down's syndrome, Turner syndrome, Fallot's tetralogy multiple disabilities), underwent antimicrobial agent administration prior to faecal collection, and/or were lacking a body mass index measurement[57]. In total, 894 samples passed the exclusion criteria and were available for analysis. For more details on the KOALA birth cohort, it has been described more thoroughly by Kummeling *et al.*, 2005[3].

3.3.1 DNA sequencing

For the sequencing of the stool samples, first total DNA was extracted with double bead-beating, after which the QIAamp DNA stool mini kit (Qiagen, Hilden, Germany) was applied following the instructions of the manufacturer. Subsequently, PCR amplification and Illumina HiSeq sequencing of the V4 region of 16S ribosomal RNA genes was performed on the 5-20 ng DNA templates. From the resulting sequencing reads, amplicon sequencing variants (ASVs) were inferred using the NG-Tax2[62] pipeline under default settings and forward and reverse reads were trimmed to a length of 80 bases. ASVs were taxonomically annotated at all taxonomic levels, from the level of species to kingdom through use of the NG-Tax2 and SILVA-132 reference databases. DNA purification and sequencing was not part of this thesis project and has been described in more detail elsewhere[15,63].

3.3.2 Clinical data preparation

The clinical data that was used for this thesis project originates from a collection of surveys that were conducted within the KOALA birth cohort study[3]. A specific selection of variables was made that were collected at approximately one month of age or earlier and were expected to associate with atopy and/or the gut microbiome to some extent. This

resulted in a total of 31 independent clinical variables. A complete overview of all independent variables is shown in Table 1. All variables were collected at one month postpartum or earlier, except for “living on a farm”, which was collected at 4 to 7 months but was expected to not have changed after one month of age. The “breastfeeding proportion” variable was estimated based on a weekly maternal questionnaire regarding the type of feeding. Atopic conditions of mother and father were summarised into a separate atopy variable which describe if one or both of the parents had asthma, dust mite allergy, pet allergy and/or hay fever. “Furry pets during pregnancy” describes if any cats, dogs and/or rodents were present during pregnancy. All categorical variables were encoded as a factor class inherent to R[52], with two or more levels according to the number of possible values the variable can take.

Table 1: Variable overview containing all clinical independent variables. The first column shows the variable name. The second column indicates the possible values that the variable can take. The third column depicts the variable’s class in R, either factor for categorical variables or numeric for continuous and discrete variables.

Variable	Possible values	Class
Older siblings	Yes/no	Factor
Birth mode and place	vaginal at home, vaginal in the hospital, or C-section	Factor
Smoke exposure during pregnancy	No smoke, some smoke, or a lot of smoke	Factor
Neonatal animal exposure	Yes/no	Factor
Antibiotics exposure during pregnancy	Yes/no	Factor
Trimester of antibiotics exposure during pregnancy	1st, 2nd, or 3rd trimester	Factor
Antibiotics	No antibiotics, direct antibiotics, antibiotics in breast milk, or antifungals	Factor
Breastfeeding proportion	linear continuous scale ranging from zero to one, 0 = purely formula fed, 0.5 = balanced mixture, 1 = purely breastfed	Numeric
Feeding type	Breastfed, formula fed, or mixed	Factor

Pregnancy duration	Weeks	Numeric
Frequency of probiotics use during pregnancy	rarely or never, multiple times per month, multiple times per week, daily, or not sure	Factor
Age at faecal collection	Days	Numeric
Maternal asthma	Yes/no	Factor
Maternal dust mite allergy	Yes/no	Factor
Maternal furry pet allergy	Yes/no	Factor
Maternal hay fever	Yes/no	Factor
Paternal asthma	Yes/no	Factor
Paternal dust mite allergy	Yes/no	Factor
Paternal furry pet allergy	Yes/no	Factor
Paternal hay fever	Yes/no	Factor
Pet dogs during pregnancy	Yes/no	Factor
Pet cats during pregnancy	Yes/no	Factor
Pet rodents during pregnancy	Yes/no	Factor
Furry pets during pregnancy	Yes/no	Factor
Birth weight	Grams	Numeric
Baby washed when held	Not washed, washed blood, or washed completely	Factor
Hospitalisation	Yes/no	Factor
Hospitalisation from birth	Yes/no	Factor
Incubator	Yes/no	Factor
Lives on a farm	Yes/no	Factor
Atopic parents	No, one, or two	Factor
Prenatal antibiotics exposure	Yes/no	Factor

The dependent variable asthma was constructed from a collection of asthma variables taken at different time points. Information was available for samples if they had ever developed asthma according to a doctor's diagnosis with symptoms or medication use at 6 to 7, 6 to 8, 7 to 9 or 8 to 10 years of age. If a sample indicated at any of these timepoints to have had developed asthma, the value of the dependent asthma variable for the sample was set to 'asthmatic'. The dependent eczema variable was collected somewhere between the age of 6 to 7 years, depicting whether the individual was ever diagnosed with eczema. Also, an eczema variable was included for eczema during the first two years

of life. Eczema was diagnosed using diagnostic criteria of the United Kingdom Working Party[58–60].

Feeding type and birth mode were included as reference dependent variables. The purpose of these dependent variables was to inspect predictive performance on variables known to highly associate with the gut microbiome composition, to validate model implementation and the adequacy of the microbiome data for predictive tasks. For the prediction task of feeding type, only purely breast or formula fed infants were used. The birth mode prediction task was performed on the “birth mode and place” variable with vaginal delivery at home and in the hospital merged into one class.

3.3.3 Microbiome data preparation

The microbiome data was loaded as a phylosec-class object, inherent to the phyloseq package[64]. On each taxonomical level, the taxonomical units that could not be accurately classified were named according to their unclassified taxonomical level, and the lowest taxonomical level for which they were classified, this was done using the microViz[65] package.

3.4 Splitting the data

The code that was written for this thesis supports two methods for splitting the data into a train and test set. The two supported methods are DUPLEX[66] and randomised stratified sampling. All work shown in this thesis was performed on the train and test set that were generated with randomised stratified sampling, which created a test set of 20% and a training set of 80%. Randomised stratified sampling was implemented using the rsample[67] package.

DUPLEX was implemented using the prospectr[68] package. The sampling technique was implemented such that first the data was imputed with Random Forest (RF) imputation[69] with 20 iterations and 200 trees. Then, the classes of the dependent variables were separated into different data frames and the variables “older siblings”, “birth mode and place”, “direct antibiotics”, antibiotics exposure during pregnancy”, “trimester of antibiotics exposure during pregnancy”, “feeding type”, “hospitalisation”, “atopic parents” and “pets during pregnancy were dummy encoded. On those variables, the DUPLEX function was applied using the Euclidean distance separately for each dependent variable class without any scaling or centring. Finally, both classes were combined for the train and test sets, resulting in stratified sets. The sample IDs were used to reconstruct the train and test sets from the original data such that no imputed values were present in the train or test set.

To evaluate the performance of both methods, a function was designed which performs Principal Component Analysis (PCA) on the training set with dummy encoded categorical variables and min-max normalised numeric variables, creates a scatter plot of the PCA scores, and then projects the test set on to the score plot by multiplying the test set with loading vector from the training set. This allows for a visual representation of the distribution of the training and test set such that it can be easily observed if all areas of the training set are captured by the test set.

3.5 Feature sets

3.5.1 Clinical feature set

The clinical feature set contains the independent variables from the KOALA birth cohort surveys. The numeric variables were min-max normalised, by linearly scaling the values of each variable such that its maximum value is one, and its minimum value is zero. The nominal variables were converted to dummy variables. These pre-processing steps were

incorporated using the recipes[70] package such that scaling is always based on the training set and applied to the test set. For the hyperparameter tuning this means that for each fold of the repeated k-fold cross validation, pre-processing is determined separately on the training set, and later applied to the corresponding validation set.

3.5.2 Microbiome genus feature set

To construct the microbiome genus feature set, the microbiome abundance data was filtered such that only genera present in at least 10% of the samples were retained. Here, a genus was considered present if it had a count of 1 or higher. Subsequently, the taxa were aggregated to the level of genera and the sequencing reads were stored in a data frame. Finally, the sequencing reads were converted to relative abundances such that all relative abundances of each sample sum up to one, resulting in compositional information for 23 genera. Figure A2 in the appendix illustrates the spread of relative abundances across the different genera in the microbiome genus feature set for the prediction task of asthma.

3.5.3 Microbiome expanded feature set

The microbiome expanded feature set contains compositional information of the gut microbiome at all taxonomical levels. To this end, the ASV read counts were filtered at 0.01 prevalence, meaning that only ASVs present in at least 1% of all the samples, with a minimum count of one, were kept. Thereafter, the ASVs were aggregated to all taxonomical levels such that the feature set contains sequencing reads on the taxonomical levels ranging from ASVs all the way up to the level of the kingdom of bacteria. Finally, the sequencing reads of each separate taxonomical level were transformed to relative abundances, such that the values of all taxonomical units of a specific taxonomical level sum up to one for each sample.

3.5.4 Combined feature set

Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) was implemented using the mixOmics[71] package to combine clinical and microbiome variables into one feature set. Prior to running the method, the clinical and microbiome data blocks were prepared as described in section 3.5.1 and 3.5.2 respectively. A more conceptual description of DIABLO is provided in the appendix section 9.2.2.1.

A full design matrix was created, which is an $m \times m$ matrix for m data blocks. Thus, in this case a 2×2 matrix filled with ones, but with a diagonal of zero. This indicates that the covariance between all data blocks will be utilised except for the covariance of a block with itself. To determine the optimal number of components to use, and how many variables to use for the two different data blocks when building the model, hyperparameter tuning was performed. Here, 1 up until 10 number of components, and any combination of 2, 5, 10, 15 and 20 variables both for the clinical and microbiome data block are evaluated on the balanced error rate based on the Mahalanobis distance of the weighted vote classification in a 10-times repeated 5-fold cross-validation. Finally, a DIABLO model with sparse Partial Least Squares (sPLS-DA) was fitted on the training data using the optimal hyperparameter values with a maximum of 10,000 iterations and a tolerance of 1e-08.

3.6 Imputation

The clinical variables had less than 3% missing values, except for the “incubator” and “birth weight” variables, which were missing for roughly ~10% and ~9% of the samples respectively, for the prediction of asthma. The number of missing values per variable and prediction task can be found in the descriptive statistics Table A1.

RF imputation was performed to impute missing clinical data. First, the samples with missing information for the dependent variable were removed. Subsequently, the missForest package[69] was used to impute the training set, with a maximum of 20 iterations and 100 trees. The test set was separately imputed in an identical fashion.

3.7 Explorative data analysis

A multitude of methodologies were equipped to reveal properties of the data and inspect if any clear separation of the data was present which could be useful for the prediction tasks. This section describes the implementation of the methods in detail.

3.7.1 Correlation heatmap

A correlation heatmap was generated for which the spearman correlation was computed for each possible combination of variables, including dependent and independent variables. The complete sample set (N=894) was used and per inter-variable comparison, all missing values were dropped. Then, ggplot2[56] was used to generate the heatmap.

3.7.2 Alpha diversity

The α-diversity was measured on the Shannon index. To this end, the microbiome data was not aggregated, and no filtering was applied. The microbiome[72] package was used to compute the Shannon diversity. The Wilcoxon rank sum test was used to compare α-diversities between different groups.

3.7.3 Circular stacked bar chart

A circular stacked bar chart of the gut microbiome composition was generated from the data for the prediction of asthma. The visualisation was created with a function from the microViz[65] package, which utilises PCA on the centre-log-ratio (CLR) transformed data at genus level, and sorts the samples based on their rotational order around the origin of the first two principal components. In this order, a circular stacked bar chart of the relative abundances of the 8 genera with the highest summed relative abundance across all samples, and all remaining genera combined in one “other” category, was visualised using ggplot2[56]. CLR transformation was performed using the microbiome[72] package.

3.7.4 Principal Component Analysis (PCA)

PCA was performed on the microbiome data. First, the microbiome data was filtered to only retain genera present in at least 10% of all samples with a minimum abundance value of one. Then, the data was aggregated to the level of genera and CLR-transformed. The microViz[65] package was used, which in turn makes use of the vegan[73] package to perform PCA and subsequently generate 2D score plots. The code also supports generation of 3D sore plots using the pca3d[74] package.

3.7.5 Factor Analysis of Mixed Data (FAMD)

Factor Analysis of Mixed Data (FAMD) was performed on the clinical data. First, the dependent variable was removed and stored in a separate object. The recipes[70] package was used to standardise all numeric variables, meaning that for each numeric variable its mean was subtracted from it, and was divided by its standard deviation. Finally, the FactoMineR[75] package was used to create scatter plots of the FAMD dimensions.

3.7.6 Unsupervised Random Forest (URF)

First, the dependent variable was removed from the clinical data frame and stored in a separate variable. Then, synthetic data was generated by randomly sampling each variable with replacement, resulting in synthetic samples with random combinations of observed variable values. Hyperparameter optimisation was then performed for the RF, to determine the optimal hyperparameter values. The hyperparameter values were chosen based on the best Area Under the Receiver Operating Characteristics (AUROC) metric after 10-fold cross validation with 20 repeats. The resulting hyperparameter values were used in the Unsupervised Random Forest (URF) which is essentially an RF on a data set containing the original data and the synthetic data, where the dependent variable describes for each sample whether it is synthetic or real. The URF was trained using the randomForest[76] package, resulting in a randomForest class object, from which the proximity matrix was extracted. The proximities of the real samples were converted into distances as is shown in Eq. 1.

$$distance = \sqrt{1 - proximity_{real}} \qquad \textit{Eq. 1}$$

Classical Multidimensional Scaling (cMDS), which in this scenario is identical to Principal Coordinates Analysis (PCoA)[77], was performed on the distance matrix, which was double centred inherently by the method. Finally, a scatter plot of the scores of the first two components was created.

3.8 Machine Learning (ML)

All ML models in this thesis have been implemented using the tidy models[78] architecture, using the parsnip[79] package to streamline model tuning and training, unless specifically stated otherwise. All machine learning models were tuned, trained and evaluated on all different feature sets: clinical (section 3.5.1), microbiome genera (section 3.5.2),

microbiome expanded feature set (section 3.5.3), and the DIABLO combined feature set (section 3.5.4). This was carried out separately for the prediction tasks of all the following dependent variables: asthma, eczema during first two years, eczema at 6-7 years, feeding type and birth mode. All models were trained and tuned with up sampling unless specifically stated otherwise. The test set was not up sampled for the final model evaluation.

3.8.1 Hyperparameter optimisation

The hyperparameter optimisation of all ML methods was performed in an identical fashion unless an alternative approach is mentioned in the section of that ML method. For each of the ML methods, a custom function was designed to perform all tasks involved in the hyperparameter optimisation. A series of input variables allow control over how the optimisation is performed. For each tuneable hyperparameter, an array of values can be supplied. Then, the function generates a grid, consisting of all possible combinations of inputted hyperparameter values. Additionally, repeated stratified k-fold cross-validation can be specified, by setting the number of folds and repeats. To ensure reproducibility of this stochastic process, a seed value can also be supplied, which is a positive integer that fixes the random state such that it will return the exact same result every time it is ran. In this thesis project, the seed was set to 8392 for repeated 5-fold cross-validation, with 10 repeats unless stated otherwise. Finally, the hyperparameter optimisation was performed by running a grid search, which was implemented with the tune[80] package. Within the grid search, a multitude of evaluation metrics are computed, including accuracy, balanced accuracy, AUROC, F1, precision, sensitivity, specificity, and the mean log loss. Within this thesis project, the optimal hyperparameter values were based on the mean log loss score, unless stated otherwise. The evaluated hyperparameter values for each model are shown in the appendix Table A2.

All hyperparameter optimisation functions designed in this thesis project also support parallelisation, which can be enabled with an input variable. To this end, a parallel socket cluster is created. By default, as many sockets are created as there are logical processing cores available to the system. These parallel sockets are registered using the doParallel[81] and future[82] packages. The grid search is specified to parallelise over "everything" rather than "resamples", meaning that a nested parallelised loop will be constructed which parallelises over the splits from the 5-fold cross-validation, but also over all the unique combinations of hyperparameter values. This can improve computational efficiency when there are more logical processing cores then splits.

3.8.2 Logistic Regression (LR)

Logistic Regression (LR) was implemented with the glmnet[83] engine under classification mode. The penalty and mixture parameter values were determined using a grid search as described in section 3.8.1 for any combination of hyperparameter values as indicated in the appendix Table A2. Here, the penalty parameter corresponds to the lambda parameter of glmnet, which represents the total amount of regularisation that is performed. The mixture parameter sets the alpha parameter of glmnet, which can take any value between zero and one, specifying the ratio of the proportion of L1 (lasso) and L2 (ridge) regularisation. A mixture value of zero, results in only using L1 regularisation, whereas a mixture value of one represents purely L2 regularisation.

3.8.3 Random Forest (RF)

RF was implemented with the ranger[84] engine. The hyperparameter optimisation was performed as described in section 3.8.1 for the "mtry", "trees" and "min_n" hyperparameters. The "mtry" parameter specifies how many randomly sampled variables should be used for each split. The number of "trees" represents how many decision trees

are built for the RF, and “min_n” controls how many samples should be present in a node for it to be split further. The exact values that were used can be found in Table A2.

3.8.4 eXtreme Gradient Boosting (XGBoost)

For the eXtreme Gradient Boosting (XGBoost) implementation, the xgboost[85] engine was used and the objective function was set to logistic regression for binary classification. The hyperparameter optimisation framework described in section 3.8.1 was utilised. The hyperparameters that were tuned are “trees”, “min_n”, “tree_depth”, “learn_rate” and “loss_reduction”. A complete overview of the evaluated hyperparameter values can be found in supplementary Table A2. The cross-validation was not repeated for XGBoost due to computation time limitations.

3.8.5 Support Vector Machine (SVM)

The Support Vector Machine (SVM) was implemented using the kernlab[86] engine under classification mode in combination with the radial basis function (rbf) kernel. Hyperparameter tuning was performed as is described in section 3.8.1. The hyperparameters that were tuned are the “cost”, which controls the cost of samples ending up in the inside or on the wrong side of the margin, and “sigma”, which controls the precision of the rbf.

3.8.6 Sparse Partial Least Squares Discriminant Analysis (sPLS-DA)

The sPLS-DA was performed using the mixOmics[71] package. For this method, the hyperparameter optimisation was not performed exactly as described in section 3.8.1 because mixOmics did not comply with the tidymodels[78] framework at the time of writing.

As a result, the hyperparameter optimisation was carried out using the sPLS-DA tuning function, inherent to the mixOmics package. Here, 5-fold cross validation was performed on the training set. No up-sampling was performed as that could lead to data leakage from the training sets into the validation sets within the 5-fold cross validation. To compensate for the imbalance under the absence of up-sampling, the sPLS-DA hyperparameter tuning was optimised for the balanced error rate. Up to ten components were evaluated and the number of used variables can be found in Table A2. sPLS-DA predictions were based on the Mahalanobis distance measure. Data was always scaled to unit variance and mean centred prior to any sPLS-DA.

3.8.7 Feature importance

The SHAPforxgboost[87] R package was used to calculate the Shapley values for the tuned XGBoost models on all feature sets. To this end, the XGBoost[85] model was trained using the optimal hyperparameter values on the corresponding training set. The resulting model, and the training set were used as input for the SHAPforxgboost functions to get the Shapley information and plots.

4. Results

4.1 Explorative data analysis

The training and test set, constructed through random stratified sampling, describe a similar distribution of the data. This can be observed in the PCA projection plot in Figure 2, where scatter plots of PCA scores are shown for the microbiome and clinical data of the training set, represented by the blue points, with the test set projected onto it as red points. Additionally, this figure illustrates that the imputed samples do not deviate from the distribution of non-imputed samples. This is shown in Figure 2 as the triangles and circles, which correspond to imputed and non-imputed samples respectively, exhibit a similar spread.

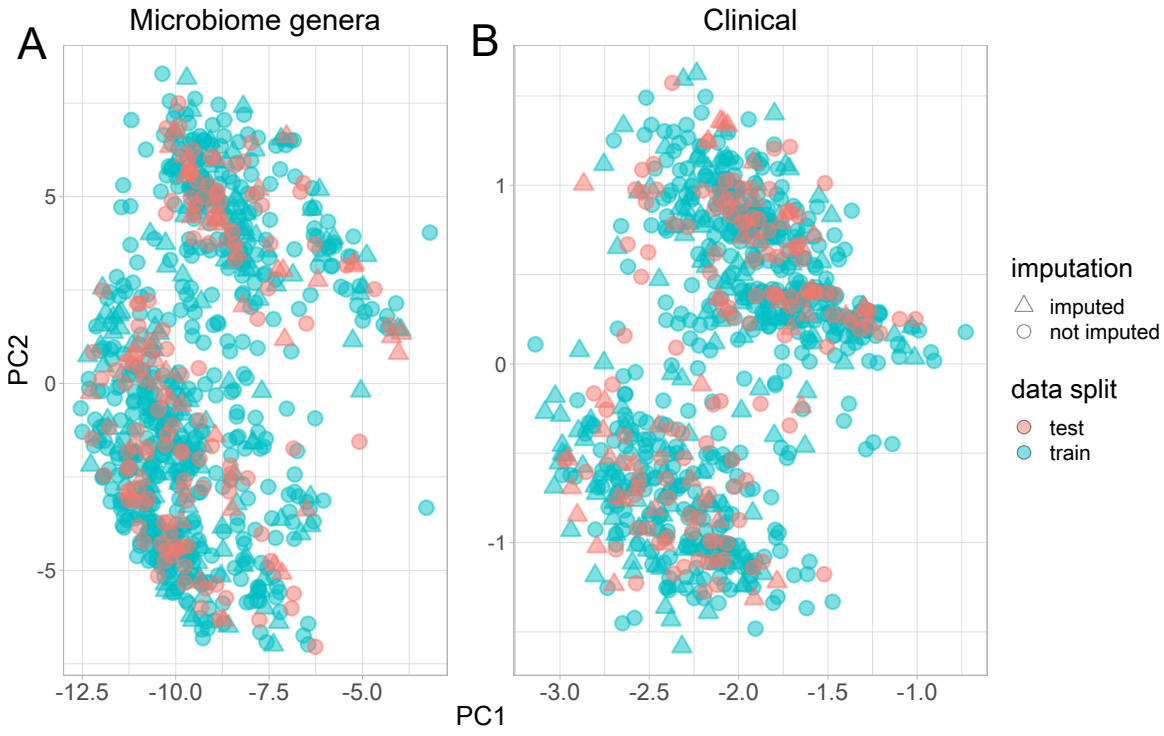


Figure 2: Scatter plot of PCA scores on **(A)** microbiome and **(B)** clinical training sets, with the according test set projected onto it. Red colours represent samples from the test set, blue colours represent samples

from the training set. Triangles indicate that the sample had at least one variable imputed, circles correspond to samples for which no variables were imputed.

Based on the spearman correlation, asthma and eczema did not have a correlation with any of the independent variables stronger than 0.21. The highest correlation was between asthma and maternal asthma, with a correlation coefficient of 0.21. A heatmap of all correlations is shown in Figure A1.

The scatter plot of the FAMD scores on the clinical data, as is illustrated in Figure 3, shows that the first two dimensions of the FAMD analysis do not exhibit different behaviour for asthmatic or eczematic samples compared to non-asthmatic or non-eczematic samples, respectively.

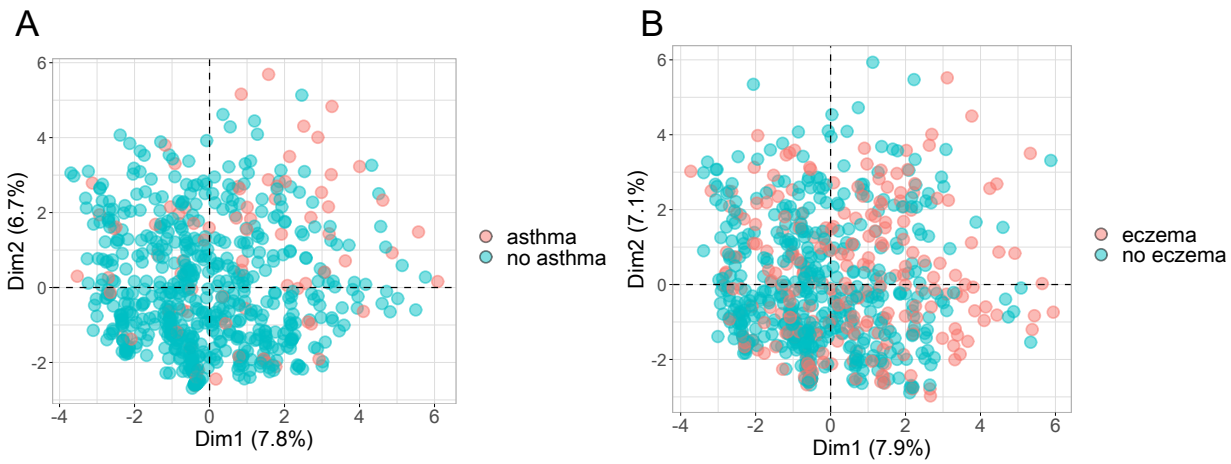


Figure 3: Scatter plot of the first two FAMD dimensions on the imputed clinical data for all samples with **(A)** information on the dependent asthma variable and **(B)** information on the dependent eczema variable. The points correspond to samples and the colours indicate to which class a sample belongs as indicated in the legend. The percentages represent the amount of variance explained by the dimension.

The URF cMDS score plot is shown in Figure 4 and illustrates that also URF does not reveal any structure in the data that can separate asthmatic from non-asthmatic samples, this holds true for both clinical and microbiome data.

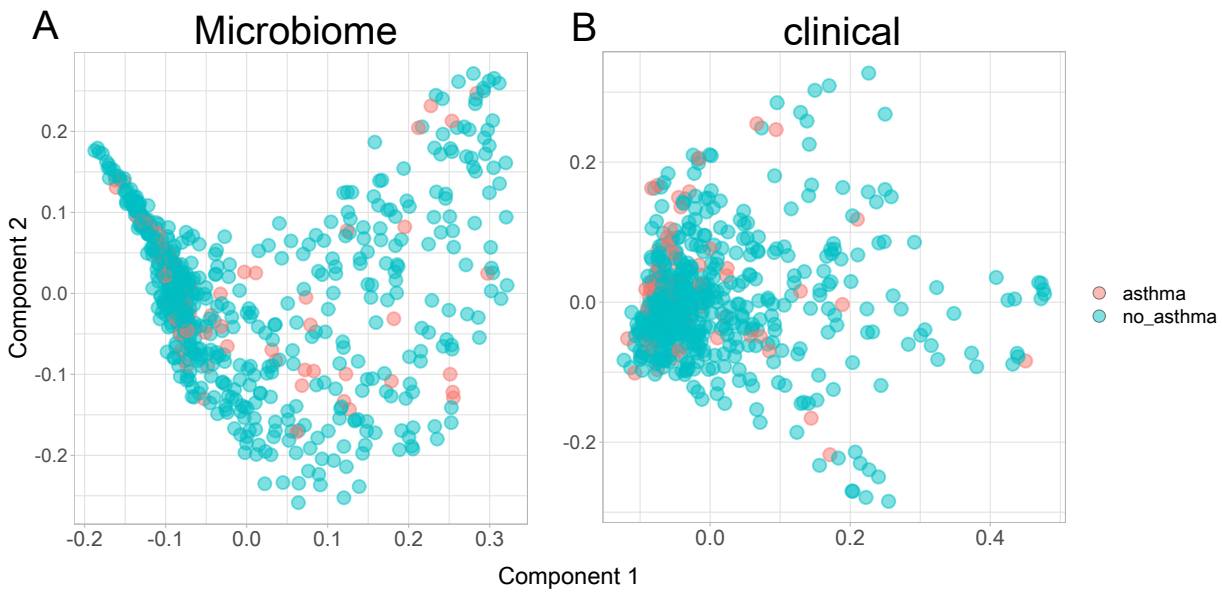


Figure 4: Scatter plot of URF cMDS scores. This figure presents the URF cMDS score plot of the **(A)** microbiome data at centre-log-ratio transformed genus level with prevalence filter of 0.1 and **(B)** the clinical data. For both plots the x-axis corresponds to the first MDS component, and the y-axis to the second MDS component. Red dots represent asthmatic samples and blue dots represent non-asthmatic samples.

The α-diversity of asthmatics, measured on the Shannon diversity index, did not differ from those of non-asthmatic samples, as tested with the Wilcoxon rank sum test ($p = 0.65$). Also, eczematic and non-eczematic (during first two years) samples showed no difference in α-diversity ($p = 0.91$). The similarity in α-diversity is illustrated in Figure 5 which shows the distribution of α-diversities for asthmatic and non-asthmatic samples in plot A, and eczematic and non-eczematic samples in plot B, which are both very similar.

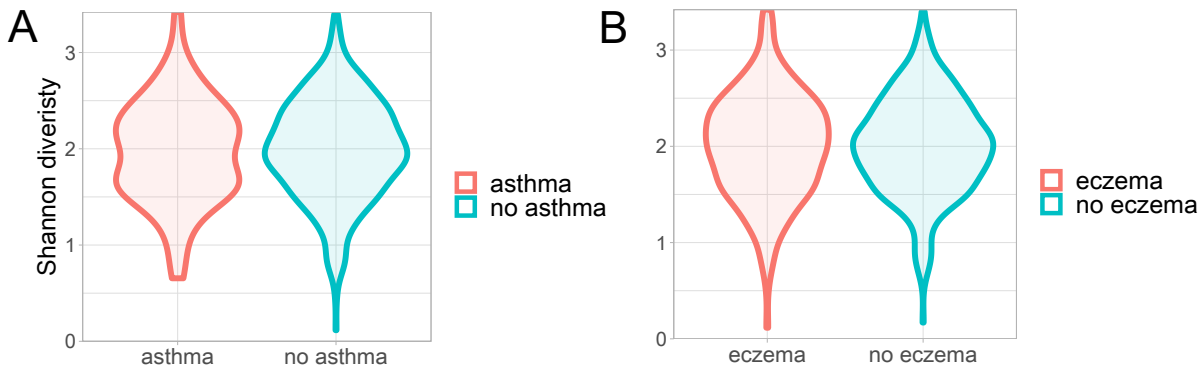


Figure 5: Violin plot of the α-diversity, measured on the Shannon diversity index. Diversities are compared between **(A)** asthmatic and non-asthmatic samples, and between **(B)** eczematic and non-eczematic (during first two years) samples.

The PCA biplot of the clr-transformed microbiome data at genus level, filtered at 10% prevalence, shows that *Bacteroides*, *Parabacteroides* and *Streptococcus* are the biggest drivers of the first principal component (PC), describing ~25% of the total variance. *Bifidobacterium*, *Clostridium sensu stricto 1* and some undefined genus of the *Enterobacteriaceae* family are the biggest drivers of the second PC, describing ~14% of the variance. The colouring of the figure reveals that the first two PCs do not reveal any separation of asthmatics from non-asthmatics, or eczematics from non-eczematics, as they have a similar distribution throughout the plot.

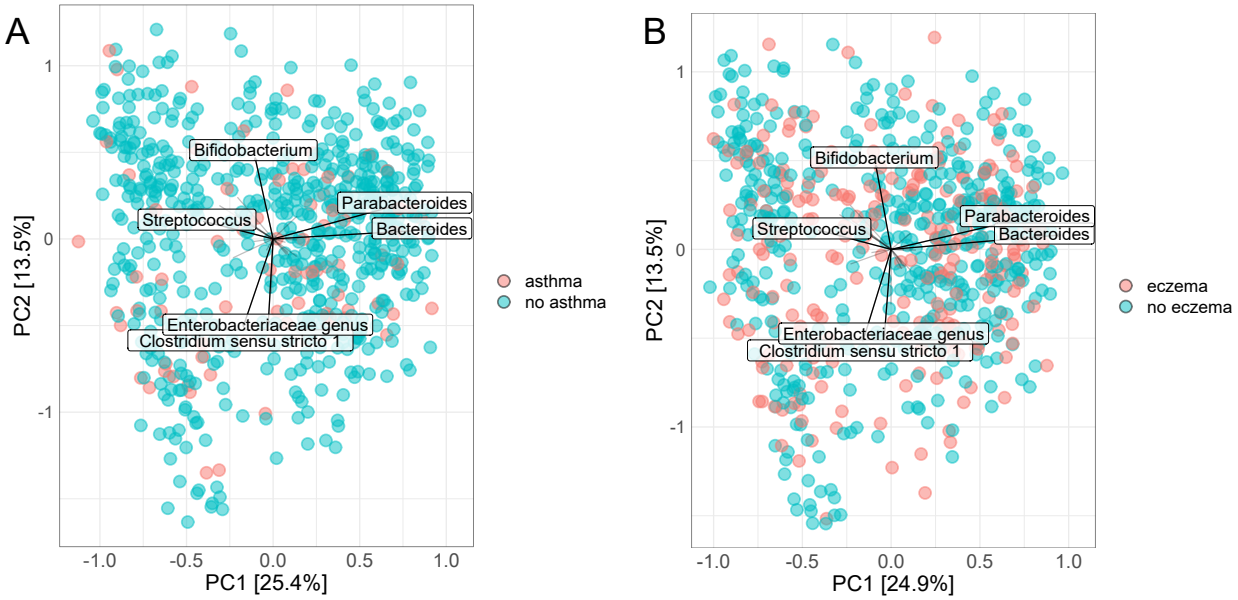


Figure 6: PCA biplot of the centre-log-ratio transformed microbiome data at genus level on samples with information on **(A)** the dependent asthma variable, and **(B)** the dependent eczema (during first two years) variable. Genera were filtered at 0.1 prevalence with a minimum total count of 1. Samples with missing values for the dependent variable were dropped.

The circular stacked bar chart, visualised in Figure 7, summarises the microbiome composition for all samples with information on the asthmatic dependent variable. Relative abundances are shown for the eight taxa with the highest summed relative abundance over all samples, the remaining taxa are combined in the 'other' category. From this chart, it can be observed that most infants were either dominated by *Bifidobacterium* or some unclassified genus that belongs to the *Enterobacteriaceae* family. Most of the samples had a relatively high abundance of *Bacteroides* as well. The outer ring shows that the asthmatic samples are spread out evenly over the different compositions, which were sorted based on rotational order around the origin of the first two PCA scores.

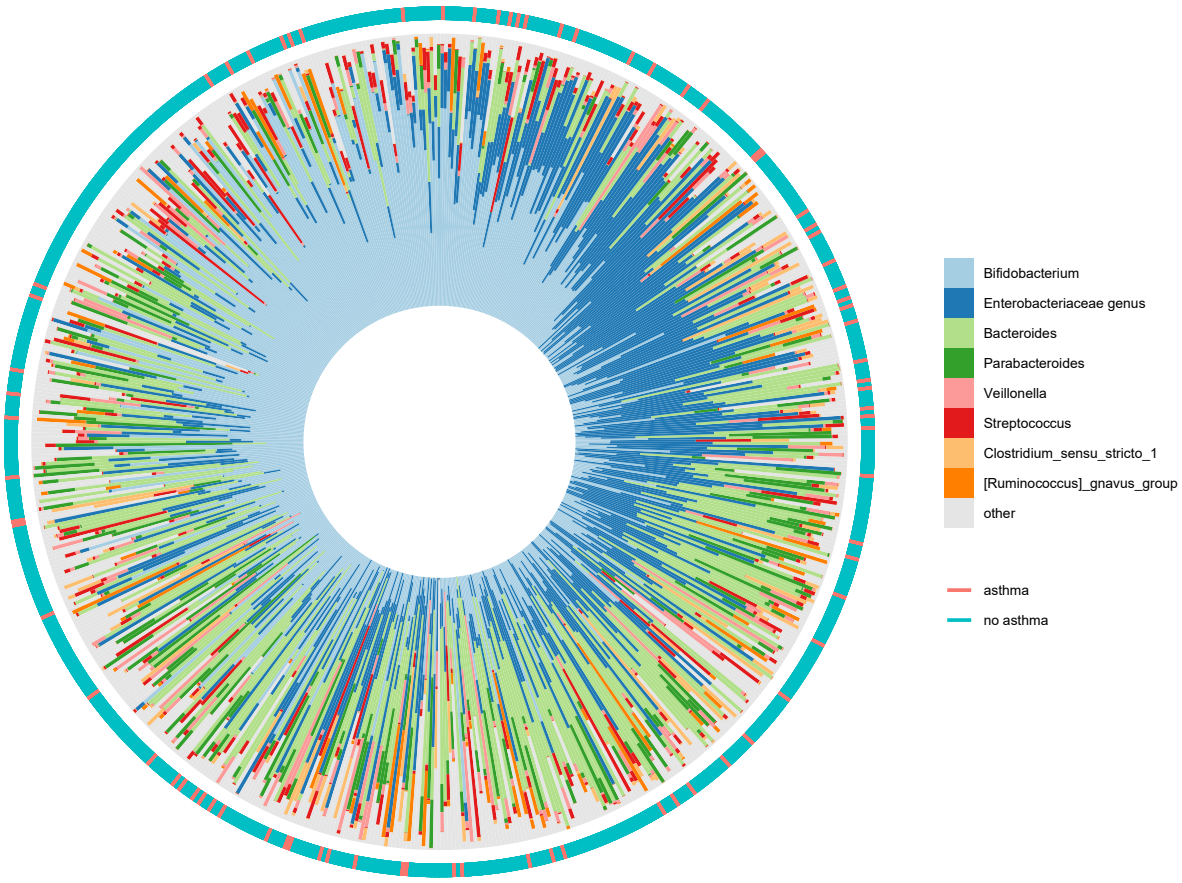


Figure 7: Circular stacked bar chart of microbiome composition at genus level. The eight most abundant genera are colour coded as indicated in legend, where grey represent the combination of all other genera present in the data. Each bar represents a sample where the colours correspond to the relative abundance of a given genus, as indicated in the legend. The outer ring shows the class of the samples, red indicates that a sample is asthmatic, blue indicates that a sample is non-asthmatic.

All dependent variables that were used in this thesis were imbalanced, but each to a different extent. Birth mode and asthma had the most imbalance, where only ~11% of the samples belonged to the minority class. Eczema during the first two years of life had the least imbalance, as roughly 1/3 of the infants showed some form of eczema during the first two years of life. The exact percentages and number of samples per class are shown in Table 2.

Table 2: Dependent variable class imbalance overview. This table contains all the dependent variables that were used. Per variable the majority and minority classes are shown. The cells show class label of majority and minority classes, and how samples were available for that class in the data prior to any splits. The percentage per class is shown in between brackets. The subset of asthma for infants of atopic parents considers parents to be atopic if any of the parents have any of the following conditions: asthma, hay fever, furry pet allergy and/or dust mite allergy.

	Minority class	**Majority class**
Asthma	*Asthma* *94 (11.1%)*	*No asthma* 751 (88.9%)
Asthma (atopic parents)	*Asthma* 68 (13.6%)	*No asthma* 431 (86.4%)
Eczema (6 to 7 years)	*Eczema* 158 (20%)	*No eczema* 633 (80%)
Eczema (first two years)	*Eczema* 297 (33.5%)	*No eczema* 590 (66.5%)
Birth mode	*C-section* 95 (10.7%)	*Vaginal* 797 (89.3%)
Feeding type	*Formula fed* 198 (24.4%)	*Breastfed* 615 (75.6%)

4.2 Prediction models

4.2.1 Asthma

None of the models showed high predictive capacity for asthma with any of the feature sets. Some of the models, including RF and XGBoost, scored a high accuracy around ~80%, with a very high specificity nearing ~100%. The AUROC and balanced accuracy, however, barely exceeded the 50% threshold. F1, precision and sensitivity were all below 50%. A comparison of all models and feature sets is presented in Figure 8.

Based on the mean log loss score, the XGBoost model trained on the expanded feature set was the best performing model across all models and feature sets. Figure 9 shows that the XGBoost model could not effectively classify asthmatic samples and classified all

samples as non-asthmatic. Additionally, the truly asthmatic samples were not predicted to be more likely to be asthmatic in general as the distribution of predicted probabilities was similar for asthmatic and non-asthmatic samples.

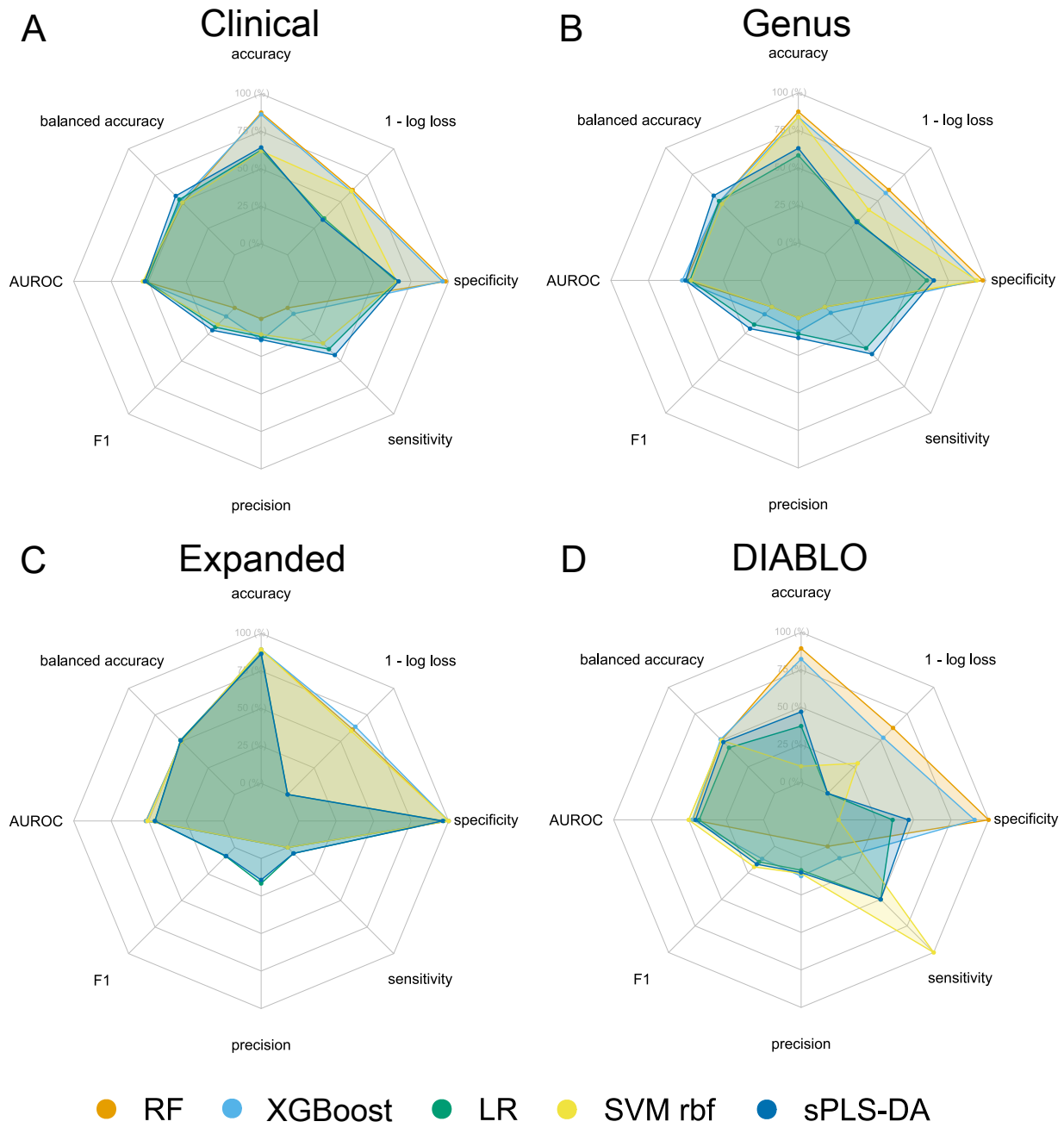


Figure 8: Radar charts of machine learning model performance for predicting asthma with the **(A)** clinical, **(B)** microbiome at genus level, **(C)** microbiome at expanded taxa level, and **(D)** DIABLO data fusion set of clinical and genus feature sets. The colours represent different machine learning models as indicated in the

legend. All metrics are shown on the scale of percentages. The mean log loss is shown as “1 – log loss” as it has been inverted, such that higher values correspond to a better model, just as all other metrics. Any mean log loss higher than one was capped at one such that in this figure they cannot go below zero.

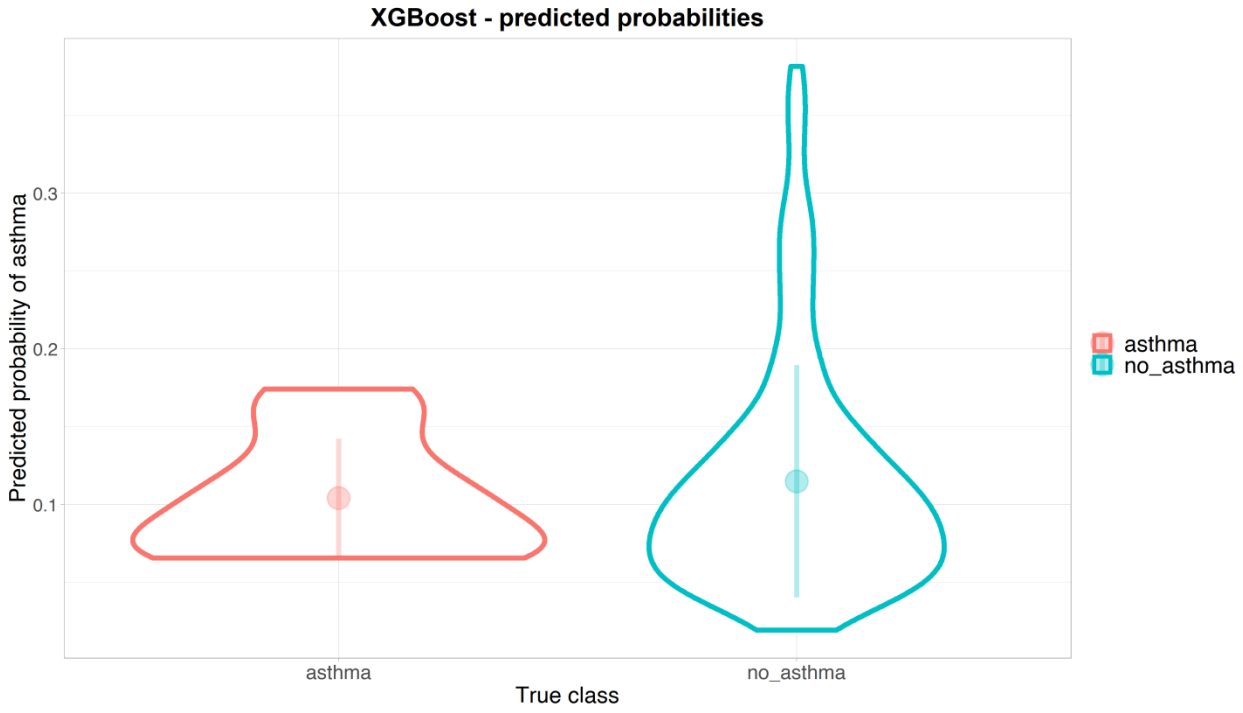


Figure 9:Violin plot of the output of XGBoost to predict asthma on the test set of the expanded feature set. The y-axis shows the output of the XGBoost model, which is the predicted probability of being asthmatic. The x-axis seperates the truly ashtmatic and non-asthmatic samples. The dot in the middle shows the mean with the standard deviation indicated by the vertical line crossing it.

4.2.2 Asthma for infants with atopic parents

This section documents the results for the prediction models of asthma on the subset of infants that have one or two parents with any of following atopic conditions: asthma, hay fever, furry pet allergy and dust mite allergy.

The model results are shown in Figure 10 which illustrates an improved performance compared to asthma on the general study population, both for clinical and microbiome feature sets. Balanced accuracy remained low for the genus and expanded feature set, at around 50%. Also, F1, precision, and recall remained below 50%. However, the AUROC surpassed the 50% threshold, and XGBoost achieved an AUROC of ~0.67 on the genus, expanded and DIABLO combined feature sets. XGBoost on the expanded feature set achieved the lowest mean log loss across all models on the microbiome feature sets. Its output is visualised in Figure 11, which shows that, although asthmatic samples were attributed a higher probability of being asthmatic on average, all samples were predicted to be non-asthmatic.

As illustrated by Figure 12, *Clostridium Sensu Stricto 1* was the most important genus for the prediction for XGBoost on the genus feature set with a mean Shapley value of 0.948. Followed by *Bacteroides*, an unclassified genus of the *Enterobacteriaceae* family and *Bifidobacterium*.

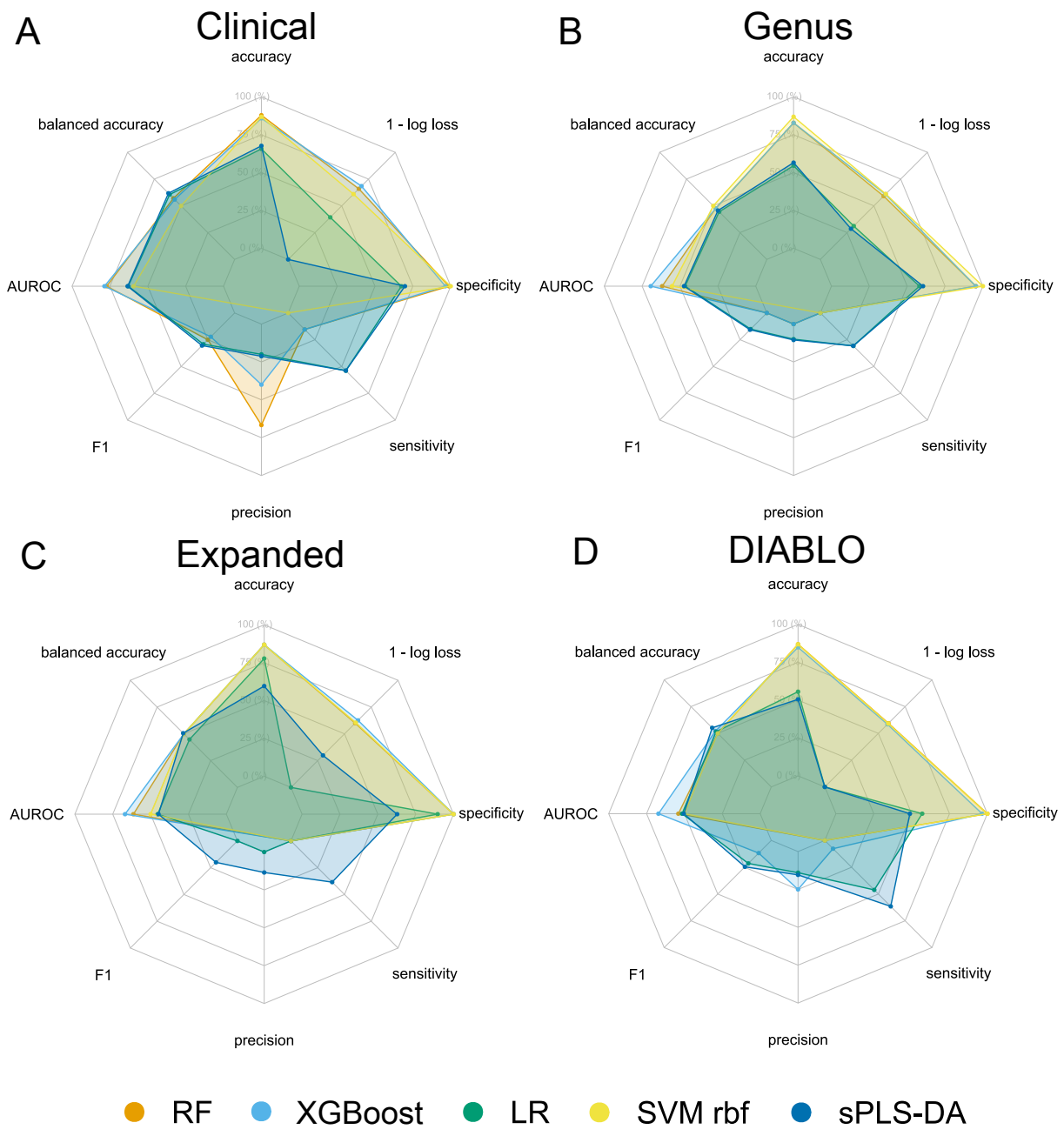


Figure 10: Radar charts of machine learning model performance for predicting asthma for infants with atopic parents. Here, parents are considered atopic if they have any of the following conditions: asthma, hay fever, furry pet allergy and dust mite allergy. Model performance metrics are shown for the **(A)** clinical, **(B)** microbiome at genus level, **(C)** microbiome at expanded taxa level, and **(D)** DIABLO data fusion set of clinical and genus feature sets. The colours represent different machine learning models as indicated in the legend. All metrics are shown on the scale of percentages. The mean log loss is shown as "1 – log loss" as

it has been inverted, such that higher values correspond to a better model, just as all other metrics. Any mean log loss higher than one was capped at one such that in this figure they cannot go below zero.

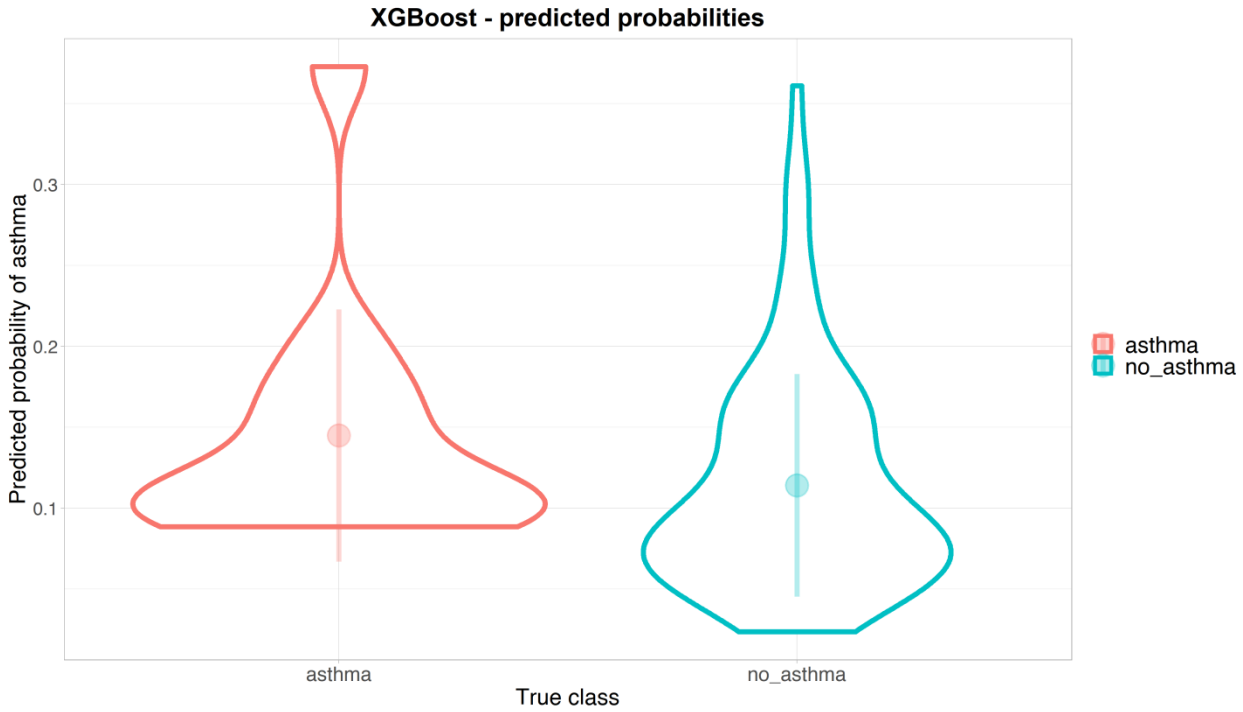


Figure 11: Violin plot of the output of XGBoost to predict asthma for infants of atopic parents on the test set of the expanded feature set. The y-axis shows the output of the XGBoost model, which is the predicted probability of being asthmatic. The x-axis seperates the truly ashtmatic and non-asthmatic samples. The dot in the middle shows the mean with the standard deviation indicated by the vertical line crossing it.

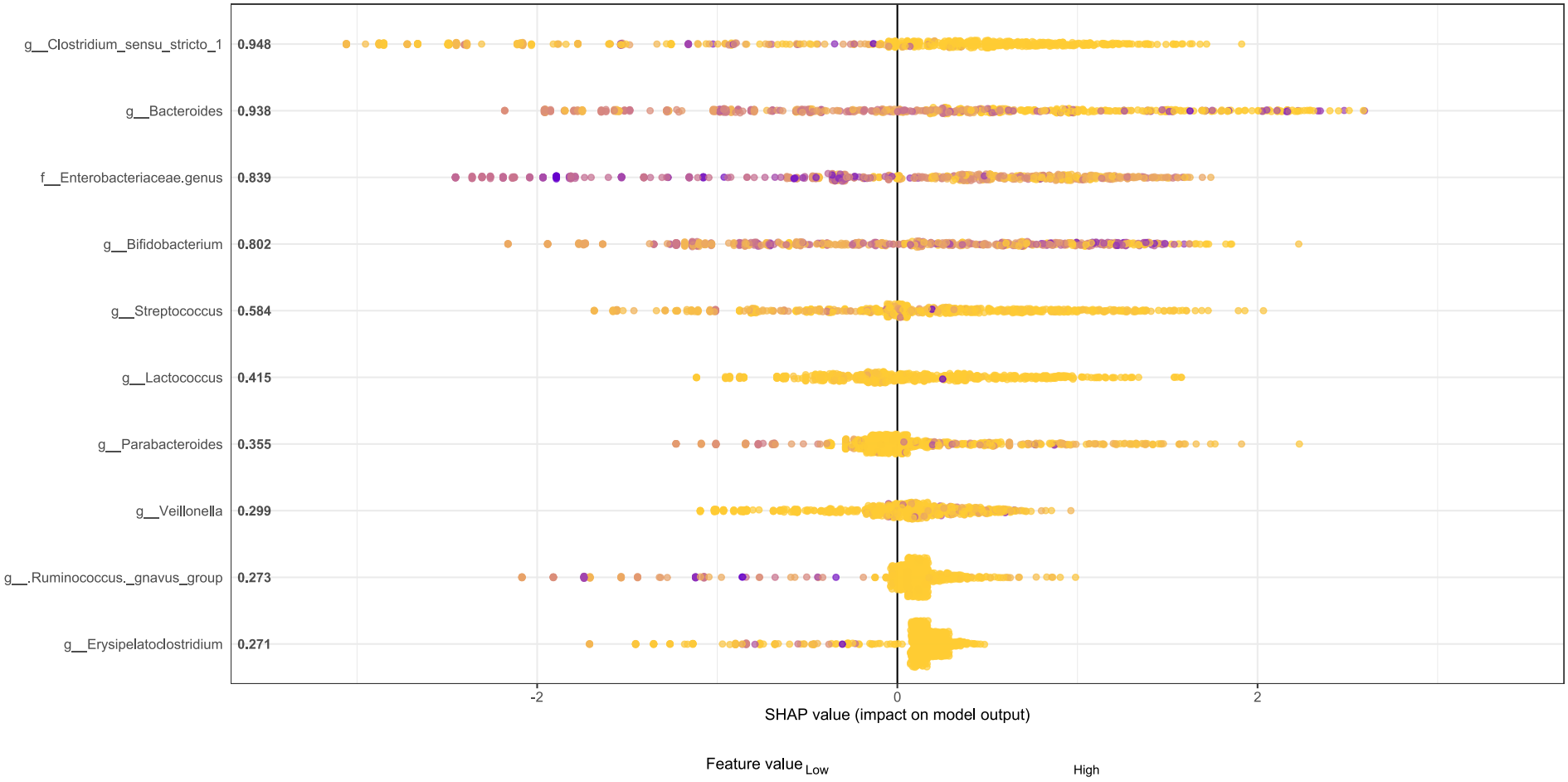


Figure 12: Shapley summary plot of XGBoost on the training set of microbiome genus feature set to predict asthma for infants with atopic parents. The top 10 most important genera based on their mean Shapley value are shown on the y-axis. The mean Shapley value for each genus is shown on the inside of the y-axis. The "g__" prefix specifies that the name corresponds to the name of the genus. The "f__" indicates that the name corresponds the family of the genus, where the specific genus is unknown. The Shapley values, which indicate how important a variable was for the model to predict a specific sample's class, are shown on the x-axis. The direction of the Shapley value, being negative or positive, corresponds to a sample being predicted to be more likely asthmatic or non-asthmatic, respectively. The

yellow-to-purple colour gradient corresponds to the actual value of a given variable for a sample. As such, purple corresponds to samples with high relative abundance of a specific genus, and yellow depicts a low relative abundance

4.2.3 Eczema at 6 to 7 years

For the prediction task of eczema at the age of six to seven years, no model showed promising results for any feature set. RF, XGBoost and SVM generally got high specificity but low sensitivity and precision. AUROC and balanced accuracy was low for all models.

The overall best performing model on the microbiome feature sets, according to the mean log loss score, was SVM with the rbf kernel on the expanded feature set (mean log loss = 0.499). The output of this model, for eczematic and non-eczematic samples, is shown in Figure 14, which illustrates that this model always outputted the same probability of a sample being eczematic, namely ~0.172.

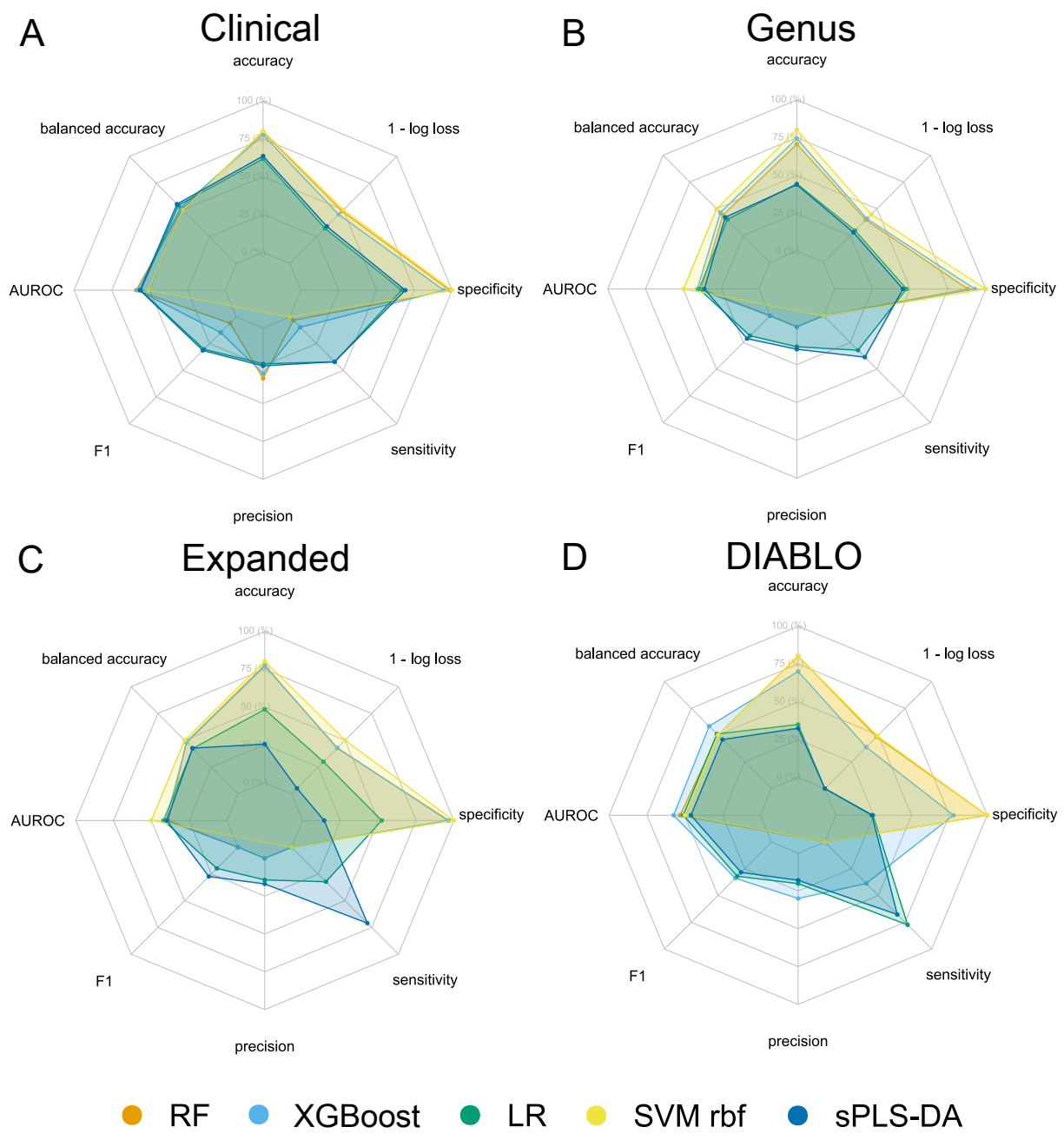


Figure 13: Radar charts of machine learning model performance for predicting eczema at the age of 6 to 7 years with **(A)** clinical, **(B)** microbiome at genus level, **(C)** microbiome at expanded taxa level, and **(D)** DIABLO data fusion set of clinical and genus feature sets. The colours represent different machine learning models as indicated in the legend. All metrics are shown on the scale of percentages. The mean log loss is shown as "1 – log loss" as it has been inverted, such that higher values correspond to a better model, just as all other metrics. Any mean log loss higher than one was capped at one such that in this figure they cannot go below zero.

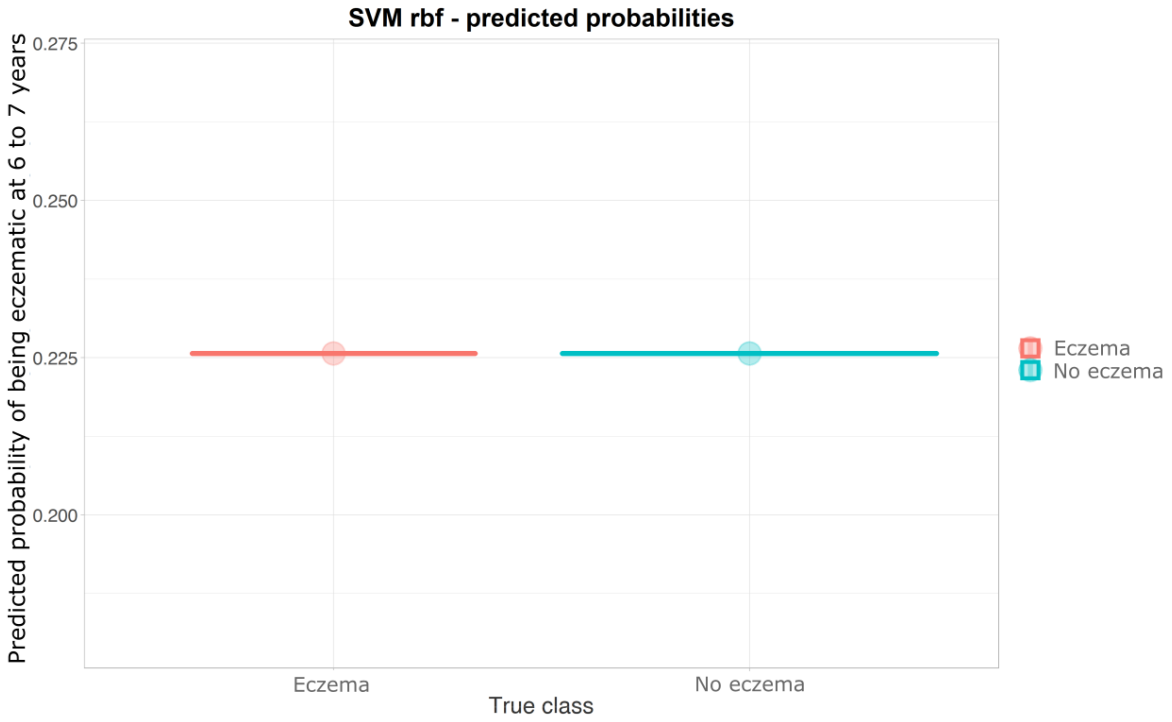


Figure 14: Violin plot of the output of SVM with the rbf kernel to predict eczema at 6 to 7 years of age on the test set of the expanded feature set. The y-axis shows the output of the XGBoost model, which is the predicted probability of developing eczema. The x-axis separates the samples that did develop eczema and those that did not develop eczema. The dot in the middle shows the mean with the standard deviation indicated by the vertical line crossing it.

4.2.4 Eczema during first two years

Development of eczema during the first two years of life could not be accurately predicted by any model, for any feature set. The highest accuracy that was obtained is 66%, which was achieved by the models which classified all samples as not eczematic. The models barely exceed 50% balanced accuracy and AUROC, as sensitivity is always low, and only increases for models with greatly reduced specificity.

The RF model, trained on the DIABLO set which integrated the clinical and genus feature sets, got the best log loss score (log loss = 0.640). This model's output is visualised in Figure 16, which shows that the model classifies all samples as non-eczematic.

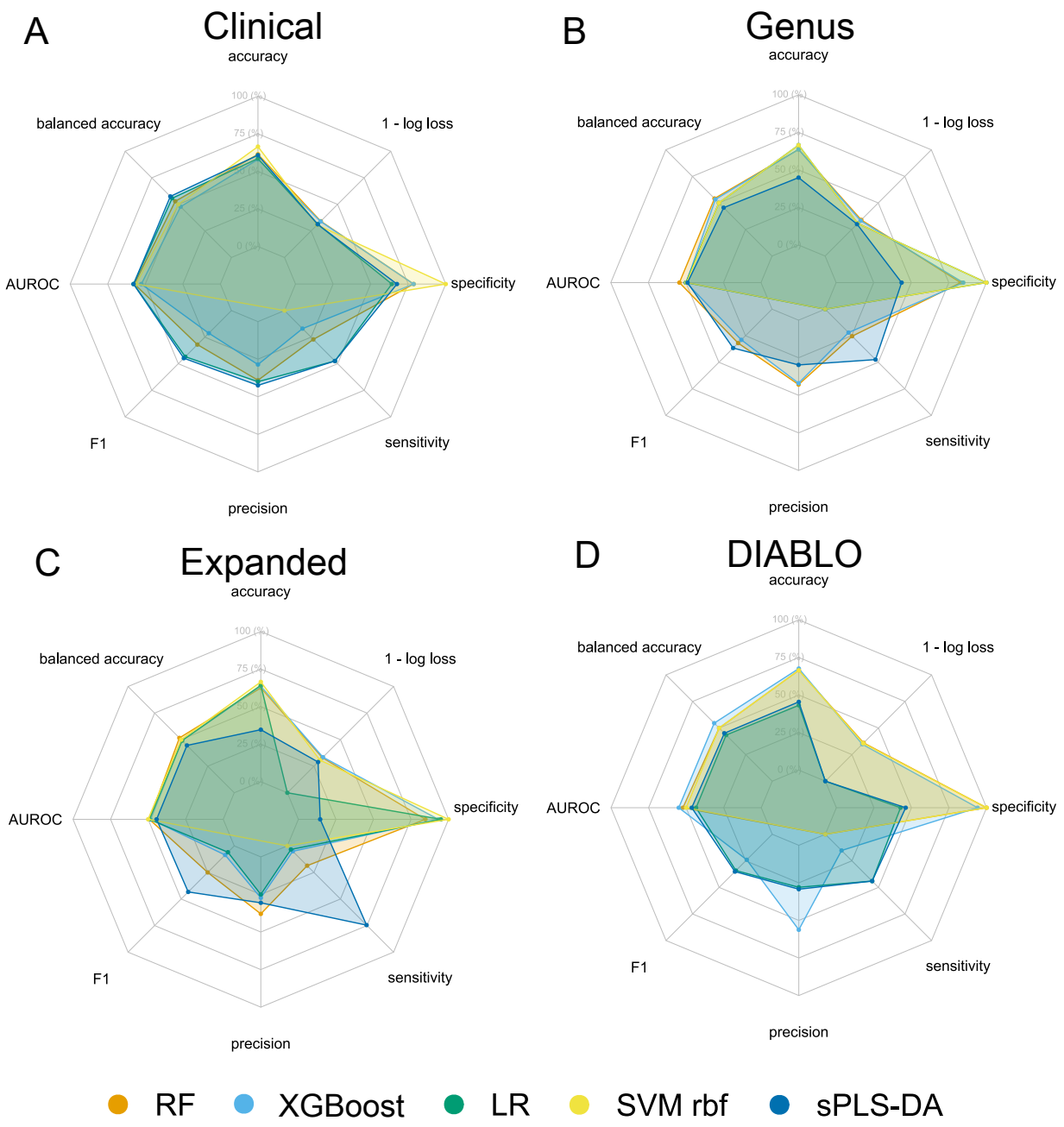


Figure 15: Radar charts of machine learning model performance for predicting eczema during the first two years of life with **(A)** clinical, **(B)** microbiome at genus level, **(C)** microbiome at expanded taxa level, and **(D)** DIABLO data fusion set of clinical and genus feature sets. The colours represent different machine learning models as indicated in the legend. All metrics are shown on the scale of percentages. The mean log loss is shown as "1 – log loss" as it has been inverted, such that higher values correspond to a better model, just as all other metrics. Any mean log loss higher than one was capped at one such that in this figure they cannot go below zero.

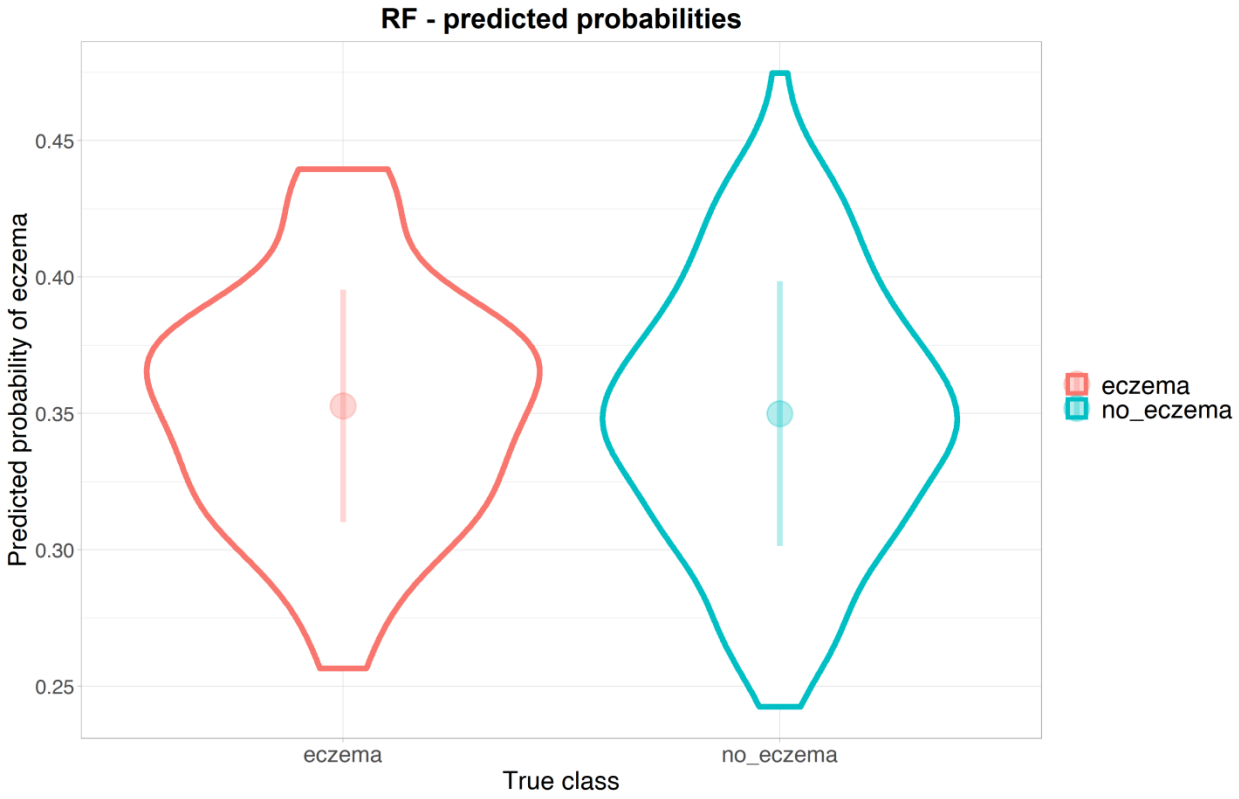


Figure 16: Violin plot of the output of RF to predict eczema during the first two years of life on the test set of the DIABLO feature set. The y-axis shows the output of the XGBoost model, which is the predicted probability of developing eczema. The x-axis separates the samples that did develop eczema and those that did not develop eczema. The dot in the middle shows the mean with the standard deviation indicated by the vertical line crossing it.

4.2.5 Feeding type

The classification task of feeding type proved challenging for all models on the clinical data, but most models got good results across all evaluation metrics for the microbiome feature sets. All models performed well on the genus feature set as all models scored an AUROC of 0.88 or higher. XGBoost, which was the best performing model on the genus feature set according to the mean log loss, performed slightly better on the expanded feature set. RF achieved the second-best log loss on the genus feature set, but a slightly worse log loss on the expanded feature set, with an increase of ~0.004. However, the AUROC did increase from ~0.97 to ~0.98. LR, SVM and sPLS-DA on the other hand dropped in overall performance on the expanded feature set compared to the genus feature set.

The best model across all models and feature sets was XGBoost on the expanded feature set (mean log loss = 0.140). It got an AUROC of 0.99 and mean log loss of 0.14. The model could with great certainty classify breastfed samples and with slightly lower confidence classify the formula fed infants, as is shown in Figure 18.

A Shapley summary plot of XGBoost on the genus feature set is shown in Figure 19, according to which *Enterococcus* was the most important genus to determine the feeding type. In general, it shows that a higher relative abundance of *Enterococcus* suggested a higher probability of being formula fed. A similar trend is observed for *Veillonella* and *Streptococcus*. Higher relative abundance of *Staphylococcus* generally meant that the model would more likely classify the sample as breastfed.

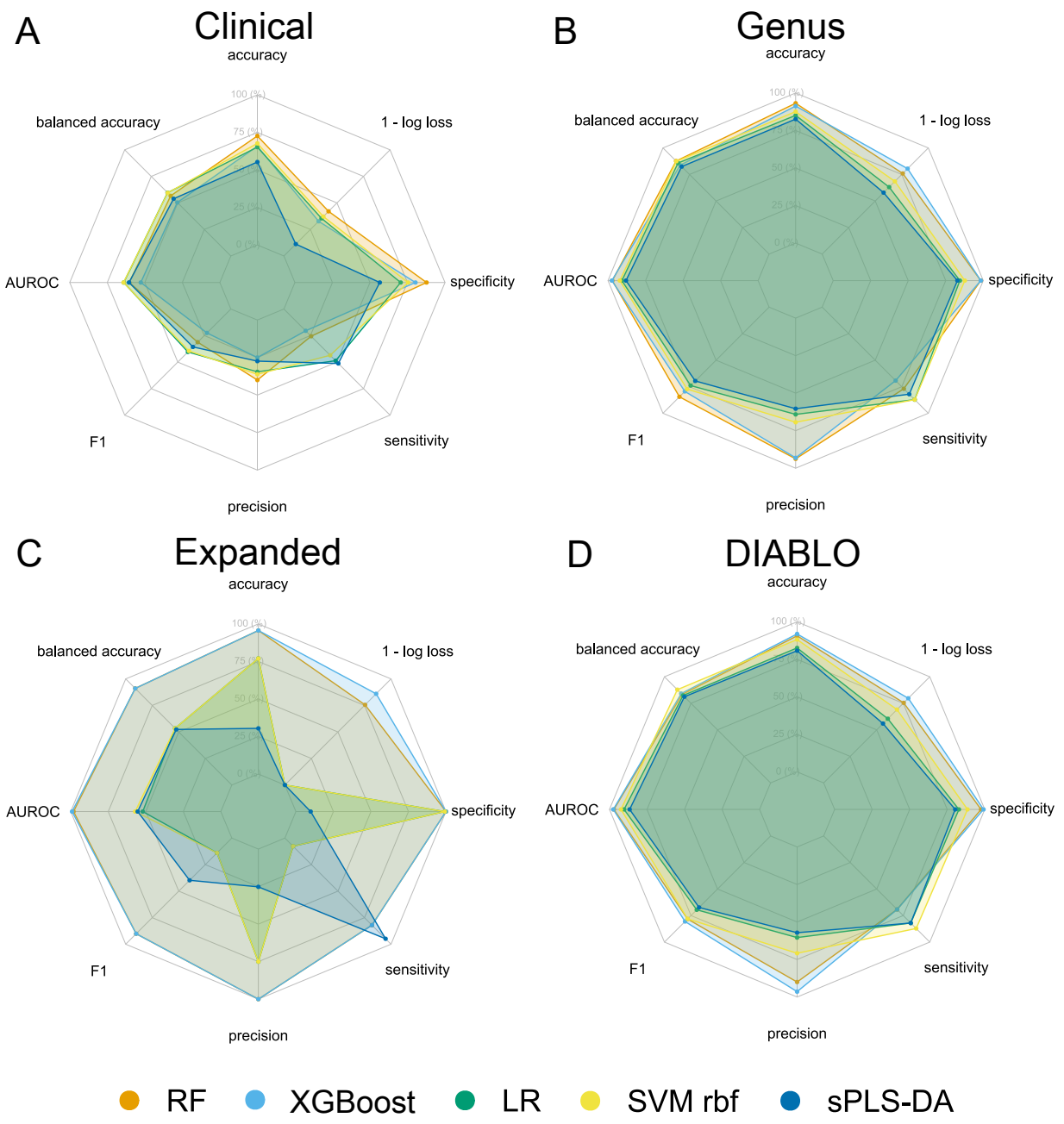


Figure 17: Radar charts of machine learning model performance for predicting feeding type (breastfeeding vs. formula feeding) at the age of one month with **(A)** clinical, **(B)** microbiome at genus level, **(C)** microbiome at expanded taxa level, and **(D)** DIABLO data fusion set of clinical and genus feature sets. The colours represent different machine learning models as indicated in the legend. All metrics are shown on the scale of percentages. The mean log loss is shown as "1 – log loss" as it has been inverted, such that higher values correspond to a better model, just as all other metrics. Any mean log loss higher than one was capped at one such that in this figure they cannot go below zero.

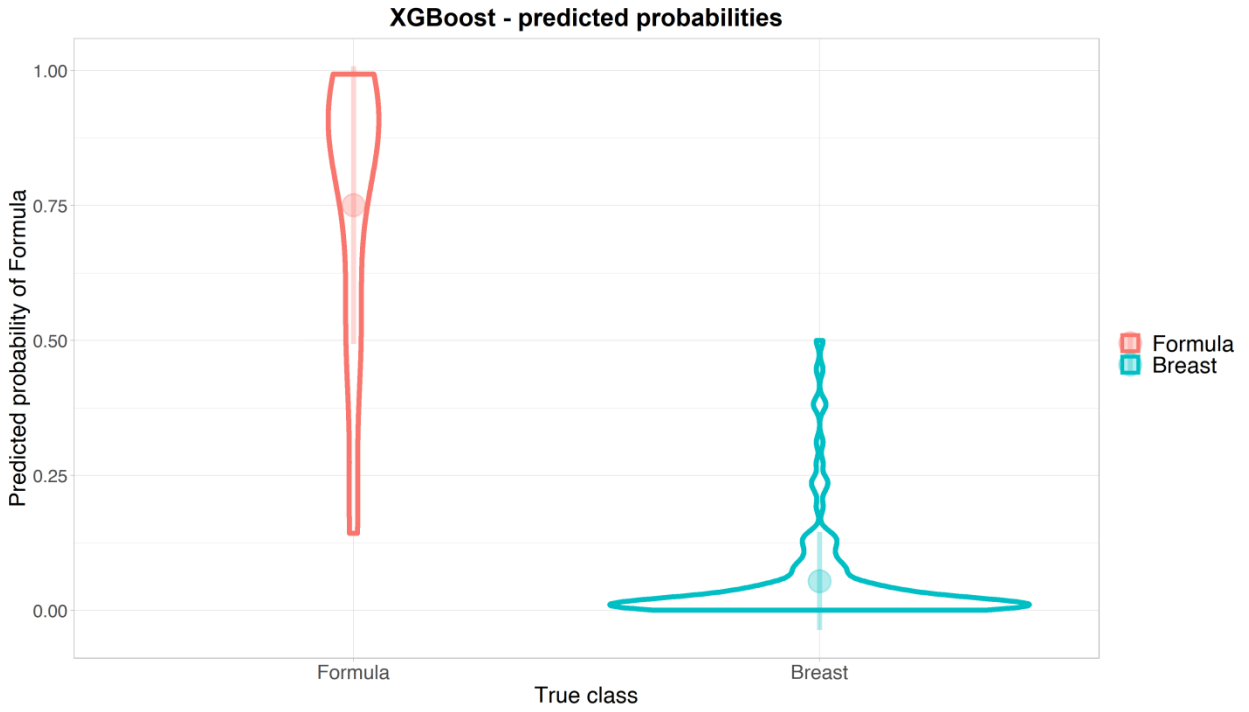


Figure 18: Violin plot of the output of XGBoost to predict feeding type (breast feeding vs. formula feeding) at the age of one month on the test set of the expanded feature set. The y-axis shows the output of the XGBoost model, which is the predicted probability of being formula fed. The x-axis separates the samples that were formula fed, and those that were breastfed. The dot in the middle shows the mean with the standard deviation indicated by the vertical line crossing it.

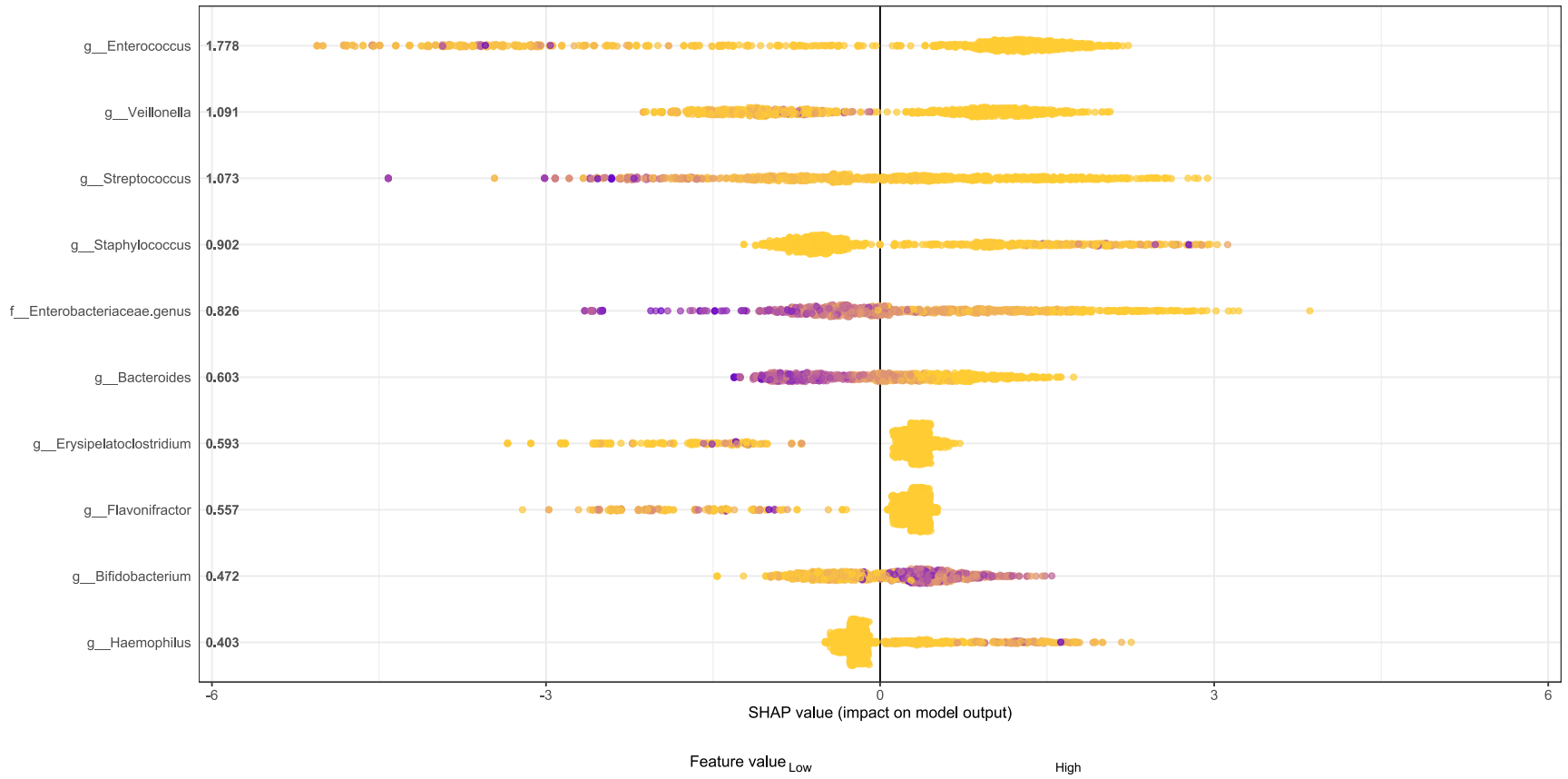


Figure 19: Shapley summary plot of XGBoost on the training set of microbiome genus feature set to predict feeding type at one month of age. The top 10 most important genera based on their mean Shapley value are shown on the y-axis. The mean Shapley value for each genus is shown on the inside of the y-axis. The "g__" prefix specifies that the name corresponds to the name of the genus. The "f__" indicates that the name corresponds the family of the genus, where the specific genus is unknown. The Shapley values, which indicate how important a variable was for the model to predict a specific sample's class, are shown on the x-axis. The direction of the Shapley value, being negative or positive, corresponds to a sample being predicted to be more likely asthmatic or non-asthmatic, respectively. The yellow-to-purple colour gradient corresponds to the actual value of

a given variable for a sample. As such, purple corresponds to samples with high relative abundance of a specific genus, and yellow depicts a low relative abundance.

4.2.6 Birth mode

For the prediction task of Birth mode, all models using the clinical feature set achieved an AUROC in between ~0.79 and ~0.83 as is shown in Figure 20. The overall model performance was lower on the microbiome feature sets. The models got AUROC values ranging from ~0.65 to ~0.77 on the genus feature set. However, precision and recall were below 0.5, except for LR and sPLS-DA, at the cost of a lower specificity. RF, XGBoost and sPLS-DA only performed slightly better on the expanded feature sets, based on their mean log loss. SVM performed slightly worse, whereas LR got much worse with respect to the mean log loss as it climbed to a high ~29.5.

The best performing ML model on the microbiome feature sets, with respect to the mean log loss, was XGBoost on the expanded feature set (mean log loss = 0.287). This model's output is visualised in Figure 21, which shows that there were some samples for which the model was certain that they were vaginally delivered, for these samples the model was correct. However, all samples got assigned a probability of being born through c-section below 0.5, meaning that none of the samples were classified as born through c-section.

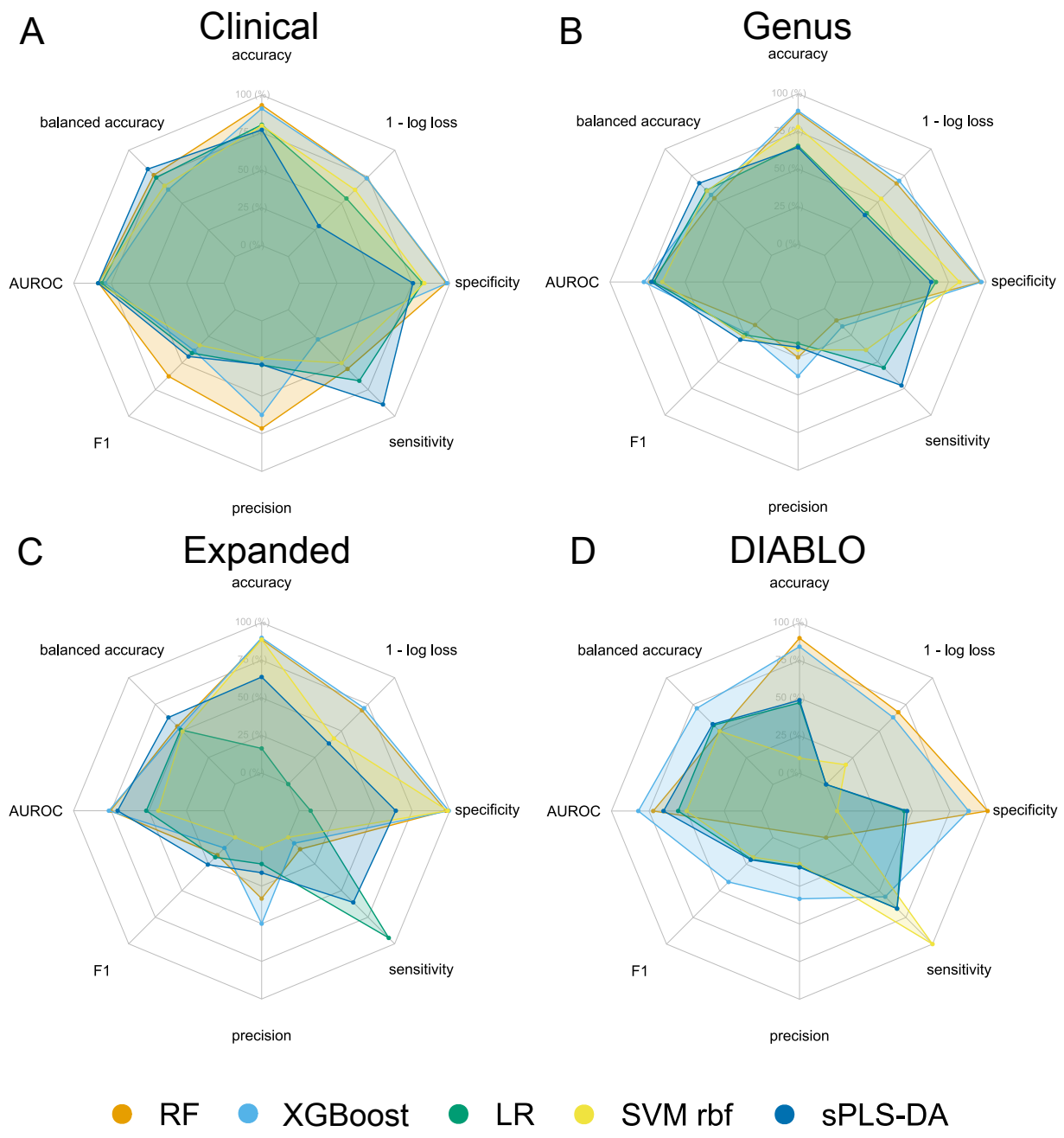


Figure 20: Radar charts of machine learning model performance for predicting birth mode (caesarean section vs. vaginal delivery) with **((A)** clinical, **(B)** microbiome at genus level, **(C)** microbiome at expanded taxa level feature, and **(D)** DIABLO data fusion set of clinical and genus feature sets. The colours represent different machine learning models as indicated in the legend. All metrics are shown on the scale of percentages. The mean log loss is shown as "1 – log loss" as it has been inverted, such that higher values correspond to a better model, just as all other metrics. Any mean log loss higher than one was capped at one such that in this figure they cannot go below zero.

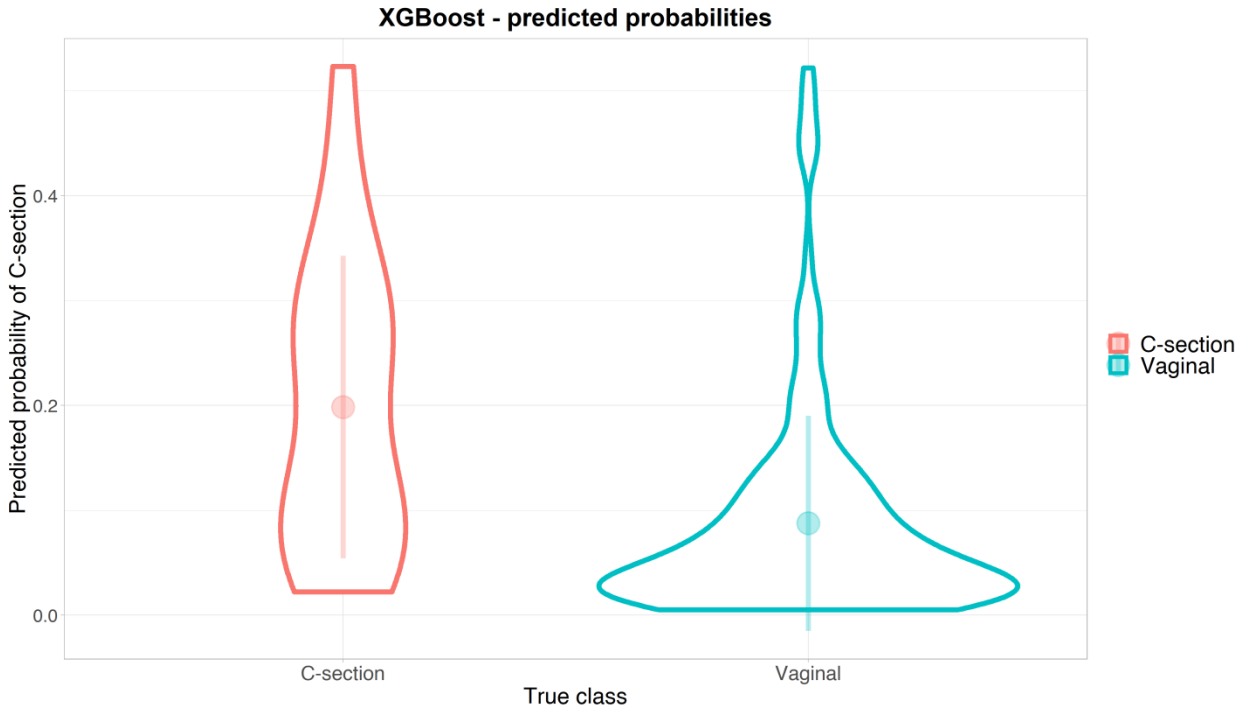


Figure 21: Violin plot of the output of XGBoost to predict birth mode (caesarean section vs. vaginal delivery) on the test set of the expanded feature set. The y-axis shows the output of the XGBoost model, which is the predicted probability of being born through caesarean section. The x-axis separates the samples that were born though caesarean section and those that were delivered vaginally. The dot in the middle shows the mean with the standard deviation indicated by the vertical line crossing it.

5. Discussion

5.1 Findings

This thesis investigated the possibility of predicting the development of asthma and eczema, using clinical and gut microbiota data of one month old infants. All models that were trained during this thesis had low predictive capacity towards both asthma and eczema, although previous research did identify associations between the gut microbiome of infants and later development of atopic conditions[5,6,8–10]. Only some predictive capacity was identified when prediction development of asthma for infants with atopic parents (AUROC = 0.67). However, the fact that this thesis was unable to provide a model capable of accurately identifying asthmatic or eczematic samples does not rule out the possibility of an interesting and complex interplay between the gut microbiome and the host's immunological development, which could eventually result in an atopic condition such as asthma or eczema.

A multitude of potential explanations are available for why little predictive performance was achieved in this thesis. First of which is the possibility that there is simply no, or too little, association between the gut microbiome composition in one month old infants and later development of atopy. Indeed, it does appear that the association between the microbial composition of the infant gut is not particularly strong, which is confirmed by the discrepancies in literature[17,18], as is already indicated in the introductory section 2.1. This implies that it is viable that the predictive models were ineffective in this classification task because the microbiome-atopy association is subtle and extremely heterogeneous due to the extensive number of (sparse) taxa involved, and the seemingly infinite number of different compositions they can form together. As a result, a large amount of data is required to model this complex system. This ties in with the second potential explanation for the lack of predictive performance: inadequacies in the data.

The data that was used for this thesis has some unique and advantageous characteristics, for example the follow-up questionnaires on the atopic conditions up until the age of 10 years, the combination of clinical and faecal collection data, and that for 1176 samples[3], of which 894 satisfied the inclusion criteria. However, the more complex a system is, the more data is required to adequately model it. Therefore, it could be assumed that given the complexity of the associations of atopic conditions with the gut microbiome, more observations are required to develop useful predictive models for atopic conditions. On top of this, the data was highly imbalanced for both the asthma and eczema prediction tasks. Although class imbalance is something that can be adjusted for, the lack of information on the minority class cannot be remedied as easily. Of course, there are techniques that generate new samples from already collected ones through synthetic up sampling, such as Synthetic Minority Oversampling Technique[88] and Adaptive Synthetic Sampling[89] but these do not allow for the collection of novel information. Therefore, if information on the system is lacking, this cannot be simply compensated for in silico. As such, a bottleneck for the predictive models could very well be the availability of merely 94 asthmatic and 158 eczematic (at 6 to 7 years) samples. However, to put this in perspective, there were 297 samples with eczema during the first two years, which could not be properly classified by any model (Figure 15), whereas there were only 198 formula fed samples, which could be classified almost perfectly (Figure 17). This suggests that a lack of minority samples is not the only reason for the lack of predictive capacity. Alternatively, it could very well be the combination of the two described potential reasons, that explain the lack in predictive performance.

As mentioned before, more complex systems require more data to be properly understood. Likewise, it is observed in this thesis that the number of samples available for the minority class alone, does not seem to explain how well ML models can classify it. This was illustrated by the high predictive capacity of feeding type and low performance on eczema, while less minority samples were available for feeding type. In this case however, feeding type is highly associated with the gut microbiome, while eczema's

association with the gut microbiome is not as well established[14–16]. Additionally, it is well known that eczema is a heterogeneous phenotype, occurring in different types and levels of severity, whereas for feeding type only purely breastfed or formula fed infants were used. However, this does not mean that the lack of minority samples for asthma and eczema is not an issue. This argument is strengthened by the prediction task of birth mode, for which the association with the gut microbiome has also been relatively well established[14–16]. Birth mode has a similar imbalance compared to asthma, with only 95 infants being delivered through c-section. For this particular classification task, it is observed that although there is some predictive capacity, the models generally struggle to identify the minority class using the microbiome data (Figure 20), illustrating the importance of plentiful minority samples. In summary, the imbalance could be a bottleneck for the ML models, particularly when the dependent variable does not exhibit a strong association with the gut microbiome composition.

Supervised ML methods are useful tools, not only because they can potentially predict various outcomes, but also since they can give insights into the underlying system[48]. This thesis used the Shapley value method to gain insights into these systems by looking at feature importance for each individual sample. There are multiple ways these Shapley values can be used to get insight into the decision making process of a specific model, but this thesis focussed solely on the Shapley values of the XGBoost models, and visualised these in Shapley summary plots as is illustrated in Figure 12 & Figure 19. Although the Shapley value in general is a model-agnostic technique, such that it can be used with any type of supervised ML model, ranging from LR to deep learning, it was decided here to use a type of Shapley value calculation that is based purely on tree learners as this offered improved computation time and handles multicollinearity to some extent[87,90]. However, it might be preferable in the future to use kernel SHAP, which is another approximation of the true Shapley value that is computationally efficient, model-agnostic, and can better handle dependent features according to Aas *et al.*, 2021[91].

Figure 19 shows that the XGBoost model generally associates higher levels of *Enterococcus*, *Veillonella*, *Streptococcus* and *Bacteroides* with formula fed infants. In contrast, it associates higher relative abundance of *Staphylococcus* and *Bifidobacterium* with breastfed infants. All of these are in accordance with previous literature[15,16,92]. Interestingly however, the four most important features (*Enterococcus*, *Veillonella*, *Streptococcus & Staphylococcus)* are only mentioned in one of the three papers used to compare microbiome associations with feeding type, namely by Guaraldi and Salvatoru, 2012[92]. Also, *Enterococcus* and *Veillonella* are not mentioned directly, only their subspecies *E. faecalis and V. parvula,* albeit with the same direction of association as indicated by the XGBoost model. Moreover, the model identifies an unknown genus belonging the family of *Enterobacteriaceae* as fifth most important genus. Since *E. coli* is a common genus belonging this family, and is known to be associated with feeding type[15], many of the sequencing reads assigned to this unclassified genus could potentially be *E. coli*. The reason why the genera that are most frequently associated with feeding type in literature are not the best predictors of feeding type for XGBoost might be because those genera, such as *Bifidobacterium* and *Bacteroides* often are always present but extremely abundant under some specific conditions (Figure 7), for instance being breast or formula fed. These differences will numerically be most pronounced as the numerical differences between these conditions will be bigger, given that these genera are simply more abundant. Rarer genera, that are generally less abundant, will have less pronounced differences between groups. Nevertheless, for the sake of classification, very sparse genera can be particularly useful, especially if they only occur, or only surpass some level of relative abundance, in one of the groups. This might be an oversimplification of the true association that occurs, yet the general concept is supported by Figure 19 as it clearly shows that the association for highly abundant genera is easily observable with the more purple samples on one side of the origin, whereas, the less abundant but very important genera, show a less obvious association. However, even slightly higher relative abundances seem to indicate to the model that they very likely belong to a specific class.

The results described in section 4.2.2 suggest that the microbiome association might be of particular interest in a specific subset of the infants. Specifically, infants with some atopy related hereditary background, as Figure 10 shows that the ML models surpass the 0.5 AUROC threshold for the prediction of asthma on infants of parents with atopic conditions using microbiome data. The results are far from perfect, with a maximum AUROC of 0.67, but still very little to no specificity as the models still struggle to identify the asthmatic samples. However, it should be considered that the models can identify higher risk infants to some extent, and that this subset of the data only contains 68 asthmatic samples. This finding of improved predictive capacity for infants with atopic parents is supported by the findings of Stokholm *et al.*, 2018[6], who found predictive capacity of 0.76 cross validated AUROC for infants of asthmatic mothers using an sPLS-DA model. Since the data used for this thesis only contains 24 asthmatic infants of asthmatic mothers, no direct comparison was made for this subgroup as such small numbers cannot give a reliable indication of true predictive performance, especially given the evaluation technique used here, namely using an independent test set of 20%, which would then only contain five asthmatic samples to evaluate the model on. Although Stokholm *et al.*, 2018[6], used cross-validation to determine the AUROC, the results were not confirmed on an independent test set, and the exact number asthmatic samples was not mentioned, they only state their total population size (N=589), the number of asthmatic samples in the total population (N=58), and the number of samples with asthmatic mothers (N=147). This stresses that, although some predictive capacity was identified, much more validation is required to make definitive claims on the actual predictive capacity. Also, this illustrates the importance of detailed descriptions of ML implementation in literature such that models can be appropriately compared and evaluated. Nevertheless, these findings do suggest some association between the gut microbiome of infants with a familial predisposition for asthma and later development of asthma, which should be further explored.

Another interesting phenomenon occurs for the prediction task of asthma using microbiome data. Namely, that the XGBoost model on the genus feature set to predict asthma (Figure 9) and on the expanded feature set to predict asthma for infants of atopic parents (Figure 11), seems to be able to identify some samples for which the model is relatively certain that they are non-asthmatic. Such findings can be very interesting for future research as even models with overall low predictive capacity, can identify interesting patterns in a selection of the samples. Here, the Shapley values could be inspected for the selection of samples that are evidently non-asthmatic, to see if a specific microbiome composition exists that is protective against asthma. However, it should be tested if this phenomenon is robust across different study populations, as it could also be an artifact of the underrepresentation of asthmatic samples due to the imbalance.

Figure 12 reveals some characteristics of the XGBoost model that was trained to predict asthma for infants with atopic parents. Prior to drawing any conclusions from the Shapley values, it is critical to realise that these values indicate how a model with low predictive capacity distinguishes asthmatics from non-asthmatics, and as such should not be treated as a representation of the "truth". Nevertheless, it can be interesting to compare the results with other research. Stokholm *et al.,* 2018[6], show that they have found higher levels of *Veillonella* to associate with development of asthma at 5 years. In children born to asthmatic mothers Conversely, they also identified for *Faecalibacterium*, *Bifidobacterium*, *Roseburia, Allistepes, Lachnospiraceae incertae sedis, Ruminococcus* and *Dialister* that lower relative abundance associates with asthma. Figure 12 shows that *Bifidobacterium* and *Veillonella* appear in the top 10 most important genera for XGBoost and are also associated with asthma according to Stokholm *et al.* However, the XGBoost model rates samples with very low relative abundance of *Veillonella* as more likely to asthmatic, whereas Stokholm *et al.* found a positive association between *Veillonella* and asthma. There are some additional discrepancies in the literature on the directionality of this association as another Canadian study[93] for example also shows a reduced amount of *Veillonella* in 3-month old infants to associate with higher risk of asthma. Furthermore,

this thesis shows that *Bifidobacterium* does not seem to follow a single directional association, instead it seems that the most extreme values, be it low or high relative abundances, promote the model to classify as non-asthmatic. This phenomenon could also be a reason why XGBoost identifies other genera as most important. Because XGBoost is non-linear and therefore not limited to one-directional associations, it can identify samples for which higher levels of some specific genus increases the likelihood of being asthmatic, and other samples for which lower levels of the same genus also increase likelihood of being asthmatic. The discrepancies between the results do not necessarily imply erroneous findings in Stokholm's nor this paper, also since this thesis looks at one month old infants while Stokholm *et al.* looked at 1 year old infants who also reside in a different geographical location. Moreover, they did not find any association with any genus and asthma for infants of one month of age[6], potentially indicating another limitation of this thesis, being the use of only a very early timepoint. In other words, caution should be used when comparing microbiome studies, as many aspects such as the subjects, timing, data collection, data preparation and analysis can each influence the results.

Ideally, predictive models for asthma and eczema should be able to identify high risk individuals as soon as possible, such that preventative measure can be taken at the optimal timepoint. This makes the early collection timepoint of the microbiome data potentially very useful and is also the reason why only clinical variables collected up until the age of one month were used. However, faecal collection at one month of age could be slightly too early. For example, exclusively breastfeeding for three or more months, which is also known to influence the gut microbiome composition, has been shown to have a protective effect against eczema during infancy for children with first-order atopic relatives[94]. Therefore, it is important to study the entire time span of the infant gut microbiome maturation, to identify aspects of the microbiome that either promote or protect against atopic manifestations, to allow for early identification of high-risk individuals, and to guide future research into probiotic use as atopy intervention.

5.2 Limitations

The work that was performed in this thesis demonstrates that the prediction of asthma and eczema using clinical and/or microbial data is a complex task. From the explorative data analysis, presented in section 4.1, it seems that the quality of the data was good, with little missing data and a representative train/test split. Additionally, imputed samples did not deviate from the distribution of non-imputed samples (Figure 2). Although, imputed samples ultimately should have been separately projected onto a PCA of only non-imputed samples, Figure 2 still shows a useful approximation of the distribution of imputed samples. When looking at some of the properties of the data, it already becomes clear, even prior to running any models, that it might prove to be challenging to predict asthma and eczema. From the correlation heatmap in Figure A1 it can be observed that some clinical variables correlate only very slightly with asthma (spearman correlation), with the highest correlation coefficient just surpassing 0.2 for having an asthmatic mother. Additionally, the FAMD score plot shown in Figure 3 illustrates the incapability of the first two dimensions to separate asthmatics from non-asthmatics. Although not shown in this thesis, no other dimensions were identified that could separate the classes. The URF cMDS plot in Figure 4-B also cannot discriminate asthmatics from non-asthmatics or eczematics from non-eczematics. This indicates that classification using clinical data will most likely not be very accurate. With respect to the microbiome data, similar properties are observed. First, no difference in α-diversity was observed between asthmatics and non-asthmatics nor between eczematics and non-eczematics. Also, PCA (Figure 6) and URF cMDS (Figure 4-A) do not reveal apparent differences between asthmatics and non-asthmatics. Although, this does not mean that the classes cannot be separated, as these explorative techniques are all unsupervised and therefore not designed to separate the classes, it does suggest that the classification could be very difficult.

There are multiple reasons why this prediction task proved so challenging. First, the similarities between asthmatic and non-asthmatic, as well as eczematic and non-eczematic samples, based on the α-diversities and PCs, already suggests that separation of the classes could be challenging. Therefore, the models will presumably need to identify complex patterns in the data to achieve good predictive performance. To train such a complex model, also more data should be available, where the biggest limitation in this thesis project was the lack of asthmatic and eczematic samples. Another factor that might make individual taxa associations less pronounced is the very early time point of data collection at one month of age. The instability of the gut microbiome at such a young age could make this prediction task particularly challenging. Ideally, multiple timepoints should be available as that would allow for the identification of the best trade-off between predictive capacity and timing of the faecal collection. Moreover, this would allow to take microbial dynamics into account, as the development of ecosystem over time could also reveal important aspects of the system. Another limitation for microbiome data in general is the use of relative abundances. Although, it is a convenient way of handling the taxonomic counts that are not representative of their true abundances, it can be a difficult unit to handle. There is the drawback of not knowing how dense the microbiota ecosystem in an individual's gut is, and thus how much of a given taxa is truly active in their gut[95]. The best way of scaling relative abundances also remains an open discussion, and the fact that relative abundances were not scaled for prediction in this thesis, might have put scale-invariant models in a preferable position. However, collection of absolute abundance is more expensive and time consuming, which is a major disadvantage in large studies. Also, the best technique of quantitative microbiome profiling is still highly debated[96,97]. Additionally, the high level of multicollinearity in the clinical and expanded feature sets will inevitably favour the models that can handle multicollinearity, such as RF and XGBoost, over methods as SVM and LR. This effect can also often be observed in the radar charts (e.g., Figure 17), when comparing model performances on the genus and expanded feature sets. Here, RF and XGBoost often

benefit from the extra features but higher multicollinearity in the expanded feature set, while LR, SVM and sPLS-DA drop in performance when extra features and multicollinearity are introduced. Such model inequalities should be considered when determining the optimal model for a specific classification task. As this thesis' aim was to explore the predictive potential of various models on these prediction tasks, rather than perfecting the best model, limited search spaces were explored in the hyperparameter optimisation, and due to limited computation power and time, less repeats were performed than would be ideal. Especially XGBoost suffers here, for which the cross validation was not repeated. Although this is a limitation, the models and their evaluations are still very useful for the purpose of this thesis. Also, the evaluation of models was tailored for the explorative setting, where each model was validated on only one test set. The first limitation here is the size of the test set, only containing roughly 20% of the total number of minority and majority samples indicated in Table 2. Ideally, a bigger test set to validate the models should be acquired, and cross-validated model evaluation should be performed to also gain information on the variance of the performance of a model when evaluated on different samples. Finally, the Shapley values that were shown and discussed are based on the training set. As a result, they do not reflect the samples on which the final performance of the model was based. Rather, they give insight into the training of the model, and the decisions that were constructed by XGBoost. It could be argued that Shapley values should be calculated on the test set, so they correspond to the model's decisions on unseen data. However, this thesis is more focussed on whether models can find predictive performance, and if so, where this comes from. Therefore, it was decided to use Shapley values based on the training set as these describe the training process of the model and have more samples available to compute Shapley values for.

The DIABLO data fusion technique, integrating the clinical and microbiome genera feature sets into one data block, did not improve predictive capacity in a useful way. This could be because of the lack of predictive performance of the individual blocks for the

prediction of asthma and eczema. However, also no improvement was observed for the prediction task of birth mode, for which both blocks had some predictive capacity, possibly because DIABLO has not yet been optimised for use with categorical variables. Another limitation could be the linear nature of DIABLO, making it incapable of identifying non-linear inter-block relationships.

In summary, the biggest limitation of this thesis project lies in the data. Ideally, more minority samples should be available to train and evaluate the models on. Also, the properties of the gut microbiome data, being relative abundances at only one timepoint, make the prediction task more challenging. Finally, some additional information, such as genetics might be lacking which could potentially influence the role of the gut microbiome in its association with atopy.

5.3 Future directions

Although, data fusion did not result in any strong improvements in predictive capacity for asthma or eczema, it could still play a key role in unravelling the role of the gut microbiome in the development of atopic conditions. However, to this end, more research is required. First, the inclusion of genetic information could shed light on person specific gut microbiome interactions. This can be expected as we observed an increase in predictive capacity for infants with an hereditary background in atopy, and as Morgan *et al.*, 2014[40], for example have also shown that genetics can influence the effectiveness of probiotics for eczema intervention. However, such a holistic approach makes the models far more complex, meaning that plentiful data is required to make sense out of the many human-microbe interactions. This is clearly a challenge as the data collection is difficult, time consuming and expensive. To remedy this, it might be beneficial to also investigate data integration on the sample level, where samples from different studies are fused to increase sample size. Unfortunately, this could introduce more challenges than it solves, as this could drastically increase the heterogeneity of the sample space, due to

differences in e.g., sample geography, collection techniques, time points and diet. Nevertheless, it would be interesting to attempt in practice. Also, different methods for data fusion should be explored, as DIABLO is only one of many ways of integrating two or more separate data blocks.

Additionally, subtypes of the atopic conditions should be explored. For example, IgE levels can be used to investigate only allergic types of asthma and eczema. Moreover, other atopic conditions could be explored. This thesis project had too few samples available with other atopic conditions to explore, but it could for example also be interesting to see if atopy in general can be classified, rather than zooming in on specific types of atopy as these might be largely overlapping. Conclusively, future research should focus on improving the data rather than the models. Only when the appropriate data is available to generate predictive models with some predictive capacity, does it make sense to shift the focus towards finetuning the models themselves.

6. Conclusion

This thesis evaluated a series of supervised ML models on the prediction task of asthma, eczema, feeding type and birth mode, using feature sets containing clinical, microbiome, and the combination of both clinical and microbiome features. No predictive capacity was identified for asthma and eczema on the complete study population. However, some predictive capacity was found when predicting asthma for the subset of infants with atopic parents (AUROC = 0.67). Correct model implementation was confirmed by the prediction task of feeding type, which achieved a high maximum AUROC of 0.99 using XGBoost on the expanded microbiome feature set. This thesis illustrates the complexity of the gut microbiome, and its association with atopic conditions such as asthma and eczema. Although, little to no predictive capacity was identified for atopic conditions, the potential use of ML models to reveal novel insights into the microbiome-atopy association, and

their possible application to aid in early identification of high-risk individuals, persists. Presumably, the lack asthmatic and eczematic samples was the biggest bottleneck in this thesis as ML models perform best on large data sets, especially for complex prediction tasks. Furthermore, the fact that only predictive capacity was identified for infants with an atopic genetic background, strengthens the argument that there is an interplay between the genome and microbiome, which associates with later development of asthma, and possibly other atopic conditions. As such, future research is encouraged to further explore data fusion techniques to incorporate genomics into the feature space and potentially combine different study populations to increase sample size.

7. Acknowledgements

David Barnett for the great support throughout the thesis project.

Ilja Arts & John Penders for their valuable feedback and inviting me into their scientific communities.

Stefan Meier for the countless discussions on machine learning.

Rachel Cavill for her input on data fusion techniques.

Carel Thijs for interesting discussions regarding the KOALA cohort and atopic conditions.

Borewicz *et al.*[63] *&* Martha Endika for sequencing of the data.

8. Reference list

1. Beasley, R. Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. *The Lancet* **351**, 1225–1232 (1998).

2. Bleecker, E. R., Postma, D. S. & Meyers, D. A. Genetic susceptibility to asthma in a changing environment. *Ciba Found. Symp.* **206**, 90–99; discussion 99-105, 106–110 (1997).

3. Kummeling, I. *et al.* Etiology of atopy in infancy: the KOALA Birth Cohort Study. *Pediatr. Allergy Immunol. Off. Publ. Eur. Soc. Pediatr. Allergy Immunol.* **16**, 679–684 (2005).

4. Strachan, D. Family size, infection and atopy: the first decade of the 'hygiene hypothesis'. *Thorax* **55**, 2S – 10 (2000).

5. Penders, J. *et al.* Gut microbiota composition and development of atopic manifestations in infancy: the KOALA Birth Cohort Study. *Gut* **56**, 661–667 (2007).

6. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).

7. Penders, J. *et al.* New insights into the hygiene hypothesis in allergic diseases: Mediation of sibling and birth mode effects by the gut microbiota. *Gut Microbes* **5**, 239–244 (2014).

8. Bengt Björkstén, Sepp, E., Julge, K., Voor, T. & Mikelsaar, M. Allergy development and the intestinal microflora during the first year of life. *J. Allergy Clin. Immunol.* **108**, 516–520 (2001).

9. Galazzo, G. *et al.* Development of the Microbiota and Associations With Birth Mode, Diet, and Atopic Disorders in a Longitudinal Analysis of Stool Samples, Collected From Infancy Through Early Childhood. *Gastroenterology* **158**, 1584–1596 (2020).

10. Bisgaard, H. *et al.* Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. *J. Allergy Clin. Immunol.* **128**, 646-652.e5 (2011).

11. Milani, C. *et al.* The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiol. Mol. Biol. Rev.* **81**, e00036-17, e00036-17 (2017).

12. Noverr, M. C. & Huffnagle, G. B. Does the microbiota regulate immune responses outside the gut? *Trends Microbiol.* **12**, 562–568 (2004).

13. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14**, (2016).

14. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690–703 (2015).

15. Penders, J. *et al.* Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy. *PEDIATRICS* **118**, 511–521 (2006).

16. Fallani, M. *et al.* Intestinal Microbiota of 6-week-old Infants Across Europe: Geographic Influence Beyond Delivery Mode, Breast-feeding, and Antibiotics. *J. Pediatr. Gastroenterol. Nutr.* **51**, 77–84 (2010).

17. Pedersen, E., Skov, L., Thyssen, J. & Jensen, P. Role of the Gut Microbiota in Atopic Dermatitis: A Systematic Review. *Acta Derm. Venereol.* 0 (2018) doi:10.2340/00015555-3008.

18. Abdel-Aziz, M. I., Vijverberg, S. J. H., Neerincx, A. H., Kraneveld, A. D. & Zee, A. H. M. der. The crosstalk between microbiome and asthma: Exploring associations and challenges. *Clin. Exp. Allergy* **49**, 1067–1086 (2019).

19. Kuruvilla, M. E., Lee, F. E.-H. & Lee, G. B. Understanding Asthma Phenotypes, Endotypes, and Mechanisms of Disease. *Clin. Rev. Allergy Immunol.* **56**, 219–233 (2019).

20. Lundbäck, B., Backman, H., Lötvall, J. & Rönmark, E. Is asthma prevalence still increasing? *Expert Rev. Respir. Med.* **10**, 39–51 (2016).

21. Eder, W. The Asthma Epidemic. *N Engl J Med* 10 (2006).

22. Nurmagambetov, T., Kuwahara, R. & Garbe, P. The Economic Burden of Asthma in the United States, 2008–2013. *Ann. Am. Thorac. Soc.* **15**, 348–356 (2018).

23. Mukherjee, M. *et al.* The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases. *BMC Med.* **14**, 113 (2016).

24. Barnes, P. J., Jonsson, B. & Klim, J. B. The costs of asthma. *Eur. Respir. J.* **9**, 636–642 (1996).

25. Aaron, S. D., Boulet, L. P., Reddel, H. K. & Gershon, A. S. Underdiagnosis and Overdiagnosis of Asthma. *Am. J. Respir. Crit. Care Med.* **198**, 1012–1020 (2018).

26. Pedersen, S. Early use of inhaled steroids in children with asthma. *Clin. Exp. Allergy* **27**, 995–1006 (1997).

27. Heffler, E. *et al.* Inhaled Corticosteroids Safety and Adverse Effects in Patients with Asthma. *J. Allergy Clin. Immunol. Pract.* **6**, 776–781 (2018).

28. Guilbert, T. W. *et al.* Long-term inhaled corticosteroids in preschool children at high risk for asthma. *N. Engl. J. Med.* **354**, 1985–1997 (2006).

29. Nielsen, K. G. & Bisgaard, H. The effect of inhaled budesonide on symptoms, lung function, and cold air and methacholine responsiveness in 2- to 5-year-old asthmatic children. *Am. J. Respir. Crit. Care Med.* **162**, 1500–1506 (2000).

30. Busse, W. W. *et al.* The Inhaled Steroid Treatment As Regular Therapy in Early Asthma (START) study 5-year follow-up: effectiveness of early intervention with

budesonide in mild persistent asthma. *J. Allergy Clin. Immunol.* **121**, 1167–1174 (2008).

31. Haahtela, T. *et al.* Thirteen-year follow-up of early intervention with an inhaled corticosteroid in patients with asthma. *J. Allergy Clin. Immunol.* **124**, 1180–1185 (2009).

32. Luo, G., Nkoy, F. L., Stone, B. L., Schmick, D. & Johnson, M. D. A systematic review of predictive models for asthma development in children. *BMC Med. Inform. Decis. Mak.* **15**, (2015).

33. Yoshihara, S. *et al.* Early intervention with suplatast tosilate for prophylaxis of pediatric atopic asthma: A pilot study. *Pediatr. Allergy Immunol.* **20**, 486–492 (2009).

34. Fukuhara, K. *et al.* Suplatast tosilate protects the lung against hyperoxic lung injury by scavenging hydroxyl radicals. *Free Radic. Biol. Med.* **106**, 1–9 (2017).

35. Warner, J. O. & ETAC Study Group. Early Treatment of the Atopic Child. A double-blinded, randomized, placebo-controlled trial of cetirizine in preventing the onset of asthma in children with atopic dermatitis: 18 months' treatment and 18 months' posttreatment follow-up. *J. Allergy Clin. Immunol.* **108**, 929–937 (2001).

36. Rożalski, M., Rudnicka, L. & Samochocki, Z. Atopic and Non-atopic Eczema. *Acta Dermatovenerol. Croat. ADC* **24**, 110–115 (2016).

37. Kim, J., Kim, B. E. & Leung, D. Y. M. Pathophysiology of atopic dermatitis: Clinical implications. *Allergy Asthma Proc.* **40**, 84–92 (2019).

38. Kim, B. E. & Leung, D. Y. M. Significance of Skin Barrier Dysfunction in Atopic Dermatitis. *Allergy Asthma Immunol. Res.* **10**, 207 (2018).

39. Silverberg, J. I. Public Health Burden and Epidemiology of Atopic Dermatitis. *Dermatol. Clin.* **35**, 283–289 (2017).

40. Morgan, A. R. *et al.* Differential modification of genetic susceptibility to childhood eczema by two probiotics. *Clin. Exp. Allergy* **44**, 1255–1265 (2014).

41. Allen, S. J. *et al.* Probiotics in the prevention of eczema: a randomised controlled trial. *Arch. Dis. Child.* **99**, 1014–1019 (2014).

42. Isolauri, E., Arvola, T., SÜtas, Y., Moilanen, E. & Salminen, S. Probiotics in the management of atopic eczema. *Clin. Exp. Allergy* **30**, 1605–1610 (2000).

43. Boyle, R. J., Bath-Hextall, F. J., Leonardi-Bee, J., Murrell, D. F. & Tang, M. L.-K. Probiotics for the treatment of eczema: a systematic review. *Clin. Exp. Allergy* **39**, 1117–1127 (2009).

44. Abrahamsson, T. R. *et al.* Probiotics in prevention of IgE-associated eczema: A double-blind, randomized, placebo-controlled trial. *J. Allergy Clin. Immunol.* **119**, 1174–1180 (2007).

45. Marras, L. *et al.* The Role of Bifidobacteria in Predictive and Preventive Medicine: A Focus on Eczema and Hypercholesterolemia. *Microorganisms* **9**, 836 (2021).

46. Caudri, D. *et al.* Predicting the long-term prognosis of children with symptoms suggestive of asthma at preschool age. *J. Allergy Clin. Immunol.* **124**, 903-910.e7 (2009).

47. Hacılar, H., Nalbantoğlu, O. U. & Bakir-Güngör, B. Machine Learning Analysis of Inflammatory Bowel Disease-Associated Metagenomics Dataset. in *2018 3rd International Conference on Computer Science and Engineering (UBMK)* 434–438 (2018). doi:10.1109/UBMK.2018.8566487.

48. Eck, A. *et al.* Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. *BMC Bioinformatics* **18**, 441 (2017).

49. Fukui, H. *et al.* Usefulness of Machine Learning-Based Gut Microbiome Analysis for Identifying Patients with Irritable Bowels Syndrome. *J. Clin. Med.* **9**, (2020).

50. Hollister, E. B. *et al.* Leveraging Human Microbiome Features to Diagnose and Stratify Children with Irritable Bowel Syndrome. *J. Mol. Diagn. JMD* **21**, 449–461 (2019).

51. Marcos-Zambrano, L. J. *et al.* Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. *Front. Microbiol.* **12**, (2021).

52. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2020).

53. RStudio Team. *RStudio: Integrated Development Environment for R*. (RStudio, PBC, 2021).

54. Ushey, K. *renv: Project Environments*. (2020).

55. Gaujoux, R. *doRNG: Generic Reproducible Parallel Backend for 'foreach' Loops*. (2020).

56. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

57. Scheepers, L. E. J. M. *et al.* The intestinal microbiota composition and weight development in children: the KOALA Birth Cohort Study. *Int. J. Obes.* **39**, 16–25 (2015).

58. Williams, H. C. *et al.* The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. I. Derivation of a minimum set of discriminators for atopic dermatitis. *Br. J. Dermatol.* **131**, 383–396 (1994).

59. Williams, H. C., Jburney, P. G., Strachan, D. & Hay, R. J. The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis II. Observer variation of clinical diagnosis and signs of atopic dermatitis. *Br. J. Dermatol.* **131**, 397–405 (1994).

60. Williams, H. C., Burney, P. G., Pembroke, A. C. & Hay, R. J. The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. III. Independent hospital validation. *Br. J. Dermatol.* **131**, 406–416 (1994).

61. Honari, G. Clinical Scoring of Atopic Dermatitis. in *Agache's Measuring the Skin* (eds. Humbert, P., Maibach, H., Fanian, F. & Agache, P.) 1–10 (Springer International Publishing, 2016). doi:10.1007/978-3-319-26594-0_94-1.

62. Poncheewin, W. *et al.* NG-Tax 2.0: A Semantic Framework for High-Throughput Amplicon Analysis. *Front. Genet.* **10**, (2020).

63. Borewicz, K. *et al.* Correlating Infant Fecal Microbiota Composition and Human Milk Oligosaccharide Consumption by Microbiota of 1-Month-Old Breastfed Infants. *Mol. Nutr. Food Res.* **63**, 1801214 (2019).

64. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).

65. Barnett, D. *microViz: Microbiome Data Analysis and Visualization*. (2021). doi:10.5281/zenodo.4667069.

66. Snee, R. D. Validation of Regression Models: Methods and Examples. *Technometrics* **19**, 415–428 (1977).

67. Kuhn, M., Chow, F. & Wickham, H. *rsample: General Resampling Infrastructure*. (2020).

68. Stevens, A. & Ramirez-Lopez, L. *An introduction to the prospectr package*. (2020).

69. Stekhoven, D. J. & Buhlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).

70. Kuhn, M. & Wickham, H. *recipes: Preprocessing Tools to Create Design Matrices*. (2020).

71. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13**, e1005752 (2017).

72. Lahti, L. & Shetty, S. *microbiome R package*. (2012).

73. Oksanen, J. *et al. vegan: Community Ecology Package*. (2019).

74. Weiner, J. *pca3d: Three Dimensional PCA Plots*. (2020).

75. Lê, S., Josse, J. & Husson, F. FactoMineR: A Package for Multivariate Analysis. *J. Stat. Softw.* **25**, 1–18 (2008).

76. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).

77. Shi, T. & Horvath, S. Unsupervised Learning With Random Forest Predictors. 22.

78. Kuhn, M. & Wickham, H. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* (2020).

79. Kuhn, M. & Vaughan, D. *parsnip: A Common API to Modeling and Analysis Functions*. (2020).

80. Kuhn, M. *tune: Tidy Tuning Tools*. (2020).

81. Corporation, M. & Weston, S. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. (2020).

82. Bengtsson, H. *A Unifying Framework for Parallel and Distributed Processing in R using Futures*. (2020).

83. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

84. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).

85. Chen, T. *et al. xgboost: Extreme Gradient Boosting*. (2021).

86. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab – An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **11**, 1–20 (2004).

87. Liu, Y. & Just, A. *SHAPforxgboost: SHAP Plots for 'XGBoost'*. (2021).

88. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

89. He, H., Bai, Y., Garcia, E. A. & Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *Int. Jt. Conf. Neural Netw.* 7 (2008).

90. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

91. Aas, K., Jullum, M. & Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.* **298**, 103502 (2021).

92. Guaraldi, F. & Salvatori, G. Effect of Breast and Formula Feeding on Gut Microbiota Shaping in Newborns. *Front. Cell. Infect. Microbiol.* **2**, (2012).

93. Arrieta, M.-C. *et al.* Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **7**, 307ra152-307ra152 (2015).

94. Gdalevich, M., Mimouni, D., David, M. & Mimouni, M. Breast-feeding and the onset of atopic dermatitis in childhood: A systematic review and meta-analysis of prospective studies. *J. Am. Acad. Dermatol.* **45**, 520–527 (2001).

95. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).

96. Galazzo, G. *et al.* How to Count Our Microbes? The Effect of Different Quantitative Microbiome Profiling Approaches. *Front. Cell. Infect. Microbiol.* **10**, (2020).

97. Jian, C., Salonen, A. & Korpela, K. Commentary: How to Count Our Microbes? The Effect of Different Quantitative Microbiome Profiling Approaches. *Front. Cell. Infect. Microbiol.* **11**, (2021).

98. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **6**, 21–45 (2006).

99. Dietterich, T. G. Ensemble Methods in Machine Learning. in *Multiple Classifier Systems* 1–15 (Springer, 2000). doi:10.1007/3-540-45014-9_1.

100. Friedman, J., Hastie, T. & Tibshirani, R. ADDITIVE LOGISTIC REGRESSION: A STATISTICAL VIEW OF BOOSTING. 71.

101. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016). doi:10.1145/2939672.2939785.

102. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **24**, 1565–1567 (2006).

103. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

104. Vert, JP., Tsuda, K. & Schölkopf, B. A Primer on Kernel Methods. in *Kernel Methods in Computational Biology* 35–70 (Biologische Kybernetik, 2004).

105. Abdi, H. Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Comput. Stat.* **2**, 97–106 (2010).

106. Cao, K.-A. L., Rossouw, D., Robert-Granié, C. & Besse, P. A Sparse PLS for Variable Selection when Integrating Omics Data. *Stat. Appl. Genet. Mol. Biol.* **7**, (2008).

107. Pérez-Enciso, M. & Tenenhaus, M. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. 12 (2003).

108. Lê Cao, K.-A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253 (2011).

109. Ribeiro, M. T., Singh, S. & Guestrin, C. Model-Agnostic Interpretability of Machine Learning. *ArXiv160605386 Cs Stat* (2016).

110. Afanador, N. L., Smolinska, A., Tran, T. N. & Blanchet, L. Unsupervised random forest: a tutorial with case studies. 10 (2016).

111. Bro, R. *et al.* Data fusion in metabolomic cancer diagnostics. *Metabolomics* **9**, 3–8 (2013).

112. Kuligowski, J. *et al.* Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE). *Analyst* **140**, 4521–4529 (2015).

113. Borràs, E. *et al.* Data fusion methodologies for food and beverage authentication and quality assessment - a review. *Anal. Chim. Acta* **891**, 1–14 (2015).

114. Smolinska, A., Engel, J., Szymanska, E., Buydens, L. & Blanchet, L. Chapter 3 - General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences. in *Data Handling in Science and Technology* (ed. Cocchi, M.) vol. 31 51–79 (Elsevier, 2019).

115. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).

116. Tenenhaus, A. *et al.* Variable selection for generalized canonical correlation analysis. *Biostatistics* **15**, 569–583 (2014).

117. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).

9. Appendix

9.1 Abbreviations

Abbreviation	Definition
ASV	Amplicon Sequencing Variant
AUROC	Area Under the Receiver Operating Characteristic
CLR	Centre-Log-Ratio
cMDS	classical Multidimensional scaling
DIABLO	Data Integration Analysis for Biomarker discovery using Latent cOmponents
FAMD	Factor Analysis of Mixed Data
LR	Logistic Regression
ML	Machine Learning
PCA	Principal Component Analysis
PCoA	Principal Coordinates Analysis
PLS	Partial Least Squares (also known as Projection to Latent Structures)
rbf	radial basis function
RF	Random Forest
ROC	Receiver Operating Characteristics
sGCCA	sparse Generalized Canonical Correlation Analysis
sPLS-DA	sparse Partial Least Squares Discriminant Analysis
SVM	Support Vector Machine
URF	Unsupervised Random Forest
XGBoost	eXtreme Gradient Boosting

9.2 Methodology

9.2.1 Machine Learning (ML) methods

9.2.1.1 Logistic Regression (LR)

The first ML method that was used in this thesis project was Logistic Regression (LR). One of the reasons why LR was included, is because it is already popular among predictive asthma models[32]. Also, the method is relatively simple and computationally inexpensive compared to other ML methods, making it an appropriate model to start with. One benefit of LR models is that the effect of specific independent variables on the prediction of the dependent variables can be easily determined, making the model easily

interpretable. On the downside however, LR models require the data to be linearly separable for it to perform well in the prediction task. Also, the model cannot handle multicollinearity very well, therefore, caution should be used especially when interpreting the contribution of a specific variable in the prediction task if it correlates with one or more other independent variables.

9.2.1.2 Random Forest (RF)

Random Forest (RF) falls under the category of ensemble learners, which means that rather than fitting one optimal model on a specific prediction task, a multitude of models, also referred to as weak learners, are trained. For an RF, these weak learners are decision trees. Eventually, the RF model will base its final classification on all the predictions of the weak learners combined. The approach with which RF generates these weak learners is called bagging, where each weak learner is trained with a bootstrapped subset of the data, meaning that each decision tree is trained on a random selection of the samples, where each sample can be selected more than once in the same subset. Additionally, it is possible to train each weak learner on a random subset of the variables. This bagging approach, where weak learners are trained on subsets of the samples and variables, improves the generalisability of the model such that it improves the model's performance on new data[98,99]. Although RF models generally take longer to train, and especially to tune, they have various benefits compared to other methods such as LR. Namely, RF models can handle multicollinearity, high dimensionality, missing values, and outliers. Additionally, they are relatively resistant against overfitting and can also be used for feature importance extraction, improving the interpretability of the models.

9.2.1.3 eXtreme Gradient Boosting (XGBoost)

Much like the RF ensemble learner described in section 9.2.1.2, eXtreme Gradient Boosting (XGBoost) is also an ensemble learner that trains multiple decision trees as weak learners to perform a classification task. Whereas RF uses the bagging approach, which trains all the weak learners in parallel, each on a different subset but with the exact same objective, XGBoost utilises the gradient boosting approach. Gradient boosting implies that the weak learners are trained sequentially such that each weak learner is specifically trained to improve the classification of samples that the previously trained weak learner was not able to correctly classify, by giving those samples a higher weight[100]. Since XGBoost supports parallelisation, it is also very computationally efficient[101]. However, since XGBoost has more tuneable hyperparameters than RF, the hyperparameter optimisation can take much longer depending on the size of the grid search.

9.2.1.4 Support Vector Machine (SVM)

The general concept of a Support Vector Machine (SVM) is that it identifies a hyperplane that can best separate the dependent variable. Here a hyperplane is nothing more than a separator, in the sense that for an n-dimensional data set, the hyperplane will have n-1 dimensions. For example, to separate datapoints in a 2-dimensional space, a 1-dimensional line is required. Similarly, you can separate a 3-dimensional volume, with a 2-dimensional plane. An SVM chooses the best hyperplane by maximising the orthogonal distance of the hyperplane to the closest samples to the hyperplane, also referred to as support vectors. Figure 22 shows a simplified diagram of how an SVM works. It shows that the hyperplane is based on the middle of the margin, which maximises the distance of the hyperplane to the support vectors. Additionally, Figure 22 shows two misclassified samples, the orange sample above the hyperplane, and the blue sample below the hyperplane. SVM has a cost hyperparameter which is essentially the regularisation factor

of SVM, where higher cost values increase model complexity, as a bigger penalty will be given to misclassifications. Lower cost values make the model less complex, where misclassification, as the two in the diagram, do not get as much of a weight in determining the optimal margin. As a result, a too high cost will result in overfitting, and cost values that are too low, result in underfitting, which is why this hyperparameter should be optimised for the best generalised performance[102,103].

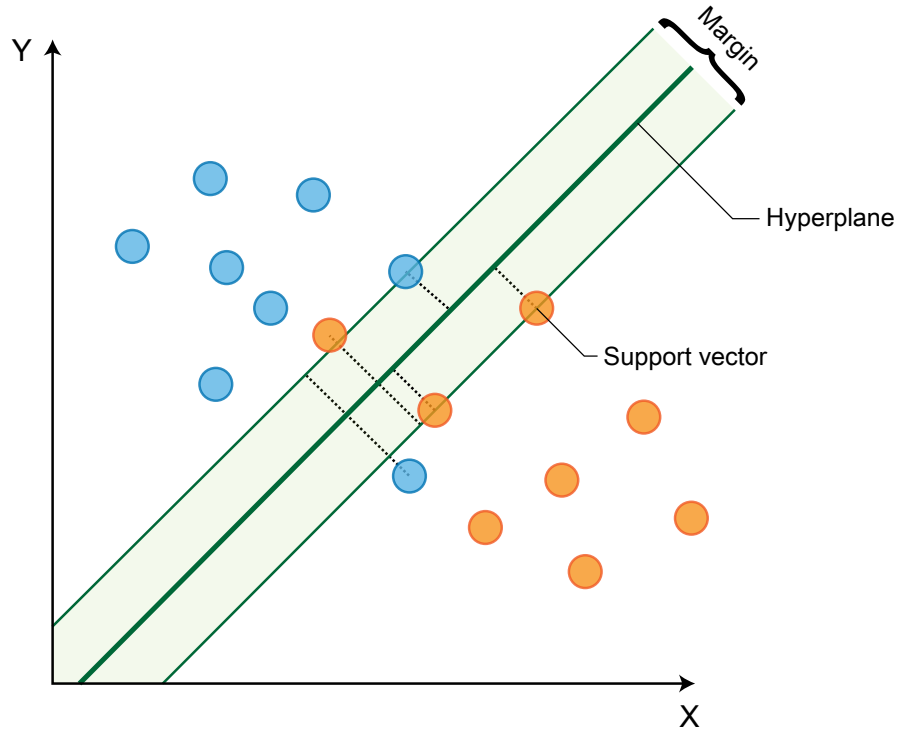


Figure 22: Example diagram of a Support Vector Machine. The circles indicate different samples, where the colours (blue and orange) represent different classes. The green area is the margin, with the hyperplane running through its centre. The support vectors are the samples closest to the hyperplane, and contribute to the placement of the margin, they can be identified by the dotted lines orthogonal to the hyperplane, indicating their distance to the hyperplane. This is a simplified 2-dimensional example with an arbitrary x and y axis.

SVM makes use of the kernel trick, which transforms the data into a high dimensional space in the attempt to make non-linearly separable data, separable. Many different kernels are available, but this thesis only focusses on the Gaussian Radial Basis Function (rbf) kernel. This is a popular kernel as it can create nonparametric decision functions through which almost any decision boundary can be acquired. Although, SVMs are very powerful techniques, normalisation of the data is advised by scaling each variable to a mean of zero and unit variance when using the rbf kernel. If the data is not properly scaled, variables with a higher numerical range can overshadow the effect of variables with smaller ranges[104].

9.2.1.5 Partial Least Squares (PLS)

Partial Least Squares (PLS), sometimes also referred to Projection to Latent Structures, is a method that is similar to PCA, as it projects observed variables into latent variables, which in the case of PCA are linear combinations of the original variables that describe the maximal variance of the original data. However, what makes PLS particularly interesting, is that it can handle two data blocks X and Y simultaneously. Here, it generates latent variables for both data blocks to optimally describe the variance in either data block. Furthermore, the method also maximises the covariance between the latent variables of X and Y. This allows to capture the joint variance between data block X and Y, such that you can best predict Y from X in the latent space[105].

Another method called sparse PLS (sPLS) was developed by Cao et. al, 2008[106], and is a derivative of PLS that performs soft-thresholding penalisation on the loading vectors to combine variable selection and modelling into one step. Since the loading vectors of the data blocks in question represent the relative importance of the variables in either data block, the sPLS method puts sparsity constraints on the loading vectors. This ensures that only variables important for the prediction of Y from X are retained in the sPLS model. Due to its inherent feature selection method, it is argued that sPLS results are more useful

in terms of interpretability, compared to PLS[106]. As sPLS results in a set of latent variables of X that maximally covariate with Y, it can be expected that this would be an appropriate set of independent variables for the prediction of Y. The sparsity constraints that are enforced upon these latent variables can be considered as simplification constraints, in the sense that the latent variables can more easily be compared to the original feature space, as they are formed by a reduced number of original variables. Therefore, the sPLS method holds the potential to deliver a balance between model performance and biological interpretation.

Although (s)PLS is not a classification ML algorithm by nature, it can be combined with Discriminant Analysis (DA) into sPLS-DA. As such, the regression properties of sPLS are translated into classification properties[107]. The sPLS-DA method applies $\ell 1$ penalisation on the loading vectors of the independent variables and has been shown to have good predictive performance, while also allowing for easy interpretation and many useful graphical interpretations of the model[108]. However, it should be mentioned that any PLS related methodology, including sPLS-DA, requires appropriately scaled input data[105].

9.2.1.6 Shapley value feature importance

Supervised ML algorithms are not only useful for the task of prediction itself but can also be used to study in the underlying system. This can be achieved through the interpretability of the models, where feature importance can give insight into the role of specific features of a system, and how these relate to some dependent variable of interest. However, it is often the case that more complex methods, such as RF and XGBoost are not as interpretable as simpler techniques as LR for instance. Nevertheless, techniques have been developed to circumvent this issue through model-agnostic explanation methods, which are methods that can explain any type of ML algorithm[109]. One of such techniques, is the Shapley value.

Within this thesis, the Shapley value method was used to interpret the ML models. This technique is an agnostic local explanation method, meaning that the Shapley value calculates sample-specific feature importance, rather than looking at the system as a whole. However, the Shapley value can also be utilised for global feature importance extraction. The fact that the Shapley value can determine feature importance per sample can be particularly useful for complex prediction tasks[91]. Due to the heterogeneity of the gut microbiome and its potential association with atopy, it could very well be the case that taxa do not simply follow a trend of increased/decreased risk of developing atopy. Instead, various taxa could have different effects on the dependent variable in specific microbiome compositions for example. This is even more likely given the fact that relative abundances are used, which means that all feature values are relative to the other values of a given sample. Therefore, the feature value of one sample cannot be directly compared to the value of the same feature for a different sample. As a result, local explanation methods are highly desirable to gain insight into the studied system.

The Shapley value is a technique based on cooperative game theory concepts, as it determines the contribution of different players toward a specific pay-out. In the context of supervised ML methods, the different features or variables are the players, and the prediction is the pay-out. The major drawback of Shapley values however, is the exponential growth of the computational complexity as more features are used[91]. Therefore, this thesis project used a TreeExplainer method, that is a derivate of the original Shapley value optimised for tree-based ML methods including RF and XGBoost, that is not hindered by larger numbers of features[90].

9.2.1.7 Unsupervised Random Forest (URF)

Comparable to other unsupervised techniques, such as PCA, Unsupervised Random Forest (URF) is a method that can give relevant insights into the structure, and underlying patterns of data. This technique is based on RF, which is described in section 9.2.1.2, but

in an unsupervised fashion. As such, URF does not require a dependent variable that it tries to predict, instead, its aim is to recognise the data-specific structure. Here, URF follows the assumption that if the data has any structure, it should be possible to distinguish the real data from a randomly permuted version of the real data. Therefore, when training an RF to distinguish the real data set from a fake synthetic one, it will inherently learn the general structure of the real data. Since RFs can quantify similarities between samples, URF can be used to produce a distance matrix, which in turn can be used in unsupervised analytical methods such as clustering, Multidimensional Scaling and PCoA. Since the similarity matrix is converted into the Euclidean distance space, making it symmetric positive definite, where each entry lies in the unit interval [0,1], PCoA is identical to cMDS in this case[77].

URF has various benefits compared to PCA. First of all, it can capture both linear and non-linear structure in the data whereas PCA is a purely linear method. Furthermore, it is insensitive to outliers and is unaffected by linear scaling. However, it is important to note that URFs will produce different results when applied to non-linear scaling techniques such as log-transforms. For microbiome data this means that for example, that centre-log-ratio (CLR) transformation can potentially reveal different characteristics of the data. Nonetheless, URF also has drawbacks, as it loses the interpretability that PCA offers with variable loadings because URF results in a similarity matrix, which contains no information on the variable level[110].

9.2.2 Data fusion

Data fusion, also known as data integration, is the procedure of combining two or more separate data blocks. These separate blocks should have some aspect in common in order to be combined, for example if there are two separate data blocks, each with different samples, but identical variables, one set of samples can be appended to the other set of samples. Alternatively, it is also possible that both blocks share identical

samples, yet have different variables, such that the variables can be combined. For this thesis, the latter option is particularly interesting as there is clinical and microbiome data available, which are different feature sets for identical samples. Data fusion can be applied to improve predictive performance, and/or to get a better understanding of the underlying system[111–113].

There are different forms of data fusion, including low-, mid- and high-level data fusion. Low-level data fusion, sometimes also called structure revealing data fusion, combines the data at the raw level, although they can be separately pre-processed beforehand. The resulting fused data block can then be used for further analysis. Mid-level data fusion on the other hand integrates features rather than variables. This means that prior to fusing the two data blocks, the variables are transformed into features. This can be achieved in various ways, for instance through dimensionality reduction, resulting in latent variables as features, or variable selection, where a filtered subset of the variables are the features. The resulting features are then combined into one data block. High-level data fusion can be used for classification purposes, where some classifier is trained on separate data blocks. Subsequently, the predictions on each separate block are combined into one final classification[114].

Another form of data fusion is known as sustainable mid-level data fusion, which does not only generate features for the blocks separately, but also considers inter-block association. In the case of generating features through a dimensionality reduction method, such as PCA, this would entail that the features are not created by only looking at intra-block variance. Instead, covariance between the two blocks is considered when creating the latent variance. This is beneficial as this allows to retain information that is shared between the two data blocks, possibly allowing for the identification of interesting associations between the two blocks. Also, this can potentially improve post hoc classification performance[114]. The data fusion technique that was equipped for this thesis is DIABLO, which is a sustainable mid-level data fusion technique.

9.2.2.1 Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO)

DIABLO[115] is a method from the mixOmics[71] R Bioconductor package. DIABLO aims at integrating two or more different types of data blocks to improve the holistic understanding of the studied system, but also to maximise the predictive performance. DIABLO is an extension of sparse Generalized Canonical Correlation Analysis (sGCCA)[116], which identifies linear combinations of variables, that maximise the co-variance between a multitude of data blocks.

$$\max_{a_h^{(1)},\ldots,a_h^{(Q)} i,\ j=1, i\neq j} \sum^{Q} c_{i,j}\, cov(x_h^{(i)} a_h^{(i)}, x_h^{(j)} a_h^{(j)}),$$
$$s.t. ||a_h^{(q)}||_2 = 1 \ and \ ||a_h^{(q)}||_1 \leq \lambda^{(q)} for\ all\ 1 \leq q \leq Q \qquad \text{Eq. 2}$$

The sGCCA model is depicted in Eq. 2, where Q represents the separate data blocks $X^{(1)}(N \times P_1)$, $X^{(2)}(N \times P_2)$, …, $X^{(Q)}(N \times P_Q)$, which should all be normalised, centred and scaled as is appropriate for their specific data types. The samples N should be identical for all data blocks, whereas, the variable sets P_1, …, P_q should describe varying properties of the studied system. C is a square design matrix of dimension $(Q \times Q)$, specifying which inter-block variation should be maximised. Here, $c_{i,j} = 1$ would specify that the co-variation between data blocks i and j should be maximised, whereas $c_{i,j} = 0$, would specify that this connection should be ignored. The model also incorporates an $\ell 1$ penalty parameter $\lambda^{(q)}$ which is a non-negative parameter which regulates the number of non-zero variable loadings.

As is described in Eq. 2, sGCCA maximises the covariance between all the combinations of data blocks as depicted in C for each dimension $h = 1, \ldots, H$. Each dimension h corresponds to the latent variable level of the output, where the optimisation for $h=1$ maximises the covariance with $X_1^{(q)}$ representing the full data

blocks $X^{(q)}$, resulting in the set of loading vectors $(a_1^{(1)}, \ldots, a_1^{(Q)})$ which contain the loadings for the first latent variable of all data blocks. Subsequently, for dimension $h=2$, the residual matrices $X_2^{(q)} = X_1^{(q)} - t_1^{(q)} a_1^{(q)}$ are used, where $t_1^{(q)}, \ldots, t_h^{(Q)}$ are the component scores of the first latent variable.

The only alteration that DIABLO applies to the sGCCA model described in Eq. 2 is adding one data block $X^{(q)}$ which is a dummy indicator matrix $Y(N \times G)$ containing the G class labels for all N samples. Additionally, the $\ell 1$ penalty parameter $\lambda^{(q)}$ is replaced with the number of selected variables for each data block and dimension[115].

9.2.3 Score metrics

9.2.3.1 Area Under the Receiver Operating Characteristics (AUROC)

To understand the Area Under the Receiver Operating Characteristics (AUROC), it is important to grasp the concept of the Receiver Operating Characteristics (ROC) graphs. An ROC graph shows the true positive rate, on the y-axis, over the false positive rate, on the x-axis. As such, the line that is plotted on the ROC graph, corresponds to the trade-off between how many of the samples of the positive class have been classified correctly (true positive rate), and how many of the negative class are classified incorrectly (false positive rate). Thus, a perfect model is visualised as a line that goes straight into the left upper corner, and only after the true positive rate of 1 is reached, would introduce false positives, and go to the right upper corner. A nearly perfect model is illustrated by the green line in the example ROC curve in Figure 23. The blue line would correspond to a sub-optimal model, that does have predictive capacity but is not perfect. The red line is the worst performing model of the three. An intuitive way to deduce that the model of the red line does not perform well is because it gets very close to the dotted diagonal line. This diagonal indicates the hypothetical performance of a model that would randomly

guess the class label for an infinite number of samples and would therefore get 50% of its classification correct. Therefore, a model with any predictive capacity should be above this diagonal.

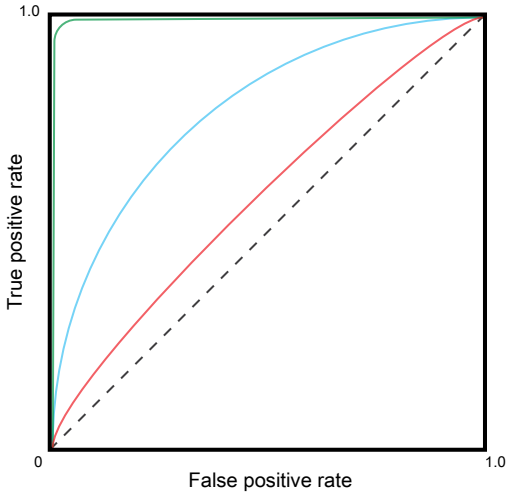


Figure 23: Example Receiver Operating Characteristics (ROC) graph. The green line illustrates a perfect model, blue represents a model with some predictive capacity and red shows a model with low predictive capacity.

ROC curves can be a convenient tool to compare different models in one plot, but they can also be utilised to calculate a single performance metric per model. To this end, the area under the line in the ROC curve can be computed, which corresponds to the AUROC. This metric is often also simply abbreviated as AUC. The AUROC measure is widely used as a robust way of summarising a model's performance. However, it does not tell the complete story as a model with a higher AUROC does not always perform better in all areas of the ROC. Furthermore, an ROC graph shows whether samples are correctly or incorrectly classified given some threshold but does not consider their actual predicted probability. Consequently, the fact that a model has a high AUROC, does not mean that the predicted probabilities reflect reality, only that samples with a higher probability of belonging to the positive class, are relatively more likely to truly belong to

the positive class[117]. This also means that it does not reflect how many samples the model was able to correctly classify, and as such, it is always best used in combination with other evaluation metrics to get a holistic understanding of a model's performance.

9.2.3.2 F-score

The F-score is an accuracy test based on the precision and recall of a model. This means that it describes both how many of the positive cases it can identify, but also how many of the positively classified samples are misclassified. Generally, the precision and recall are of equal weight in this calculation, this is referred to as the F_1-score. However, also the F_β-score can be computed where β specifies the weight given to the recall which is useful when a specific trade-off between precision and recall is desired. However, throughout this thesis only the F_1-score is considered as is indicated in Eq. 3.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \qquad \text{Eq. 3}$$

9.2.3.3 Log loss

Log loss is a so-called proper scoring rule, which means that this metric does not only consider whether a model makes right or wrong predictions, it also considers how far off the model is. For the prediction of categorical variables, this means that the model should be careful with its prediction of the likelihood of some sample belonging to a specific class. For example, a prediction that states that a sample is 80% likely to belong to class 1, while it actually belongs to class 2, will have a higher impact on the loss function compared to if it would have predicted 60% likelihood of belonging to class 1. Conversely, proper scoring rules also prefer higher predicted likelihoods if the predictions are correct. Using a proper scoring rule is often preferred as it will give more meaning to the output of a prediction model, where the predicted likelihood will better reflect the true probability of

a sample belonging to some class, as opposed to improper scoring rules such as accuracy.

$$L_{log}(y,p) = -(y\ln(p) + (1-\mathrm{y})\ln(1-p)) \quad \text{Eq. 4}$$

The formula to compute the log loss of one sample is shown in Eq. 4. Here, the log loss for some sample of true class y and predicted probability p is depicted as $L_{log}(y,p)$. For the sake of interpretability of the visualisations in this thesis, a slightly adjusted version of the log loss score has been introduced. This adjusted metric is referred to as "1 – log loss". To compute this metric, the log loss was simply capped at one, such that any log loss higher than one was set to one, forcing the metric into the [0, 1] interval. Thereafter, the metric was inversed by subtracting it from one. This was done so higher metric values correspond to a 'better' model, just as the other evaluation metrics used in this thesis.

9.3 Descriptive statistics

Table A1: Descriptive statistics table. The first column shows all the variables in bold with their corresponding values (for categorical variables) or mean with standard deviation and range (for numeric variables). The other columns show the results for the different sample subsets for each prediction task.

	Asthma (N=845)	**Asthma (N=499)** *Atopic parents*	**Eczema (N=791)** *6 to 7 years*	**Eczema (N=887)** *First two years*	**Feeding type (N=813)**	**Birth mode (N=892)**
Older siblings						
Has at least one older sibling	347 (41.1%)	207 (41.5%)	327 (41.3%)	364 (41.0%)	324 (39.9%)	367 (41.1%)
Has no older siblings	498 (58.9%)	292 (58.5%)	464 (58.7%)	523 (59.0%)	489 (60.1%)	525 (58.9%)
birth mode and place						
Missing	1	0	1	1	1	/
Vaginal delivery at home	388 (46.0%)	235 (47.1%)	371 (47.0%)	410 (46.3%)	389 (47.9%)	/
Vaginal delivery in the hospital	365 (43.2%)	209 (41.9%)	331 (41.9%)	381 (43.0%)	338 (41.6%)	/
Caesarean section	91 (10.8%)	55 (11.0%)	88 (11.1%)	95 (10.7%)	85 (10.5%)	/
Smoke exposure during pregnancy						

Missing	8	8	7	7	8	8
No smoke exposure	502 (60.0%)	299 (60.9%)	474 (60.5%)	528 (60.0%)	492 (61.1%)	530 (60.0%)
Some smoke exposure	243 (29.0%)	149 (30.3%)	228 (29.1%)	256 (29.1%)	229 (28.4%)	257 (29.1%)
High smoke exposure	92 (11.0%)	43 (8.8%)	82 (10.5%)	96 (10.9%)	84 (10.4%)	97 (11.0%)
Neonatal animal exposure						
No exposure to animals as neonate	500 (59.2%)	301 (60.3%)	469 (59.3%)	522 (58.9%)	478 (58.8%)	525 (58.9%)
Exposure to animals as neonate	345 (40.8%)	198 (39.7%)	322 (40.7%)	365 (41.1%)	335 (41.2%)	367 (41.1%)
Direct antibiotics						
Antibiotics through breastmilk	14 (1.7%)	9 (1.8%)	13 (1.6%)	14 (1.6%)	12 (1.5%)	14 (1.6%)
Antifungals	34 (4.0%)	18 (3.6%)	33 (4.2%)	35 (3.9%)	33 (4.1%)	35 (3.9%)
Direct antibiotics	27 (3.2%)	19 (3.8%)	25 (3.2%)	28 (3.2%)	26 (3.2%)	28 (3.1%)
No antibiotics	770 (91.1%)	453 (90.8%)	720 (91.0%)	810 (91.3%)	742 (91.3%)	815 (91.4%)
Antibiotics exposure during pregnancy						
No antibiotics during pregnancy	720 (85.2%)	410 (82.2%)	677 (85.6%)	753 (84.9%)	693 (85.2%)	756 (84.8%)
Antibiotics during pregnancy	125 (14.8%)	89 (17.8%)	114 (14.4%)	134 (15.1%)	120 (14.8%)	136 (15.2%)
Trimester of antibiotics exposure during pregnancy						
missing	4	2	4	5	4	5
No antibiotics during pregnancy	717 (85.3%)	408 (82.1%)	674 (85.6%)	750 (85.0%)	690 (85.3%)	753 (84.9%)
Antibiotics during first trimester of pregnancy	25 (3.0%)	16 (3.2%)	23 (2.9%)	27 (3.1%)	23 (2.8%)	27 (3.0%)
Antibiotics during second trimester of pregnancy	46 (5.5%)	34 (6.8%)	43 (5.5%)	47 (5.3%)	45 (5.6%)	49 (5.5%)
Antibiotics during third trimester of pregnancy	53 (6.3%)	39 (7.8%)	47 (6.0%)	58 (6.6%)	51 (6.3%)	58 (6.5%)
Breast feeding proportion						
Mean (SD)	0.772 (0.354)	0.779 (0.352)	0.776 (0.351)	0.772 (0.355)	/	0.770 (0.356)
Range	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	0.000 - 1.000	/	0.000 - 1.000
Feeding type						
Missing	2	1	2	2	/	2
Breastfeeding	585 (69.4%)	355 (71.3%)	553 (70.1%)	613 (69.3%)	/	615 (69.1%)
Formula feeding	184 (21.8%)	106 (21.3%)	167 (21.2%)	195 (22.0%)	/	197 (22.1%)
Mixed feeding	74 (8.8%)	37 (7.4%)	69 (8.7%)	77 (8.7%)	/	78 (8.8%)
Age at faecal collection (days)						
Mean (SD)	31.801 (3.693)	31.858 (3.825)	31.779 (3.680)	31.899 (3.780)	31.889 (3.783)	31.896 (3.773)

Range	22.000 - 49.000	23.000 - 49.000	22.000 - 49.000	22.000 - 49.000	22.000 - 49.000	22.000 - 49.000
pregnancy duration (weeks)						
Missing	3	2	2	3	3	3
Mean (SD)	39.633 (1.286)	39.618 (1.327)	39.643 (1.276)	39.633 (1.283)	39.617 (1.270)	39.632 (1.280)
Range	33.000 - 43.000	33.000 - 42.000	33.000 - 43.000	33.000 - 43.000	33.000 - 43.000	33.000 - 43.000
Frequency of probiotics use during pregnancy						
Missing	4	3	4	4	4	4
Rarely or never	666 (79.2%)	391 (78.8%)	617 (78.4%)	707 (80.1%)	648 (80.1%)	710 (80.0%)
Multiple times per month	80 (9.5%)	49 (9.9%)	77 (9.8%)	79 (8.9%)	74 (9.1%)	80 (9.0%)
Multiple times per week	51 (6.1%)	28 (5.6%)	50 (6.4%)	52 (5.9%)	46 (5.7%)	52 (5.9%)
Daily	19 (2.3%)	13 (2.6%)	19 (2.4%)	19 (2.2%)	18 (2.2%)	19 (2.1%)
Unsure	25 (3.0%)	15 (3.0%)	24 (3.0%)	26 (2.9%)	23 (2.8%)	27 (3.0%)
Maternal asthma						
Missing	1	1	1	1	1	1
No maternal asthma	768 (91.0%)	422 (84.7%)	716 (90.6%)	800 (90.3%)	731 (90.0%)	805 (90.3%)
Maternal asthma	76 (9.0%)	76 (15.3%)	74 (9.4%)	86 (9.7%)	81 (10.0%)	86 (9.7%)
Maternal dust mite allergy						
No maternal dust mite allergy	698 (82.6%)	352 (70.5%)	650 (82.2%)	730 (82.3%)	668 (82.2%)	734 (82.3%)
Maternal dust mite allergy	147 (17.4%)	147 (29.5%)	141 (17.8%)	157 (17.7%)	145 (17.8%)	158 (17.7%)
Maternal furry pet allergy						
Missing	1	1	1	1	1	1
No maternal furry pet allergy	698 (82.7%)	352 (70.7%)	651 (82.4%)	729 (82.3%)	667 (82.1%)	733 (82.3%)
Maternal pet allergy	146 (17.3%)	146 (29.3%)	139 (17.6%)	157 (17.7%)	145 (17.9%)	158 (17.7%)
Maternal hay fever						
No maternal hay fever	641 (75.9%)	295 (59.1%)	602 (76.1%)	673 (75.9%)	615 (75.6%)	677 (75.9%)
Maternal hay fever	204 (24.1%)	204 (40.9%)	189 (23.9%)	214 (24.1%)	198 (24.4%)	215 (24.1%)
Paternal asthma						
Missing	16	11	16	16	15	16
No paternal asthma	736 (88.8%)	395 (80.9%)	690 (89.0%)	776 (89.1%)	713 (89.3%)	780 (89.0%)
Paternal asthma	93 (11.2%)	93 (19.1%)	85 (11.0%)	95 (10.9%)	85 (10.7%)	96 (11.0%)
Paternal dust mite allergy						
Missing	19	15	18	19	18	19
No paternal dust mite allergy	694 (84.0%)	352 (72.7%)	651 (84.2%)	728 (83.9%)	664 (83.5%)	732 (83.8%)
Paternal dust mite allergy	132 (16.0%)	132 (27.3%)	122 (15.8%)	140 (16.1%)	131 (16.5%)	141 (16.2%)
Paternal furry pet allergy						

Missing	14	10	14	15	14	15
No paternal furry pet allergy	691 (83.2%)	349 (71.4%)	647 (83.3%)	726 (83.3%)	663 (83.0%)	730 (83.2%)
Paternal furry pet allergy	140 (16.8%)	140 (28.6%)	130 (16.7%)	146 (16.7%)	136 (17.0%)	147 (16.8%)
Paternal hay fever						
Missing	22	15	21	23	20	23
No Paternal hay fever	591 (71.8%)	252 (52.1%)	554 (71.9%)	628 (72.7%)	574 (72.4%)	630 (72.5%)
Paternal hay fever	232 (28.2%)	232 (47.9%)	216 (28.1%)	236 (27.3%)	219 (27.6%)	239 (27.5%)
Pet dogs during pregnancy						
No pet dogs during pregnancy	706 (83.6%)	412 (82.6%)	664 (83.9%)	742 (83.7%)	678 (83.4%)	745 (83.5%)
Pet dogs during pregnancy	139 (16.4%)	87 (17.4%)	127 (16.1%)	145 (16.3%)	135 (16.6%)	147 (16.5%)
Pet cats during pregnancy						
No pet cats during pregnancy	640 (75.7%)	391 (78.4%)	596 (75.3%)	666 (75.1%)	616 (75.8%)	671 (75.2%)
Pet cats during pregnancy	205 (24.3%)	108 (21.6%)	195 (24.7%)	221 (24.9%)	197 (24.2%)	221 (24.8%)
Pet rodents during pregnancy						
No pet rodents during pregnancy	791 (93.6%)	468 (93.8%)	741 (93.7%)	825 (93.0%)	754 (92.7%)	830 (93.0%)
Pet rodents during pregnancy	54 (6.4%)	31 (6.2%)	50 (6.3%)	62 (7.0%)	59 (7.3%)	62 (7.0%)
Birth weight (grams)						
Missing	75	43	71	79	70	79
Mean (SD)	3564.682 (473.003)	3557.476 (473.023)	3567.960 (469.412)	3565.496 (471.557)	3553.945 (470.922)	3564.048 (471.776)
Range	1755.000 - 4960.000	2140.000 - 4900.000	1755.000 - 4960.000	1755.000 - 4960.000	1755.000 - 4960.000	1755.000 - 4960.000
Baby washed when held						
Missing	13	9	10	11	11	12
Not washed	786 (94.5%)	460 (93.9%)	737 (94.4%)	829 (94.6%)	762 (95.0%)	833 (94.7%)
Washed off blood	20 (2.4%)	12 (2.4%)	18 (2.3%)	21 (2.4%)	18 (2.2%)	21 (2.4%)
Washed completely	26 (3.1%)	18 (3.7%)	26 (3.3%)	26 (3.0%)	22 (2.7%)	26 (3.0%)
Hospitalisation						
Missing	12	8	9	10	10	11
Baby was not hospitalised	581 (69.7%)	340 (69.2%)	549 (70.2%)	610 (69.6%)	570 (71.0%)	614 (69.7%)
Baby was hospitalised	252 (30.3%)	151 (30.8%)	233 (29.8%)	267 (30.4%)	233 (29.0%)	267 (30.3%)
Hospitalisation from birth						
Missing	12	8	9	10	10	11
Baby was not immediately hospitalised after birth	603 (72.4%)	353 (71.9%)	571 (73.0%)	635 (72.4%)	592 (73.7%)	639 (72.5%)

Baby was immediately hospitalised after birth	230 (27.6%)	138 (28.1%)	211 (27.0%)	242 (27.6%)	211 (26.3%)	242 (27.5%)
Incubator						
Missing	83	50	75	85	74	86
Baby did not go into the incubator	731 (95.9%)	434 (96.7%)	689 (96.2%)	768 (95.8%)	708 (95.8%)	772 (95.8%)
Baby was put in an incubator	31 (4.1%)	15 (3.3%)	27 (3.8%)	34 (4.2%)	31 (4.2%)	34 (4.2%)
Lives on a farm						
Missing	14	9	13	10	12	14
Does not live on a farm	807 (97.1%)	476 (97.1%)	756 (97.2%)	852 (97.1%)	778 (97.1%)	853 (97.2%)
Lives on a farm	24 (2.9%)	14 (2.9%)	22 (2.8%)	25 (2.9%)	23 (2.9%)	25 (2.8%)
Atopic parents						
No atopic parents	346 (40.9%)	/	321 (40.6%)	364 (41.0%)	326 (40.1%)	365 (40.9%)
One atopic parent	377 (44.6%)	377 (75.6%)	356 (45.0%)	397 (44.8%)	372 (45.8%)	401 (45.0%)
Two atopic parents	122 (14.4%)	122 (24.4%)	114 (14.4%)	126 (14.2%)	115 (14.1%)	126 (14.1%)
Furry pets during pregnancy						
No furry pets during pregnancy	499 (59.1%)	300 (60.1%)	467 (59.0%)	519 (58.5%)	477 (58.7%)	522 (58.5%)
Furry pets during pregnancy	346 (40.9%)	199 (39.9%)	324 (41.0%)	368 (41.5%)	336 (41.3%)	370 (41.5%)
Asthma (dependent variable)						
asthma	94 (11.1%)	68 (13.6%)	/	/	/	/
no asthma	751 (88.9%)	431 (86.4%)	/	/	/	/
Eczema (dependent variable)						
eczema	/	/	158 (20.0%)	297 (33.5%)	/	/
no eczema	/	/	633 (80.0%)	590 (66.5%)	/	/
Feeding type (dependent variable)						
formula milk	/	/	/	/	198 (24.4%)	/
breast milk	/	/	/	/	615 (75.6%)	/
Birth mode (dependent variable)						
Caesarean section	/	/	/	/	/	95 (10.7%)
Vaginal delivery	/	/	/	/	/	797 (89.3%)

9.4 Results

9.4.1 Explorative data analysis

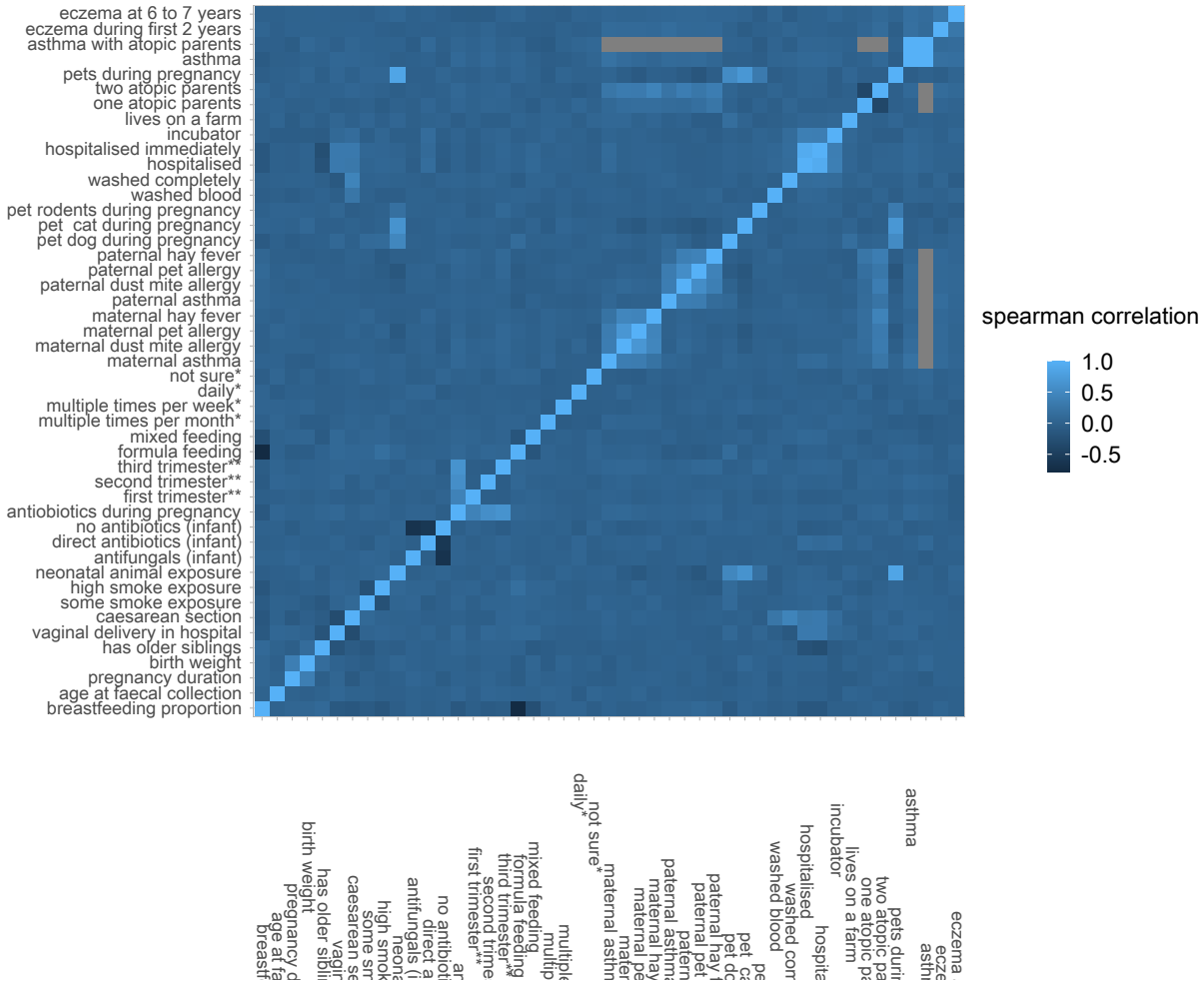


Figure A1: Correlation heatmap of spearman correlation on all clinical dependent and independent variables. All samples passing the exclusion criteria were used (N=894), missing values were dropped per comparison. * Refers to the “Frequency of probiotics use during pregnancy” variable. ** Refers to the “Trimester of antibiotics exposure during pregnancy” variable.

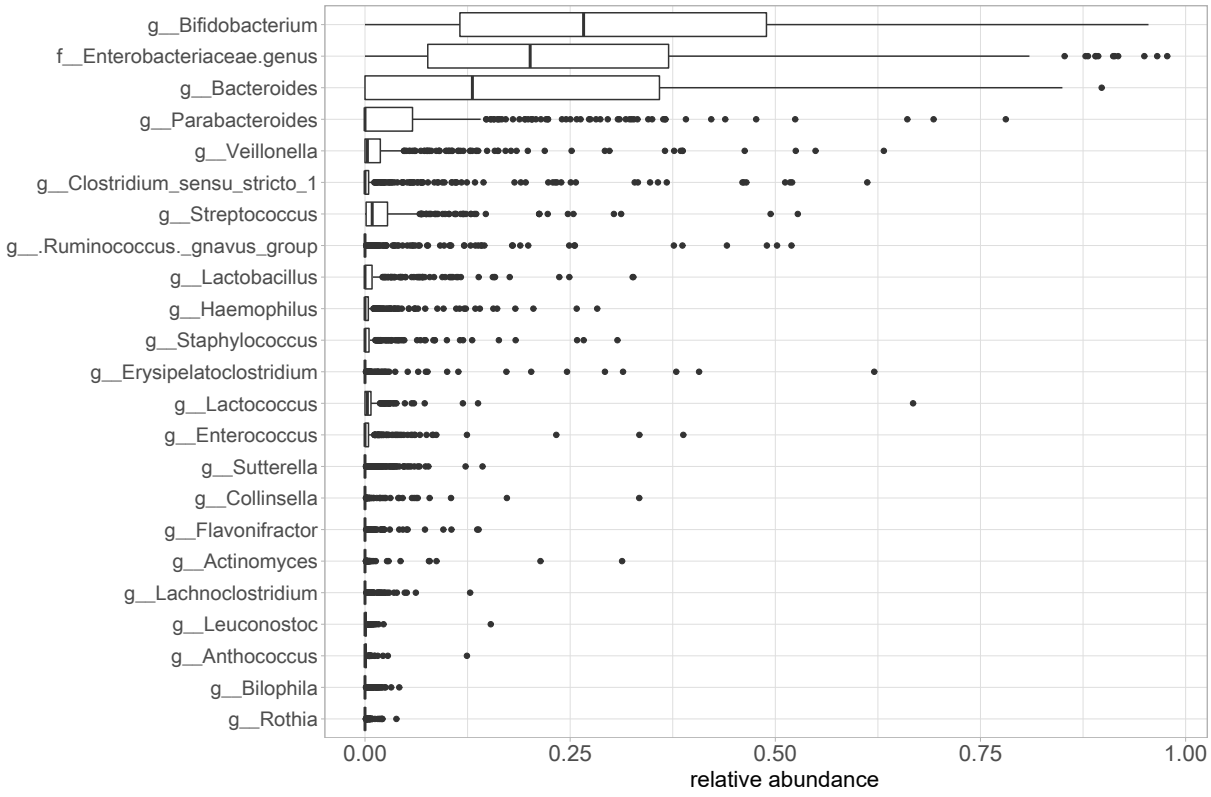


Figure A2: Boxplot of the genera in the microbiome genus feature set for the prediction task of asthma prior to any splits. These are the genera that remain after performing the 0.1 prevalence filter, which retains only genera that are present in at least 10% of all samples. The genera are shown on the y-axis, x-axis shows the relative abundance. The “g__” prefix specifies that the name corresponds to the name of the genus. The “f__” indicates that the name corresponds the family of the genus, where the specific genus is unknown.

9.4.2 Hyperparameter optimisation

Table A2: The hyperparameter values that were evaluated for all the machine learning models.

Model	Hyperparameter	Values											
LR	penalty	0	0.001	0.01	0.1	1	5	10	2	20	50	80	100
	mixture	0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1			
RF	mtry	1	10	20	50	100	150	180	200				
	trees	10	100	300	1000	2000							
	min_n	0	1	10	20								
XGBoost	trees	10	100	1000	2000								
	mtry	1	10	20									
	min_n	0	1	2	5	10	20	40					
	tree_depth	1	2	5	10	20							
	learn_rate	1e-8	1e-4	0.01	0.1	0.5	1						
	loss_reduction	1e-10	1e-5	0.1	10								
SVM (rbf kernel)	cost	1e-4	1e-3	0.01	0.1	1	10	100					
	Rbf_sigma	1e-4	1e-3	0.01	0.1	1	10	100					
sPLS-DA	n_comp	1	2	3	4	5	6	7	8	9	10		
	keepX	1	2	3	4	5	6	7	8	9	10	20	30
		40	50	60	70	80	90	100					

Table A3: Overview of all supervised machine learning prediction tasks results. The first column depicts which model was trained, here XGB refers to XGBoost and rbf to the radial basis function kernel. The second for which dependent variable. The transformation column shows how the microbiome data was transformed, either "compositional" which means relative abundances were used, or "clr" which means that a centre-log-ratio transformation was applied. The "repeats" column shows how many repeats were applied for tuning with repeated 5-fold cross validation. The "Acc." column represents the accuracy, "B. Acc." balanced accuracy, "AUC" Area Under the Receiver Operating Characteristic measures. The last column shows how long the hyperparameter optimisation took in minutes.

model	Dependent variable	Feature set	transformation	repeats	Acc.	B. acc.	AUC	F1	precision	sensitivity	specificity	Mean log loss	tuning time (min)
RF	asthma	clinical	compositional	10	0.875	0.490	0.530	0.000	0.000	0.000	0.980	0.388	5.8
XGB	asthma	clinical	compositional	1	0.863	0.508	0.534	0.080	0.143	0.056	0.960	0.400	32.2
LR	asthma	clinical	compositional	10	0.625	0.521	0.513	0.182	0.119	0.389	0.653	0.656	1.6
SVM (rbf)	asthma	clinical	compositional	10	0.619	0.493	0.539	0.158	0.103	0.333	0.653	0.398	4.2
sPLS-DA	asthma	clinical	compositional	10	0.643	0.556	0.526	0.211	0.138	0.444	0.667	0.669	4.2
RF	asthma	genus	compositional	10	0.875	0.490	0.499	0.000	0.000	0.000	0.980	0.396	8.7
XGB	asthma	genus	compositional	1	0.839	0.494	0.525	0.069	0.091	0.056	0.933	0.426	33.9
LR	asthma	genus	compositional	10	0.583	0.498	0.473	0.167	0.106	0.389	0.607	0.690	3.2
SVM (rbf)	asthma	genus	compositional	10	0.839	0.470	0.465	0.000	0.000	0.000	0.940	0.586	5.9
sPLS-DA	asthma	genus	compositional	10	0.631	0.549	0.502	0.205	0.133	0.444	0.653	0.701	4.9
RF	asthma	expanded	compositional	10	0.893	0.500	0.519	NA	NA	0.000	1.000	0.387	103.5
XGB	asthma	expanded	compositional	1	0.893	0.500	0.517	NA	NA	0.000	1.000	0.361	207.8
LR	asthma	expanded	compositional	10	0.869	0.511	0.458	0.083	0.167	0.056	0.967	3.729	69.6
SVM (rbf)	asthma	expanded	compositional	10	0.893	0.500	0.508	NA	NA	0.000	1.000	0.402	590.4
sPLS-DA	asthma	expanded	compositional	10	0.863	0.508	0.459	0.080	0.143	0.056	0.960	3.750	11.6
RF	Eczema (at 6 to 7 years)	clinical	compositional	10	0.796	0.508	0.590	0.059	0.333	0.032	0.984	0.511	5.8
XGB	Eczema (at 6 to 7 years)	clinical	compositional	1	0.777	0.521	0.580	0.146	0.300	0.097	0.944	0.543	30.9
LR	Eczema (at 6 to 7 years)	clinical	compositional	10	0.618	0.543	0.559	0.302	0.236	0.419	0.667	0.672	1.6
SVM (rbf)	Eczema (at 6 to 7 years)	clinical	compositional	10	0.803	0.500	0.508	NA	NA	0.000	1.000	0.499	3.9
sPLS-DA	Eczema	clinical	compositional	10	0.637	0.555	0.565	0.313	0.250	0.419	0.690	0.654	5.2

	(at 6 to 7 years)												
RF	Eczema (at 6 to 7 years)	genus	compositional	10	0.707	0.440	0.382	0.000	0.000	0.000	0.881	0.604	9.0
XGB	Eczema (at 6 to 7 years)	genus	compositional	1	0.745	0.464	0.408	0.000	0.000	0.000	0.929	0.593	33.4
LR	Eczema (at 6 to 7 years)	genus	compositional	10	0.446	0.399	0.396	0.187	0.132	0.323	0.476	0.704	3.5
SVM (rbf)	Eczema (at 6 to 7 years)	genus	compositional	10	0.803	0.500	0.498	NA	NA	0.000	1.000	0.552	6.1
sPLS-DA	Eczema (at 6 to 7 years)	genus	compositional	10	0.439	0.420	0.361	0.214	0.148	0.387	0.452	0.720	4.1
RF	Eczema (at 6 to 7 years)	expanded	compositional	10	0.777	0.484	0.386	0.000	0.000	0.000	0.968	0.574	22.2
XGB	Eczema (at 6 to 7 years)	expanded	compositional	1	0.771	0.480	0.403	0.000	0.000	0.000	0.960	0.571	187.5
LR	Eczema (at 6 to 7 years)	expanded	compositional	10	0.484	0.423	0.420	0.198	0.143	0.323	0.524	0.701	23.7
SVM (rbf)	Eczema (at 6 to 7 years)	expanded	compositional	10	0.803	0.500	0.500	NA	NA	0.000	1.000	0.499	301.6
sPLS-DA	Eczema (at 6 to 7 years)	expanded	compositional	10	0.255	0.426	0.398	0.273	0.169	0.710	0.143	0.949	10.6
RF	feeding type	clinical	compositional	10	0.727	0.567	0.604	0.313	0.400	0.256	0.877	0.579	5.7
XGB	feeding type	clinical	compositional	1	0.658	0.504	0.527	0.225	0.250	0.205	0.803	0.673	30.0
LR	feeding type	clinical	compositional	10	0.652	0.596	0.640	0.404	0.345	0.487	0.705	0.646	1.5
SVM (rbf)	feeding type	clinical	compositional	10	0.677	0.595	0.642	0.395	0.362	0.436	0.754	0.628	3.9
sPLS-DA	feeding type	clinical	compositional	10	0.553	0.539	0.605	0.357	0.274	0.513	0.566	0.888	3.7
RF	feeding type	genus	compositional	10	0.932	0.876	0.971	0.845	0.938	0.769	0.984	0.241	8.7
XGB	feeding type	genus	compositional	1	0.913	0.838	0.974	0.794	0.931	0.692	0.984	0.195	32.3
LR	feeding type	genus	compositional	10	0.851	0.858	0.908	0.739	0.642	0.872	0.844	0.369	3.4
SVM (rbf)	feeding type	genus	compositional	10	0.876	0.874	0.922	0.773	0.694	0.872	0.877	0.315	6.4
sPLS-DA	feeding type	genus	compositional	10	0.826	0.824	0.880	0.696	0.604	0.821	0.828	0.421	3.5

RF	feeding type	expanded	compositional	10	0.957	0.910	0.981	0.901	1.000	0.821	1.000	0.244	13.0
XGB	feeding type	expanded	compositional	1	0.957	0.910	0.991	0.901	1.000	0.821	1.000	0.140	162.3
LR	feeding type	expanded	compositional	10	0.770	0.534	0.520	0.140	0.750	0.077	0.992	7.658	8.6
SVM (rbf)	feeding type	expanded	compositional	10	0.770	0.534	0.571	0.140	0.750	0.077	0.992	1.047	39.6
sPLS-DA	feeding type	expanded	compositional	10	0.304	0.524	0.554	0.398	0.252	0.949	0.098	24.665	6.4
RF	birth mode	clinical	compositional	10	0.932	0.765	0.834	0.625	0.714	0.556	0.975	0.264	5.6
XGB	birth mode	clinical	compositional	1	0.910	0.629	0.791	0.385	0.625	0.278	0.981	0.264	31.2
LR	birth mode	clinical	compositional	10	0.802	0.742	0.814	0.407	0.293	0.667	0.818	0.454	1.6
SVM (rbf)	birth mode	clinical	compositional	10	0.797	0.665	0.838	0.333	0.250	0.500	0.830	0.373	4.2
sPLS-DA	birth mode	clinical	compositional	10	0.768	0.822	0.838	0.438	0.291	0.889	0.755	0.713	4.9
RF	birth mode	genus	compositional	10	0.876	0.537	0.729	0.154	0.250	0.111	0.962	0.324	8.5
XGB	birth mode	genus	compositional	1	0.887	0.568	0.775	0.231	0.375	0.167	0.969	0.300	33.3
LR	birth mode	genus	compositional	10	0.655	0.611	0.708	0.247	0.159	0.556	0.667	0.604	3.2
SVM (rbf)	birth mode	genus	compositional	10	0.780	0.606	0.653	0.264	0.200	0.389	0.824	0.468	5.9
sPLS-DA	birth mode	genus	compositional	10	0.644	0.679	0.724	0.292	0.183	0.722	0.635	0.621	4.5
RF	birth mode	expanded	compositional	10	0.887	0.543	0.754	0.167	0.333	0.111	0.975	0.307	21.6
XGB	birth mode	expanded	compositional	1	0.898	0.525	0.766	0.100	0.500	0.056	0.994	0.286	208.2
LR	birth mode	expanded	compositional	10	0.164	0.510	0.517	0.187	0.104	0.944	0.075	29.575	29.3
SVM (rbf)	birth mode	expanded	compositional	10	0.887	0.494	0.437	0.000	0.000	0.000	0.987	0.571	58.1
sPLS-DA	birth mode	expanded	compositional	10	0.638	0.626	0.708	0.256	0.162	0.611	0.642	0.618	8.3
RF	Eczema (during first two years)	clinical	compositional	10	0.614	0.529	0.573	0.320	0.390	0.271	0.786	0.659	5.8
XGB	Eczema (during first two years)	clinical	compositional	1	0.580	0.478	0.527	0.213	0.286	0.169	0.786	0.661	31.0
LR	Eczema (during first two years)	clinical	compositional	10	0.585	0.558	0.578	0.434	0.400	0.475	0.641	0.687	1.6
SVM (rbf)	Eczema (during first two years)	clinical	compositional	10	0.665	0.500	0.562	NA	NA	0.000	1.000	0.693	4.1
sPLS-DA	Eczema (during first two years)	clinical	compositional	10	0.608	0.575	0.578	0.448	0.424	0.475	0.675	0.688	4.2
RF	Eczema	genus	compositional	10	0.636	0.542	0.543	0.319	0.429	0.254	0.829	0.660	8.1

	(during first two years)												
XGB	Eczema (during first two years)	genus	compositional	1	0.636	0.533	0.498	0.289	0.419	0.220	0.846	0.667	32.7
LR	Eczema (during first two years)	genus	compositional	10	0.665	0.500	0.500	NA	NA	0.000	1.000	0.693	2.8
SVM (rbf)	Eczema (during first two years)	genus	compositional	10	0.665	0.500	0.496	NA	NA	0.000	1.000	0.693	5.0
sPLS-DA	Eczema (during first two years)	genus	compositional	10	0.449	0.455	0.489	0.366	0.298	0.475	0.436	0.700	4.2
RF	Eczema (during first two years)	expanded	compositional	10	0.625	0.516	0.492	0.250	0.379	0.186	0.846	0.670	14.7
XGB	Eczema (during first two years)	expanded	compositional	1	0.636	0.491	0.483	0.086	0.273	0.051	0.932	0.665	162.3
LR	Eczema (during first two years)	expanded	compositional	10	0.642	0.491	0.491	0.060	0.250	0.034	0.949	11.126	13.2
SVM (rbf)	Eczema (during first two years)	expanded	compositional	10	0.665	0.500	0.500	NA	NA	0.000	1.000	0.693	31.6
sPLS-DA	Eczema (during first two years)	expanded	compositional	10	0.347	0.446	0.444	0.433	0.306	0.746	0.145	0.711	9.5
RF	asthma	genus	clr	10	0.881	0.493	0.514	0.000	0.000	0.000	0.987	0.402	5.7
XGB	asthma	genus	clr	1	0.845	0.498	0.523	0.071	0.100	0.056	0.940	0.434	29.0
LR	asthma	genus	clr	10	0.589	0.501	0.451	0.169	0.108	0.389	0.613	0.695	1.8
SVM (rbf)	asthma	genus	clr	10	0.851	0.501	0.408	0.074	0.111	0.056	0.947	0.570	3.7
sPLS-DA	asthma	genus	clr	10	0.595	0.529	0.482	0.190	0.121	0.444	0.613	0.710	4.7
RF	asthma	expanded	clr	10	0.893	0.500	0.530	NA	NA	0.000	1.000	0.382	22.3
XGB	asthma	expanded	clr	1	0.893	0.500	0.530	NA	NA	0.000	1.000	0.361	208.5
LR	asthma	expanded	clr	10	0.881	0.518	0.457	0.091	0.250	0.056	0.980	3.726	29.2
SVM (rbf)	asthma	expanded	clr	10	0.893	0.500	0.572	NA	NA	0.000	1.000	0.420	54.4
sPLS-DA	asthma	expanded	clr	10	0.863	0.508	0.459	0.080	0.143	0.056	0.960	3.744	8.9
RF	feeding type	genus	clr	10	0.901	0.821	0.964	0.765	0.897	0.667	0.975	0.248	5.2
XGB	feeding type	genus	clr	1	0.907	0.825	0.979	0.776	0.929	0.667	0.984	0.206	25.7

LR	feeding type	genus	clr	10	0.832	0.828	0.901	0.703	0.615	0.821	0.836	0.391	1.9
SVM (rbf)	feeding type	genus	clr	10	0.882	0.861	0.922	0.771	0.727	0.821	0.902	0.311	3.6
sPLS-DA	feeding type	genus	clr	10	0.783	0.787	0.868	0.639	0.534	0.795	0.779	0.442	3.6
RF	feeding type	expanded	clr	10	0.950	0.897	0.985	0.886	1.000	0.795	1.000	0.243	22.8
XGB	feeding type	expanded	clr	1	0.938	0.881	0.992	0.857	0.968	0.769	0.992	0.141	157.7
LR	feeding type	expanded	clr	10	0.770	0.534	0.520	0.140	0.750	0.077	0.992	7.661	21.0
SVM (rbf)	feeding type	expanded	clr	10	0.764	0.522	0.548	0.095	0.667	0.051	0.992	0.894	49.2
sPLS-DA	feeding type	expanded	clr	10	0.298	0.511	0.554	0.389	0.247	0.923	0.098	24.667	6.3
RF	Asthma (infants with atopic parents)	clinical	compositional	10	0.879	0.571	0.772	0.250	0.667	0.154	0.988	0.340	5.1
XGB	Asthma (infants with atopic parents)	clinical	compositional	1	0.859	0.559	0.786	0.222	0.400	0.154	0.965	0.315	23.9
LR	Asthma (infants with atopic parents)	clinical	compositional	10	0.657	0.606	0.630	0.292	0.200	0.538	0.674	0.607	1.5
SVM (rbf)	Asthma (infants with atopic parents)	clinical	compositional	10	0.869	0.500	0.594	NA	NA	0.000	1.000	0.390	3.7
sPLS-DA	Asthma (infants with atopic parents)	clinical	compositional	10	0.677	0.618	0.633	0.304	0.212	0.538	0.698	2.012	2.8
RF	Asthma (infants with atopic parents)	genus	compositional	10	0.828	0.477	0.618	0.000	0.000	0.000	0.953	0.409	7.3
XGB	Asthma (infants with atopic parents)	genus	compositional	1	0.828	0.477	0.693	0.000	0.000	0.000	0.953	0.391	26.1
LR	Asthma (infants with atopic parents)	genus	compositional	10	0.545	0.445	0.466	0.151	0.100	0.308	0.581	0.688	3.0
SVM (rbf)	Asthma (infants with atopic parents)	genus	compositional	10	0.869	0.500	0.552	NA	NA	0.000	1.000	0.389	5.1
sPLS-DA	Asthma (infants with atopic parents)	genus	compositional	10	0.566	0.456	0.473	0.157	0.105	0.308	0.605	0.714	2.4
RF	Asthma (infants with atopic parents)	expanded	compositional	10	0.869	0.500	0.614	NA	NA	0.000	1.000	0.400	13.3
XGB	Asthma	expanded	compositional	1	0.869	0.500	0.669	NA	NA	0.000	1.000	0.374	133.3

	(infants with atopic parents)												
LR	Asthma	expanded	compositional	10	0.778	0.448	0.453	0.000	0.000	0.000	0.895	4.148	14.4
	(infants with atopic parents)												
SVM (rbf)	Asthma	expanded	compositional	10	0.869	0.500	0.500	NA	NA	0.000	1.000	0.395	36.4
	(infants with atopic parents)												
sPLS-DA	Asthma	expanded	compositional	10	0.596	0.506	0.448	0.200	0.135	0.385	0.628	0.701	7.3
	(infants with atopic parents)												