Comparative Evaluation of Machine Learning Models for BMI Prediction from Gut Microbiome Data

University of Turku
Department of Computing
Master of Science Thesis
Data Analytics
April 2026
Vilho Helenius

Supervisors:
Leo Lahti, DSc
Ville Laitinen, PhD

UNIVERSITY OF TURKU
Department of Computing

Vilho Helenius: Comparative Evaluation of Machine Learning Models for BMI Prediction from Gut Microbiome Data

Master of Science Thesis, 52 p., 1 app. p.
Data Analytics
April 2026

This thesis investigates the feasibility of predicting body mass index (BMI) from gut microbiome composition using modern machine learning approaches. The human gut microbiome has been increasingly linked to metabolic health, but the extent to which microbial community profiles can predict host phenotypes such as BMI remains an open question.

Using a large-scale dataset of 9,709 gut microbiome samples, this study compares the predictive performance of several machine learning models representing different modeling paradigms for tabular data. The evaluated models include Lasso regression as a linear baseline, gradient-boosted decision tree methods (XGBoost and CatBoost), and a recently proposed transformer-based foundation model for tabular data, TabPFN. In addition to model comparison, the study examines how dataset size and feature set composition influence predictive performance.

The results indicate that BMI prediction from gut microbiome data remains challenging, with overall predictive performance being moderate across all models. Among the evaluated methods, TabPFN achieved the highest predictive accuracy on the full dataset, suggesting that transformer-based foundation models may offer advantages in large-scale microbiome prediction tasks. However, tree-based ensemble models performed competitively and exhibited stronger performance in small-sample regimes. Additional experiments showed that microbiome features provide substantially more predictive signal for BMI than basic demographic variables alone, while the combination of microbiome and host metadata produced the best overall performance.

Overall, the findings highlight both the potential and the limitations of current machine learning approaches for microbiome-based phenotype prediction. While meaningful predictive signal can be extracted from microbiome composition, the results suggest that microbiome data alone is insufficient for accurate prediction of complex metabolic traits such as BMI.

Keywords: Gut microbiome, machine learning, BMI prediction, tabular data, transformer models, TabPFN

Contents

1 Introduction

The human gut microbiome has emerged as a central factor in understanding health and disease, influencing a wide range of physiological processes including metabolism, immune function, and energy homeostasis. Advances in high-throughput sequencing technologies and large-scale microbiome initiatives have led to the rapid accumulation of complex, high-dimensional microbiome datasets. At the same time, machine learning methods have become increasingly prominent in biomedical research as tools for extracting predictive and mechanistic insights from such data. This thesis investigates how modern machine learning models, including recent transformer-based approaches, perform in predicting body mass index (BMI) from gut microbiome profiles.

Obesity and overweight are among the most significant public health challenges worldwide, contributing to increased risk of cardiovascular disease, type 2 diabetes, and numerous other chronic conditions. Body mass index (BMI) is a widely used, albeit imperfect, proxy for adiposity and metabolic health at the population level. While BMI is influenced by genetic, environmental, and lifestyle factors, accumulating evidence suggests that the gut microbiome plays a measurable role in host energy balance and metabolic regulation. Understanding how microbial community composition relates to BMI may therefore provide insights into the biological mechanisms underlying obesity and open avenues for personalized interventions.

From a data-analytic perspective, microbiome datasets pose several unique chal-

lenges. They are typically high-dimensional, sparse, compositional, and subject to substantial inter-individual variability. Classical statistical approaches often struggle in such settings due to the high dimensionality and compositional constraints of the data, motivating the use of more flexible machine learning models capable of capturing non-linear relationships and complex feature interactions. In recent years, tree-based ensemble methods such as XGBoost and CatBoost have become a widely used standard for predictive modeling on structured tabular data, including biomedical datasets. These models combine strong predictive performance with a degree of interpretability and robustness, making them attractive for applied research.

More recently, transformer-based foundation models for tabular data have been proposed as a potential paradigm shift in machine learning for structured datasets. TabPFN represents a recent development, as it reframes supervised learning as a form of in-context inference using a model pretrained on a large distribution of synthetic datasets. Unlike conventional models that are trained separately for each task, TabPFN leverages prior knowledge learned during pretraining to perform rapid inference on new datasets with minimal tuning. This raises the question of whether such foundation models can outperform or complement established tree-based methods in real-world biomedical applications, especially in data regimes that are common in microbiome research.

The motivation for this thesis is therefore twofold. First, from a biomedical perspective, it aims to contribute to the growing body of work investigating the relationship between the gut microbiome and host metabolic traits by assessing how well BMI can be predicted from microbial composition alone. Second, from a methodological perspective, it seeks to provide an empirical evaluation of modern machine learning models for microbiome-based prediction tasks, with a particular focus on comparing traditional gradient boosting methods with emerging transformer-based approaches. By combining large-scale microbiome data with state-of-the-art ma-

chine learning techniques, this work aims to evaluate both the opportunities and limitations of current predictive modeling approaches in microbiome research.

Selected research questions this thesis explores are:

1. Can transformer-based models (TabPFN) outperform traditional tree-based methods (XGBoost, CatBoost) in predicting BMI from gut microbiome data?

2. How does dataset size affect the relative performance of TabPFN compared to tree-based models in gut microbiome prediction tasks?

These research questions were selected to address both applied and methodological aspects of microbiome-based machine learning. The first question targets the core empirical comparison between different modeling paradigms, assessing whether recent advances in foundation models for tabular data translate into practical performance gains in a challenging biomedical task. Given the widespread use of gradient boosting models in microbiome studies, this comparison is particularly relevant for evaluating whether newer approaches offer practical benefits over established baselines.

The second research question is motivated by the observation that dataset size and feature dimensionality strongly influence the relative performance of different machine learning models. Transformer-based models such as TabPFN are designed to leverage prior knowledge and may be particularly effective in small- to medium-sized datasets, whereas tree-based ensemble methods are known to scale well with increasing data volume. Investigating how model performance changes as a function of sample size and feature selection provides insight into the regimes in which different modeling approaches are most appropriate and informs practical recommendations for future microbiome studies.

This thesis is structured as follows. Chapter 2 presents a literature review covering key aspects of microbiome research, the characteristics of microbiome data,

and some of the existing machine learning approaches for tabular biological data. This chapter provides the conceptual and methodological background necessary for understanding the challenges addressed in this work.

Chapter 3 describes the dataset and preprocessing steps used in the study, including data acquisition, filtering criteria, compositional data transformations, and feature selection. The modeling framework and experimental setup, including model configurations and evaluation protocols, are also detailed.

Chapter 4 presents the empirical results of the comparative modeling experiments. Predictive performance of Lasso regression, XGBoost, CatBoost, and TabPFN is reported and analyzed, including additional experiments examining the effect of dataset size and feature dimensionality.

Finally, Chapter 5 discusses the implications of the results, highlights methodological limitations, and outlines potential directions for future research. The thesis concludes with a summary of key findings and their relevance for both microbiome research and machine learning applications in biomedical data analysis.

2 Literature review

2.1 Microbiome research

The human microbiome has become an important subject of research due to its fundamental role in host physiology and disease [1]. Advances in sequencing technologies have enabled large-scale characterization of microbial communities, producing complex datasets that require specialized analytical approaches [2]. This section provides an overview of microbiome biology, the generation of microbiome sequencing data, and the statistical challenges associated with analyzing such datasets. Finally, the role of machine learning methods in microbiome research is introduced.

2.1.1 Biological relevance of the microbiome

In addition to human cells, the human body harbors an enormous number of microscopic organisms collectively known as microbes. The total number of microbes inhabiting the human body is estimated to be in the order of trillions. Although their sizes vary, microbes are invisible to the naked eye, typically ranging from approximately 0.02 to 20 µm in diameter. For comparison, a human hair is roughly 30 µm thick. Microbes encompass a diverse range of microorganisms, including bacteria, viruses, fungi, algae, and protozoa.

The human microbiome refers to the collective community of microorganisms and their genetic material residing in and on the human body. These microbial

communities inhabit various body sites, as shown in Figure 2.1, including the gastrointestinal tract, skin, oral cavity, respiratory tract, and urogenital tract. Among these, the gastrointestinal tract hosts the most diverse and densely populated microbial ecosystem, containing the majority of the body's microbial biomass. Due to its extensive metabolic capacity and close interaction with the host immune system, the gut microbiome has been the primary focus of microbiome research.

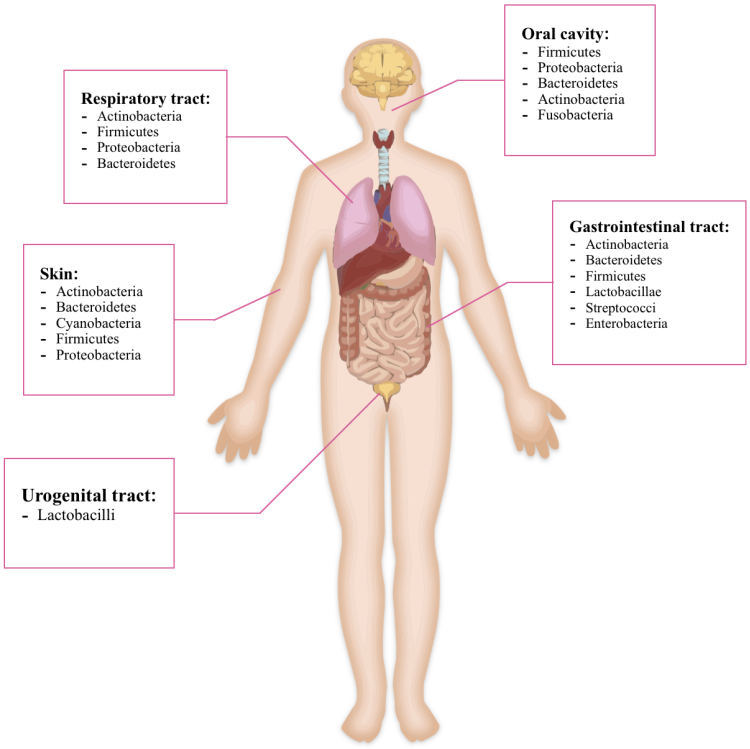


Figure 2.1: Distribution of dominant microbial communities across major human body sites, illustrating how microbial composition varies between environments such as the gut, skin, oral cavity, and respiratory tract [3].

The microbiome plays a critical role in human health and disease, and our understanding of its importance has evolved substantially over the past decades. Advances in sequencing technologies and computational methods have enabled researchers to move beyond simple taxonomic profiling toward functional and systems-level analyses. While early microbiome research primarily focused on the discovery and cataloging of microbial species, more recent studies aim to illustrate the functional roles

of microbial communities and their interactions with the human host. [4]

According to Vos et al. [5], the gut microbiota is considered to be one of the vital elements in regulating human health. Irregularities in the gut microbiota have been linked with many serious diseases like different types of cancer, obesity and type 2 diabetes. [5] The focus on microbiome research has been on describing the diversity and composition of microbiomes and finding correlations between host phenotypes and microbes. The recent advances are driven by multiple factors: the rapidly increasing number of published studies, as well as major technological improvements that enable more comprehensive and reproducible analyses. For instance, R packages and cloud-based platforms facilitate large-scale data integration and visualization [6] [7]. The gut microbiota has an important role in human health and this has led to a shift in microbiome research toward understanding the functional roles of microbial communities rather than solely describing their taxonomic composition.

Although microbiome research has revealed numerous associations between microbial communities and human health, the analysis of microbiome data presents several methodological challenges. These challenges arise from the statistical properties of microbiome datasets as well as from technical variability introduced during sequencing and data processing.

2.1.2 Microbiome data generation

Microbiome data are typically obtained from biological samples such as stool, which provide a non-invasive means of studying the gut microbial community [8]. After sample collection, microbial DNA is extracted and sequenced to identify the microorganisms present in the sample. Modern microbiome studies commonly rely on high-throughput sequencing (HTS) technologies, which allow large numbers of DNA fragments to be sequenced in parallel. These sequencing approaches enable the detection and classification of microbial taxa within a sample and generate profiles

describing the relative abundance of different microbial species. [2] The resulting sequencing reads are processed computationally and compared against reference databases to assign taxonomic identities. Several sequencing strategies have been developed for this purpose, with 16S rRNA gene sequencing and shotgun metagenomic sequencing being among the most widely used approaches [9].

One widely used approach is 16S rRNA gene sequencing, which targets a conserved bacterial gene containing both conserved and variable regions. The conserved regions allow amplification of the gene across many bacterial species, while the variable regions provide taxonomic information that enables identification of different microorganisms. After sequencing, reads are clustered into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) and matched against reference databases to estimate microbial composition. [10]

Although 16S rRNA sequencing is widely used due to its accessibility and reproducibility, it has limitations in taxonomic resolution and provides limited functional information. More comprehensive approaches such as shotgun metagenomic sequencing can provide higher taxonomic resolution and functional insights, but they are typically more computationally demanding [11]. Regardless of the specific sequencing technology used, the resulting microbiome data are typically represented as relative abundances of microbial taxa across samples.

In addition to challenges created by the complex nature of the microbiome, there are also technical challenges that play a significant role in the quality and interpretability of microbiome research [12]. Differences in sample collection, DNA extraction protocols, sequencing platforms and bioinformatics pipelines can all introduce bias and create difficulties in reproducibility across studies [13]. Even minor changes throughout these steps can lead to variation in detected microbial composition, making cross-study comparisons difficult and potentially unreliable. Efforts to standardize experimental protocols and bioinformatics pipelines aim to reduce

these sources of variability, although differences in available resources and research objectives continue to complicate full standardization across studies. These technical and methodological factors contribute to the complexity of microbiome datasets and must be considered when selecting appropriate statistical and computational analysis methods.

2.1.3 Statistical characteristics of microbiome data

Even with proper data analysis methods, the interpretation of the results is still a critical limitation. Most studies remain correlational, describing associations between microbial taxa and host phenotypes without clear evidence what drives these associations. Also temporal variability, like daily fluctuations, add another layer of complexity to the data [14]. These kinds of challenges highlight the need for robust and generalizable computational approaches that are capable of capturing nonlinear patterns in the data [15].

Despite the rapid progress and growing interest in microbiome research in the past years, there are still several methodological and biological challenges. One of the biggest issues the field is facing is the high degree of interindividual variability [5]. The composition of the microbiome varies substantially even between healthy individuals. This variation is due to factors like genetics, diet, environment, medication and age, making it challenging to establish universal microbial biomarkers or causal links [16].

Microbiome data have many characteristics that make it difficult to analyze, and the methods used to analyze it have to be chosen carefully to get valid results. The most notable challenges in microbiome data are compositionality, high dimensionality, and sparsity[12][17][18]. When working with compositional datasets, the components represent, for example, percentages or proportions of a whole, and they sum to a constant [12]. This comes from the fact that abundances measured

in HTS are converted to relative proportions, rather than absolute counts. This constant-sum constraint invalidates many standard statistical methods predicated on independent and unbound measurements. To address the compositionality issue, log-ratio transformation methods such as centered log-ratio (CLR) or additive log-ratio (ALR) have been developed.[19] These methods open the data from the simplex space back to Euclidean space, meaning that each sample of the data isn't scaled to sum to 1 anymore [20].

Microbiome datasets usually contain hundreds to thousands of taxa, or features, while the number of samples is comparatively low, resulting in high dimensionality. This kind of imbalance makes classical inference not an option and introduces other issues like covariance estimation and overfitting with predictive models. To handle high dimensionality, common methods include regularization, dimensionality reduction, and feature selection.[21]

The third main challenge when working with microbiome data is its sparsity. Sparsity arises from the large number of zero values, and in the context of microbiome data, this means that many taxa are either absent or below the detection threshold in most samples [19]. Sparsity in the data must be taken into account when doing analysis, and methods for that are, for example, zero-inflated models, imputation, or careful filtering of rare taxa. However, each of these has its trade-offs regarding bias and interpretability.[22]

Microbiome features are also organized hierarchically across multiple taxonomic levels, including species, genus, family, and higher ranks, which further complicates statistical analysis. In practice, these challenges require a cohesive workflow from selecting the proper transformation method (CLR/ALR) to applying dimensionality reduction and managing zero values via filtering or model choice. By acknowledging these issues, the validity of the research can be improved. In addition, it enhances the quality of downstream analysis, such as differential abundance tests

or machine learning predictions, improving reproducibility and interpretability of the research. These statistical characteristics motivate the use of machine learning methods that can capture complex, nonlinear relationships and operate effectively in high-dimensional feature spaces.

2.1.4 Machine learning in microbiome research

In recent years, machine learning has become an increasingly important tool in microbiome research. Advances in sequencing technologies have enabled the generation of large-scale microbiome datasets, creating opportunities to apply data-driven modeling approaches to investigate relationships between microbial communities and host phenotypes. Machine learning methods are particularly well-suited for microbiome studies because microbial ecosystems are inherently complex and high-dimensional, often involving hundreds or thousands of microbial taxa with intricate interaction structures. As a result, traditional statistical approaches may struggle to capture nonlinear relationships and multivariate patterns present in microbiome data [23].

A major application of machine learning in microbiome research has been the prediction of disease states and host phenotypes from microbial community composition. Several studies have demonstrated that microbiome profiles can be used to classify individuals according to disease status, including colorectal cancer, inflammatory bowel disease, and metabolic disorders. For example, Wirbel et al. [24] conducted a large-scale meta-analysis of colorectal cancer microbiome studies and demonstrated that machine learning models trained on microbial abundance profiles can identify robust microbial signatures associated with disease across different cohorts [24]. These findings highlight the potential of machine learning models to detect biologically meaningful signals within complex microbial ecosystems.

Beyond disease classification, machine learning has also been applied to predict

host physiological traits and lifestyle-related phenotypes. One of the most widely studied traits in microbiome research is body mass index (BMI) and obesity. Although numerous studies have reported associations between gut microbiome composition and obesity-related traits, the predictive performance of microbiome-based models remains moderate in most cases. Large cohort studies have shown that microbiome composition explains only a small fraction of the overall variation in BMI compared to factors such as diet, genetics, and lifestyle [25]. Nevertheless, microbial community composition may still provide complementary predictive information when combined with host-level metadata.

Recent work has also emphasized the importance of methodological best practices when applying machine learning to microbiome datasets. Microbiome data exhibit several statistical properties that can lead to biased or overly optimistic results if not handled carefully. For example, improper cross-validation strategies, feature selection performed outside the training folds, or inadequate control for confounding variables can substantially inflate predictive performance estimates. Topçuoğlu et al. [23] provide a systematic evaluation of common pitfalls in microbiome machine learning studies and propose guidelines for building more robust and reproducible predictive models [23]. These recommendations include careful separation of training and test data, appropriate handling of compositional data, and transparent reporting of model evaluation procedures.

In addition to classical machine learning methods, recent studies have begun exploring more advanced modeling approaches for microbiome analysis. Deep learning architectures and transformer-based models have shown promising results in various biological domains, although their application to microbiome data remains relatively limited. One challenge is that many deep learning methods require very large training datasets, whereas microbiome studies often operate in moderate sample size regimes. Nevertheless, emerging foundation models for tabular data, such as

transformer-based architectures pretrained on synthetic datasets, offer a potential avenue for improving predictive modeling in biological datasets with limited labeled data [26].

Despite the growing use of machine learning in microbiome research, several challenges remain. Predictive signals derived from microbiome composition are often subtle and distributed across many microbial taxa rather than dominated by a small number of highly predictive features. In addition, variability between cohorts, sequencing pipelines, and environmental factors can limit the generalizability of microbiome-based models. Consequently, the development of robust and generalizable machine learning methods for microbiome data remains an active area of research.

This thesis contributes to this research direction by empirically evaluating several machine learning approaches for predicting BMI from gut microbiome composition. In particular, the study compares established tree-based ensemble models with a recently proposed transformer-based foundation model for tabular data. By systematically comparing model performance across different dataset sizes and feature configurations, this work aims to provide insights into the suitability of modern machine learning models for microbiome-based prediction tasks.

2.2 Machine learning models for tabular data

Tabular data is among the most common data formats across various scientific and industrial domains, including biomedicine, physics, and economics [27]. In this representation, observations correspond to rows while variables correspond to columns, making it conceptually simpler than structured signals such as images or text [28]. However, despite its apparent simplicity, tabular data presents several algorithmic challenges. Unlike images or sequential data, tabular datasets typically lack explicit spatial or temporal structure that models can exploit through architectural

inductive biases. As a result, models must learn complex feature interactions and dependencies without the benefit of strong structural inductive biases [27].

One of the main challenges in applying machine learning to tabular data lies in its heterogeneity. Features can be numerical, categorical, or ordinal, often measured on vastly different scales and statistical distributions. Moreover, tabular datasets typically contain fewer observations compared to image or language corpora, limiting the applicability of deep learning methods that rely on large training samples. Additional complications arise from missing or noisy values, class imbalance, and the presence of outliers, all of which can significantly affect model robustness and generalization. [29]

Over the years, a variety of methods have been developed for predictive modeling with tabular data. Classical statistical models, such as regularized linear regression, remain widely used due to their interpretability and strong baseline performance. However, their limited ability to capture nonlinear relationships and complex feature interactions can restrict their predictive performance on more complex datasets. In contrast, machine learning methods such as tree-based ensemble methods, particularly gradient boosting algorithms like XGBoost and CatBoost, have emerged as the state-of-the-art due to their flexibility, robustness, and strong empirical performance [30]. These models will be examined in more detail in Sections 2.2.3 and 2.2.4.

While deep learning has historically underperformed on tabular data, recent research has introduced novel architectures, such as transformer-based models including TabNet and TabPFN, that aim to bridge this performance gap by capturing intricate feature interactions and probabilistic dependencies. The models selected for this study represent different modeling paradigms for tabular data, including a linear baseline model (Lasso regression), tree-based ensemble methods (XGBoost and CatBoost), and a transformer-based foundation model (TabPFN).

2.2.1 Lasso Regression

Lasso regression (Least Absolute Shrinkage and Selection Operator) is a regularized linear regression method designed to improve prediction accuracy and interpretability in high-dimensional datasets. The method was originally introduced by Tibshirani [31] and has since become a widely used baseline model in many machine learning applications, particularly in domains where the number of features may be large relative to the number of observations [32].

In contrast to ordinary least squares regression, which minimizes the residual sum of squares without any explicit constraint on model complexity, Lasso introduces an L_1 penalty on the regression coefficients. The objective function of Lasso regression can be written as

$$\min_{\beta_0,\beta} \left(\sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

where y_i denotes the response variable, x_i represents the vector of predictor variables for observation i, and β denotes the vector of regression coefficients. The regularization parameter λ controls the strength of the penalty and determines the degree of shrinkage applied to the model coefficients. [31]

A key property of the L_1 penalty is that it encourages sparsity in the estimated coefficients. As the value of λ increases, more coefficients are driven exactly to zero, effectively performing automatic feature selection. This property is particularly valuable in high-dimensional biological datasets such as microbiome data, where many features may contribute little predictive information. By shrinking irrelevant coefficients to zero, Lasso can reduce model complexity and improve interpretability while mitigating overfitting. [31]

Despite its advantages, Lasso regression is fundamentally limited to modeling linear relationships between predictors and the response variable. This limitation

can reduce predictive performance in complex biological systems where nonlinear relationships and interactions between microbial taxa may play an important role. Due to its simplicity, interpretability, and built-in feature selection capability, Lasso still remains an important baseline model for evaluating more flexible machine learning approaches. In addition, regularized linear models such as Lasso are particularly useful in biological datasets where the number of features may greatly exceed the number of samples.

2.2.2 TabPFN

TabPFN represents a shift in tabular data modeling by treating the prediction of labels from tabular data sets as a form of in-context learning. The model proposed by Hollmann et al. [30] is pretrained on millions of synthetic datasets generated via structural causal models. This allows learning a generic mapping from training sets and test instances to predictions without requiring dataset-specific tuning. At inference time, the training set and new test instance are jointly input to the network, which outputs a prediction for the test instance in a single forward pass.[30] This property makes TabPFN particularly interesting for scientific applications where labeled datasets are often limited, such as microbiome studies.

In contrast to traditional gradient-boosted decision trees or neural networks trained from scratch, TabPFN operates as a foundation model. It doesn't learn from the target dataset directly, but instead infers the most probable underlying data-generating process based on its pretraining distribution [30]. This allows the model to make strong predictions even with limited training data, which is particularly valuable in scientific applications where datasets are often limited in size. The model's probabilistic formulation also enables uncertainty estimation, providing confidence scores for each prediction, which can improve interpretability and trustworthiness in biological contexts. Furthermore, TabPFN's plug-and-play na-

ture means that it doesn't require hyperparameter optimization or feature scaling, which simplifies the modeling pipeline [27].

The architecture of TabPFN further addresses the heterogeneity of tabular data by employing randomized feature tokens, where each attribute is represented as a perturbation of a shared embedding vector, allowing the transformer to ingest varying numbers of features and types across datasets, while maintaining a fixed input format [27]. The model is evaluated on hundreds of small- to medium-scale tabular tasks (up to around 10000 samples and 500 features), achieving state-of-the-art performance, exceeding tuned gradient-boosted trees, such as CatBoost and XGBoost, and doing so with significantly lower computational cost [30].

Rather than explicitly estimating model parameters for each new dataset, TabPFN learns to approximate the posterior predictive distribution directly using a transformer architecture trained on a large collection of synthetically generated tabular datasets. During training, the model observes numerous simulated learning tasks and learns to map training datasets and query inputs directly to predictive distributions. As a result, the trained model can approximate Bayesian inference for new tabular datasets without explicitly performing posterior computation. [30]

In practice, TabPFN performs inference by treating the training dataset and query instance as a sequence processed by a transformer architecture, encoding both features and labels as tokens, and modeling dependencies among samples through the transformer's attention mechanism. Its causal self-attention ensures each token attends only to previous ones, while positional encoding enforces order invariance essential for tabular data. Consequently, the model functions as a neural prior-to-posterior approximator, replacing parameter estimation with direct posterior prediction. [27]

TabPFN can also be interpreted from a Bayesian perspective as approximating the posterior predictive distribution of a supervised learning problem. In the

Bayesian formulation, the goal is to compute the probability of a target value y given a new input x and an observed dataset D. This predictive distribution can be written as

$$p(y \mid x, D) = \int p(y \mid x, \theta)\, p(\theta \mid D)\, d\theta.$$

This formulation expresses the predictive distribution as an average over all possible data-generating processes consistent with the observed dataset. Here, x denotes the feature vector of a new observation and y the corresponding target variable to be predicted. The dataset D consists of n observed training samples

$$D = \{(x_i, y_i)\}_{i=1}^{n},$$

where x_i represents the feature vector of the i-th sample and y_i its associated target value. The variable θ denotes latent parameters describing the underlying data-generating process.

In the expression above, $p(\theta \mid D)$ denotes the posterior distribution of the model parameters given the observed dataset, while $p(y \mid x, \theta)$ represents the likelihood of observing the target value y for input x under parameter configuration θ. The integral therefore corresponds to Bayesian model averaging over all possible parameter values, weighted by their posterior probabilities. [30]

TabPFN still poses some limitations. Its design exhibits a significant performance drop on large-scale datasets and high-dimensional feature spaces, likely caused by its transformer backbone scaling and its pretraining regime tailored for smaller scales [27]. To address this, Ye et al. [27] proposes extensions, such as divide-and-conquer token subsets and ensemble strategies, which split features or samples into more manageable chunks and recombine predictions [27]. In a microbiome modeling context, where high-dimensionality, sparsity, and compositionality in data are

often issues, TabPFN's capability to generalize without extensive tuning offers a compelling advantage, even though careful preprocessing is still essential.

2.2.3 XGBoost

Gradient boosting, or tree boosting, is an effective machine learning method that is widely used in different areas, for example, on tabular data. In 2016, Chen and Guestrin [33] presented a model called XGBoost (Extreme Gradient Boosting), which extends the classical gradient boosting framework by combining second-order optimization, regularization, and a series of system-level engineering enhancements that together achieve high predictive performance with good scalability and computational effectiveness. The model builds an additive ensemble of decision trees, where each new tree corrects the residual errors of the previous ensemble, progressively improving predictive performance. This model structure has made XGBoost a cornerstone model for structured data across different fields such as bioinformatics, finance, and physics. [33]

Mathematically, XGBoost can be interpreted as optimizing the ensemble function

$$\hat{y}i = \sum k = 1^K f_k(x_i),$$

where each f_k belongs to the space of regression trees $\mathcal{F}$. The optimization process seeks to minimize the global loss

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k),$$

effectively balancing predictive accuracy and model simplicity. The inclusion of the regularization term $\Omega(f_k)$ makes XGBoost closer in spirit to statistical learning theory, embedding control of model complexity directly into the objective function. [33]

In contrast to deep learning models that learn global feature representations, XGBoost focuses on learning local decision rules that capture nonlinear dependencies between variables. Each iteration adds a new tree $f_t(x)$ that minimizes a regularized objective function

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\big(y_i, \hat{y}i^{(t-1)} + f_t(x_i)\big) + \Omega(f_t),$$

where l is a differentiable loss function, and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum j = 1^T w_j^2$ controls model complexity through the number of leaves T and leaf weights w_j. This explicit regularization is a key innovation that prevents overfitting and stabilizes model behavior, especially on small or noisy datasets. [33]

Classical boosting algorithms usually use only first-order gradients, but XGBoost leverages both the gradient and Hessian of the loss function through a second-order Taylor expansion. This allows the model to estimate not only the direction, but also the curvature of the loss landscape, resulting in faster convergence and more stable optimization. Each potential tree split is evaluated by its information gain, ensuring that only the most statistically meaningful partitions are selected. [33]

One of the main features of XGBoost is its capability to handle missing and sparse data natively. During training, the algorithm automatically learns a default direction for missing feature values, meaning it can assign them to the left or right branch of a split based on which direction improves performance. [33] This design feature makes XGBoost especially useful for biological datasets, where missing values are a common problem. This also makes it a useful algorithm for microbiome data, where data sparsity usually brings challenges.

When compared to neural methods such as TabPFN, XGBoost relies less on data representation learning and more on explicit structural partitioning of the feature space. This difference also explains its continued dominance in small- to medium-scale tabular problems, where the number of samples is limited and interpretability

is essential. Feature importance metrics derived from split statistics, including gain, cover, and frequency, enable transparent insight into which variables most influence model predictions. [33]

Despite its advantages, XGBoost still has some limitations. The model relies on handcrafted decision boundaries, which can limit its expressiveness compared to deep learning models that capture higher-order feature interactions. In addition, XGBoost often requires extensive hyperparameter tuning to reach its optimal performance. [33] Compared to TabPFN, which doesn't require hyperparameter tuning, that is a significant disadvantage. Because of its interpretability, robustness, and speed, XGBoost is still a valid choice for microbiome data analysis, where sparsity, compositionality, and high-dimensionality are common challenges.

In summary, XGBoost combines the principles of gradient boosting with rigorous regularization and modern computational efficiency. It has set a standard for machine learning models on tabular data, providing a balance between predictive capabilities, interpretability, and computational cost. It continues to be used as a benchmark when evaluating new model structures, such as TabPFN, which demonstrates its relevance and impact as a powerful model.

2.2.4 CatBoost

CatBoost was proposed in 2018 by Prokhorenkova et al. [34]. It is a gradient boosting framework that introduces algorithmic innovations specifically designed to eliminate prediction shift and reduce target leakage during training. CatBoost is built around the insight that both gradient estimation and categorical feature encoding can unintentionally allow information from the label of the current sample to influence its own prediction. This leads to biased gradients, suboptimal tree construction, and degraded generalization performance. CatBoost addresses this issue by using ordered boosting, a training procedure that ensures each model update for a given

data point is computed without access to its own label, thereby aligning the distribution of predictions between training and inference.[34]

Another major contribution by Prokhorenkova et al. [34] is CatBoost's approach to handling categorical features via ordered target statistics. Instead of one-hot encoding or naive target encoding, which use information from the entire dataset and therefore leak label information, CatBoost computes for each sample a smoothed estimate of the target mean using only those data points that precede it in a random permutation of the training set. Formally, for a sample k in permutation σ and categorical feature f, CatBoost computes

$$\hat{x}f, k = \frac{\sum j < k \mathbf{1}\{x_{f,j} = x_{f,k}\} y_j + a \cdot p}{\sum_{j<k} \mathbf{1}\{x_{f,j} = x_{f,k}\} + a},$$

where a is a prior strength parameter and p is the global mean target value [34]. This formulation ensures that category encodings are unbiased and prevents the artificial inflation of model accuracy during training. When combined with CatBoost's symmetric tree structure, where each level of the tree splits on the same feature across all nodes, the model becomes computationally more efficient and helps to avoid overfitting.

CatBoost demonstrates strong performance, especially on heterogeneous tabular datasets. In the benchmarks reported by Prokhorenkova et al. [34], CatBoost consistently outperforms XGBoost and LightGBM on tasks from different domains, such as e-commerce, biology, and recommendation systems. The main advantages of CatBoost are its principled handling of categorical variables, robustness to noisy or unbalanced data, and reduced need for extensive hyperparameter tuning, even though it is still required to get the most accurate results. [34] CatBoost's native handling of categorical data makes it especially appealing in real-world scenarios, where categorical metadata is abundant and preprocessing capabilities are limited.

CatBoost still has many limitations and practical challenges. One difficulty is

computational cost. Ordered boosting requires maintaining multiple permutations and training auxiliary models, which increases memory consumption and slows down training, especially on large-scale datasets. Another challenge is CatBoost's interpretability. While tree-based models are generally more interpretable than, for example, neural networks, CatBoost's symmetric tree structure and reliance on target statistics make feature importance less straightforward to interpret. Additionally, even though CatBoost performs well on categorical variables, it has challenges with purely numerical, high-sparsity datasets, such as microbiome data. To overcome this challenge, CatBoost may require extensive preprocessing to avoid overfitting and instability. [34] [35]

In summary, CatBoost offers strong performance on heterogeneous datasets with robust handling of categorical variables, but falls short on numerical data. With extensive preprocessing, it can still be a valid model for tasks across various domains, such as microbiome research.

2.3 Summary and gaps in research

Microbiome research has become more data-driven as newer technologies, such as high-throughput sequencing, have enabled larger-scale profiling of microbial communities [36]. At the same time, the resulting data exhibits properties like compositionality, sparsity, and high dimensionality, which complicates statistical analysis and challenges traditional machine learning approaches [12]. Recent methodological developments in microbiome analysis focus on resolving these issues, for example, through log-ratio transformations and dimensionality reduction pipelines. Despite the recent progress, there is still no universally accepted workflow for preprocessing and modeling microbiome data [37].

For tabular machine learning models, gradient-boosted decision trees remain the dominant choice in practical applications due to their robustness, interpretability,

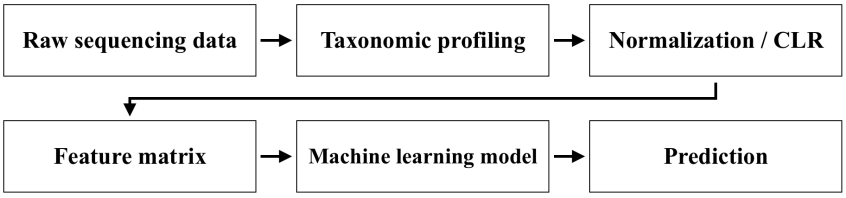


Figure 2.2: Typical workflow in predictive microbiome research, illustrating the steps from sequencing and preprocessing to feature extraction, machine learning modeling, and biological interpretation.

and ability to handle heterogeneous feature spaces with relatively little preprocessing [27]. TabPFN represents a recent shift towards transformer-based models. These models show promising results, especially in smaller-scale datasets, and require less dataset-specific tuning. Transformer-based models are currently limited by scalability and training distributions that may not fully capture the statistical properties of real-world scientific datasets.

There are still several research gaps across both microbiome research and machine learning applications. Microbiome-specific adaptations of modern machine learning methods are still relatively underdeveloped, and many models struggle with the compositional nature of microbiome data [12]. In addition, the rapid development of new analytical pipelines has led to a lack of standardized workflows, making systematic comparisons across studies difficult [37]. Model interpretability also remains a challenge. Although tree-based models provide some degree of feature attribution, understanding causal relationships in high-dimensional microbial ecosystems remains difficult.

At the same time, recent advances in machine learning have introduced foundation models that demonstrate strong performance in domains such as image and language processing [30]. However, general-purpose foundation models for tabular biological data remain relatively limited. In particular, relatively little work has systematically compared recent foundation models for tabular data, such as TabPFN,

with established tree-based approaches in microbiome-based prediction tasks. Furthermore, it remains unclear how the performance of these models changes as a function of dataset size and feature configuration in microbiome prediction problems.

Altogether, the literature highlights both the advances and limitations of current technologies. Machine learning methods from classical models to transformers achieve strong predictive performance, but their applicability to microbiome data requires careful adaptation. Future research will benefit from the development of models that natively incorporate compositional structures, scalable approaches trained specifically on biologically realistic data distributions, and standardized benchmarking frameworks that improve comparability and reproducibility across studies. Motivated by these gaps, this thesis evaluates several machine learning models representing different paradigms for tabular data, including Lasso regression, gradient-boosted decision trees, and the transformer-based TabPFN model. The goal is to assess their suitability for predicting body mass index (BMI) from gut microbiome composition and to examine how predictive performance changes under different data regimes.

3 Methodology

This chapter describes the methodological framework used in this study for predicting body mass index (BMI) from gut microbiome data. First, the microbiome dataset and preprocessing steps are introduced, including filtering criteria and compositional data transformations. Next, the machine learning models evaluated in this work are presented together with their training procedures and hyperparameter settings. Finally, the evaluation metrics and experimental setup used to compare model performance are described.

3.1 Datasets and preprocessing

The gut microbiome data used in this study were obtained from the Metalog database [38], a large-scale resource that aggregates standardized human microbiome profiles together with curated sample metadata. The database integrates microbiome sequencing data processed using uniform bioinformatics pipelines, enabling large-scale comparative analyses across cohorts and studies. In this work, species-level taxonomic profiles were generated using MetaPhlAn4 [39], which estimates the relative abundance of microbial taxa based on clade-specific marker genes.

Microbial profiles were provided as relative abundances at the species level for each sample. Corresponding metadata included demographic variables such as age and sex, as well as body mass index (BMI), which served as the primary outcome variable in all predictive modeling experiments.

Only adult fecal samples were considered in this study, as gut microbiome composition is known to vary substantially across developmental stages and body sites. Samples annotated as non-fecal or environmental were excluded, as were artificial or control samples. In addition, samples with missing BMI, age, or sex information were removed to ensure consistent availability of target and covariate variables across all experiments.

After applying these filters, microbiome profiles were matched with metadata using unique sample identifiers. Samples for which a one-to-one correspondence between taxonomic profiles and metadata could not be established were discarded. This matching step ensured that each observation in the final dataset represented a single individual with both microbiome composition and relevant phenotypic information available.

Microbiome data are inherently high-dimensional and sparse, with many taxa appearing only in a small fraction of samples. To reduce noise and improve model robustness, microbial species present in fewer than 25 % of samples were removed prior to modeling. This prevalence-based filtering is commonly applied in microbiome studies to limit the influence of extremely rare taxa, which may contribute limited predictive signal while substantially increasing model complexity and susceptibility to overfitting.

However, this filtering step also introduces a potential limitation. Rare microbial species may in some cases carry biologically meaningful information related to host phenotypes such as BMI. Consequently, removing low-prevalence taxa could theoretically exclude features that contribute to predictive performance. Despite this possibility, extremely rare taxa often exhibit unstable abundance estimates and may reflect sequencing noise or cohort-specific variation rather than robust biological signals [40]. In practice, prevalence filtering therefore represents a trade-off between retaining potentially informative rare taxa and improving statistical stabil-

ity in high-dimensional datasets.

Following this filtering step, the dataset retained 214 microbial species as features. This dimensionality was considered suitable for both classical regression models and more expressive machine learning approaches, while still capturing a broad representation of the gut microbial community.

Because microbiome abundance data are compositional in nature, representing relative rather than absolute quantities, standard statistical and machine learning methods may produce misleading results if applied directly. To address this issue, relative abundance profiles were transformed using the centered log-ratio (CLR) transformation, which maps compositional data from the simplex to real-valued Euclidean space. A small pseudocount was added to all abundance values prior to transformation to avoid undefined logarithms caused by zero counts. The CLR transformation was applied separately to each sample, ensuring that transformed feature values captured relative differences in microbial abundance while preserving comparability across samples.

The final machine learning dataset consisted of 9,709 samples and 214 CLR-transformed microbial species features, with BMI serving as the continuous target variable. Each row represented a single individual sample, and each column corresponded to a microbial species feature or a phenotypic variable. In subsequent experiments, alternative feature sets were also constructed, including models based solely on microbiome features, solely on metadata (age and sex), and on their combination, to assess the relative contribution of microbiome information to BMI prediction.

This dataset served as the common input for all downstream modeling experiments, including baseline regression and advanced machine learning models. By applying consistent preprocessing steps across all models and experimental settings, differences in predictive performance could be attributed primarily to model choice and learning capacity rather than differences in data handling.

3.2 Models and hyperparameters

This section describes the machine learning models used in this study, along with their parameter settings and training procedures. The selected models represent different modeling paradigms, ranging from linear regression with regularization to tree-based ensemble methods and a recent foundation model for tabular data. This diverse model set enables a comprehensive comparison of predictive performance and model behavior on high-dimensional microbiome data.

All models were trained on the same training splits and evaluated using identical performance metrics to ensure a fair and consistent comparison. Hyperparameter tuning was applied where applicable using cross-validation on the training data, while final performance was reported on a held-out test set.

3.2.1 Lasso regression

Lasso regression was used as the baseline linear model in this study. In high-dimensional settings such as microbiome data, where the number of features can be large relative to the number of samples, regularization is essential to prevent overfitting and improve model interpretability. The L1 penalty encourages sparsity in the regression coefficients, effectively selecting a subset of microbial species that are most strongly associated with the target variable.

Model training was performed using the glmnet package. The regularization parameter λ was selected via 10-fold cross-validation by minimizing the cross-validated mean squared error. Predictor variables were standardized prior to model fitting, as the scale of input features influences the strength of regularization in penalized linear models.

Although Lasso regression is limited to linear relationships between features and the outcome variable, it provides a transparent and interpretable baseline against which more flexible non-linear models can be compared. In addition, the sparsity of

the learned coefficients facilitates biological interpretation by highlighting a limited number of microbial taxa with non-zero contributions to BMI prediction.

3.2.2 XGBoost

XGBoost was used to model non-linear relationships and interactions between microbial features. XGBoost is a tree-based ensemble method that constructs an additive model of decision trees, where each subsequent tree is trained to correct the residual errors of the previous ensemble. This boosting framework has demonstrated strong performance across a wide range of structured prediction tasks, including biological and biomedical applications.

Model hyperparameters, including learning rate, maximum tree depth, subsampling rate, and column sampling rate, were optimized using 5-fold cross-validation with early stopping based on validation RMSE. Early stopping was used to determine the optimal number of boosting iterations, preventing overfitting by terminating training once performance on validation folds ceased to improve. The optimal number of boosting rounds was then used to train the final model on the full training set.

XGBoost is particularly well-suited for microbiome data due to its ability to handle non-linear associations, interactions between microbial taxa, and heterogeneous feature distributions. In contrast to linear models, tree-based ensembles can capture threshold effects and conditional dependencies between microbial species, which may reflect more complex ecological or host–microbe relationships relevant to BMI.

3.2.3 CatBoost

CatBoost was included as an additional gradient boosting model for comparison with XGBoost. While CatBoost was originally designed to handle categorical features ef-

fectively, it has also shown competitive performance on numerical, high-dimensional datasets such as microbiome profiles.

Model training was conducted using CatBoost's built-in training procedure with early stopping. To mitigate overfitting, early stopping was applied based on validation performance, and the number of boosting iterations was selected accordingly. Key hyperparameters, such as tree depth and learning rate, were chosen to balance model expressiveness and computational efficiency.

The use of symmetric trees constrains the model structure, which can improve stability and reduce variance compared to fully unconstrained tree ensembles. This property is particularly relevant in microbiome applications, where feature correlations and noise may lead to unstable splits in standard decision tree models. CatBoost thus serves as a complementary tree-based approach to XGBoost, enabling comparison between different boosting strategies on the same prediction task.

3.2.4 TabPFN

The TabPFN regressor was included to evaluate the performance of transformer-based foundation models on microbiome prediction tasks. Rather than learning model parameters directly from the target dataset, TabPFN performs inference by conditioning on the training examples provided at prediction time, effectively framing tabular prediction as a form of in-context learning.

In contrast to traditional machine learning models, TabPFN requires minimal hyperparameter tuning and can produce competitive predictions even in low-data regimes. This property makes it particularly interesting for scientific applications, where labeled datasets are often limited in size. In this study, the TabPFN regressor was applied using default settings provided by the authors.

Model training and inference were performed in Python, and predictions were evaluated using the same performance metrics as the other models to ensure com-

parability. The use of TabPFN enables exploration of whether foundation models pretrained on generic tabular distributions can generalize to real-world biological data such as microbiome profiles. While TabPFN offers limited transparency compared to classical models, its inclusion provides an informative contrast between conventional supervised learning approaches and emerging foundation models for tabular data.

3.3 Evaluation metrics

Model performance was evaluated using multiple regression metrics in order to capture different aspects of predictive accuracy and error behavior. Relying on a single evaluation metric can provide a limited or potentially misleading view of model performance, particularly in continuous prediction tasks where different types of errors may have different practical and scientific implications.

The coefficient of determination (R^2) was used to quantify the proportion of variance in BMI explained by the model predictions. As a normalized measure of goodness-of-fit, R^2 provides an intuitive summary of overall predictive performance. However, it does not directly convey the magnitude of prediction errors and can be sensitive to the variance of the target variable, which limits its interpretability when used in isolation.

To complement R^2, two error-based metrics were also included. Root mean squared error (RMSE), defined as the square root of mean squared error (MSE), expresses prediction error in the same units as the target variable (BMI), which facilitates interpretation of model accuracy in practical terms. Mean absolute error (MAE) provides a more robust estimate of typical prediction error by weighting all deviations linearly, making it less sensitive to extreme outliers than RMSE.

During model selection and hyperparameter tuning, performance was assessed using cross-validation on the training data only. RMSE was used as the primary

optimization criterion, as it emphasizes larger prediction errors, which is particularly relevant in biomedical and health-related applications where substantial deviations in predicted BMI may be undesirable. Final model performance was evaluated on an independent held-out test set using all reported metrics, enabling a comprehensive and transparent comparison of predictive performance across models.

3.4 Experimental setup

The experimental setup was designed to enable a fair, controlled, and reproducible comparison between different machine learning models for predicting body mass index (BMI) from gut microbiome data.

The final dataset was randomly partitioned into training and test sets using an 80/20 split. A fixed random seed was used throughout all experiments to ensure reproducibility and consistency across model runs. The test set was held out entirely during model training and hyperparameter optimization and was used exclusively for final performance evaluation.

Model selection and hyperparameter tuning were conducted using cross-validation applied only to the training data. For models supporting internal cross-validation procedures, such as XGBoost and Lasso regression, k-fold cross-validation was used to identify optimal hyperparameters. Early stopping was employed where supported to mitigate overfitting and to determine the appropriate number of training iterations. Importantly, no information from the test set was used at any stage of model tuning or selection, ensuring a strict separation between training and evaluation data.

All models were trained using the same preprocessed feature matrix and the same continuous BMI target variable. This ensured that observed differences in predictive performance could be attributed primarily to model characteristics rather than differences in data preprocessing or feature selection. For each method, the

final model was retrained on the full training set using the best-performing hyperparameters identified during cross-validation.

Model performance was evaluated on the held-out test set using a consistent set of evaluation metrics, including the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). This combination of metrics was chosen to capture both overall explanatory power and the magnitude of prediction errors, enabling a comprehensive and transparent comparison across models.

In addition to predictive performance, model interpretability was also examined. Feature importance measures and SHAP values were computed for tree-based models and TabPFN to assess the contribution of individual microbial species to BMI predictions. These analyses were performed post hoc and did not influence model training or model selection.

All analyses were conducted using a combination of R and Python. R was primarily used for data preprocessing, statistical modeling, gradient boosting models, and visualization, while Python was used for implementing the TabPFN model and selected post-processing steps. Trained models, predictions, and evaluation results were systematically saved to disk to ensure reproducibility and to support downstream analyses and visualization.

4 Results

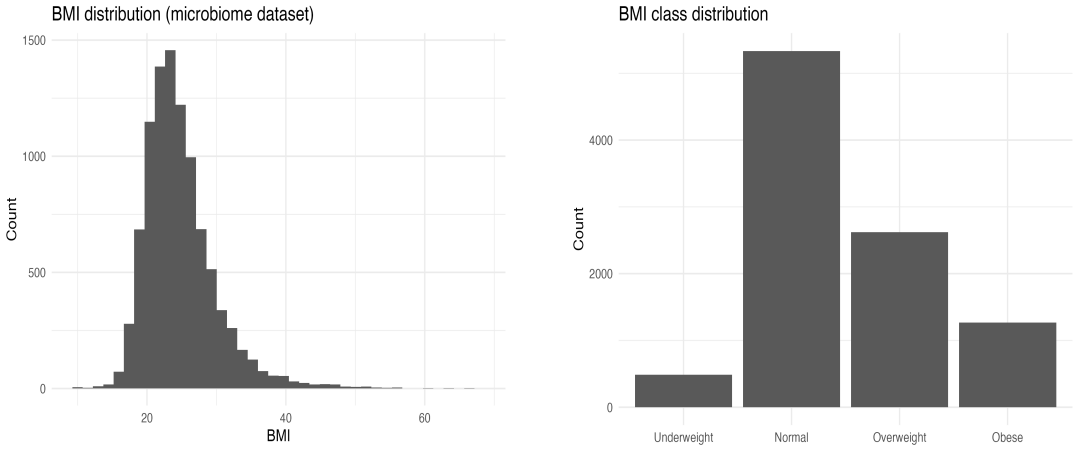
4.1 Descriptive analysis

The descriptive analysis provides an overview of the study cohort and the structure of the gut microbiome data prior to predictive modeling. The final dataset comprised 9,709 adult fecal samples with complete microbiome profiles and associated body mass index (BMI) information.

The distribution of BMI values in the study population is shown in Figure 4.1a. The mean BMI was 24.8 (SD = 5.25), with a median of 23.9 and an interquartile range of 5.70, indicating substantial variability in body mass across individuals. BMI values ranged from 10 to 67.9, reflecting the inclusion of both underweight and severely obese individuals.

Participants were stratified into four BMI categories following standard clinical thresholds [41]: underweight (5.0 %), normal weight (54.9 %), overweight (27.0 %), and obese (13.0 %). The majority of samples thus originated from individuals with normal weight or overweight status, while the underweight and obese categories were comparatively smaller (Figure 4.1b). This class imbalance reflects typical population-level BMI distributions and should be considered when interpreting model performance, particularly for extreme BMI values.

The age distribution of the study population is shown in Figure 4.2. The dataset consists exclusively of adult participants, with a mean age of 50.5 years and a median



(a) BMI distribution (b) BMI classes

Figure 4.1: Distribution of body mass index (BMI) in the study cohort. Panel (a) shows the continuous BMI distribution, while panel (b) illustrates the distribution of individuals across standard BMI categories.

of 53 years. The interquartile range spans from 35 to 65 years, indicating that the majority of samples originate from middle-aged and older adults. This age distribution is relevant for the interpretation of the predictive models, as both gut microbiome composition and BMI are known to vary across the lifespan. Including age as a covariate in selected modeling setups allows separating microbiome-specific predictive signal from broader demographic effects.

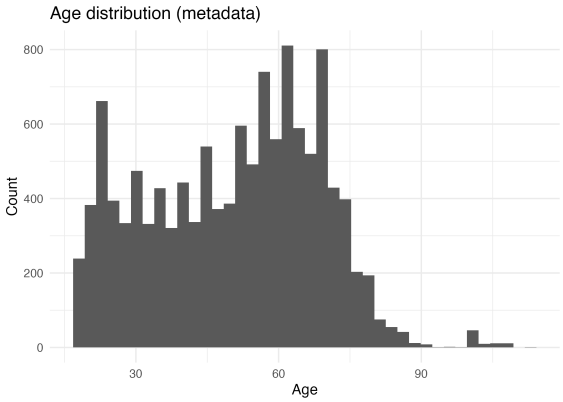


Figure 4.2: Age distribution

To explore large-scale patterns in gut microbiome composition, beta diversity was computed using Bray–Curtis dissimilarity, a commonly used ecological distance

metric for comparing community composition between samples and visualized using principal coordinates analysis (PCoA) (Figure 4.3). Samples were colored according to BMI class. While no clear separation between BMI categories was visually apparent in the low-dimensional embedding, a small but statistically significant association between BMI class and microbiome composition was observed.

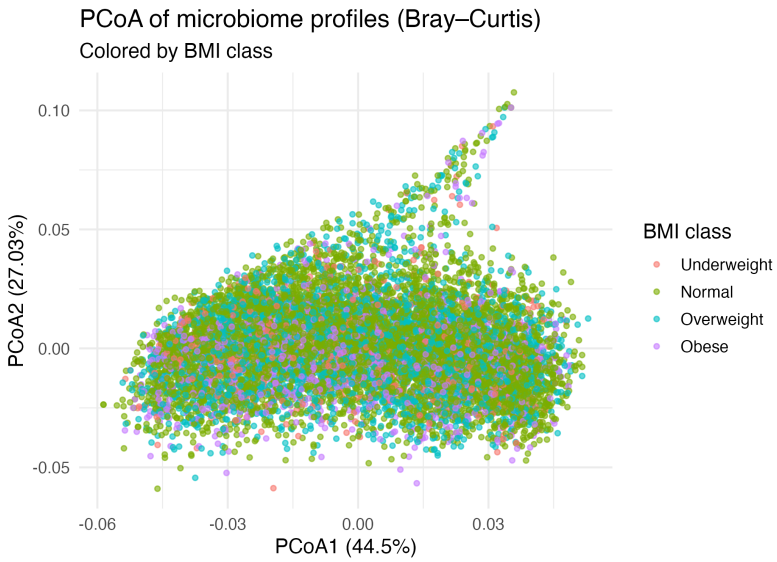


Figure 4.3: PCoA of microbiome profiles

Permutational multivariate analysis of variance (PERMANOVA) revealed that BMI class explained approximately 0.58 % of the total variance in microbial community composition ($R^2 = 0.00579$, p = 0.001). Although this effect size is small, the highly significant p-value reflects the large sample size and indicates that BMI is associated with subtle but detectable shifts in gut microbiome structure. These results are consistent with previous large-cohort microbiome studies, in which host phenotypes often account for only a small fraction of overall microbial variance [42].

To further characterize the relationship between individual microbial taxa and BMI, Spearman rank correlations were computed between species-level CLR-transformed abundances and BMI. This univariate analysis provides an initial view of how individual microbial species are associated with host body mass. Figure 4.4 presents the top 20 microbial species with the strongest positive or negative correlations with

BMI.

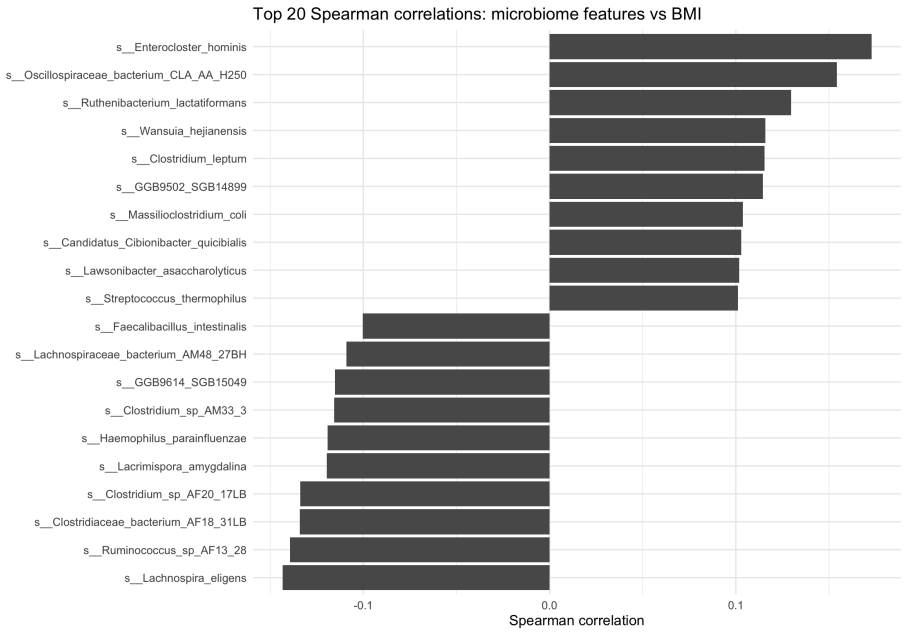


Figure 4.4: Top 20 microbial species ranked by Spearman correlation with BMI. Positive correlations indicate taxa associated with higher BMI values, while negative correlations indicate taxa associated with lower BMI values.

Overall, the observed correlations were modest in magnitude, with absolute Spearman correlation coefficients ranging approximately from 0.10 to 0.17. The strongest positive association with BMI was observed for Enterocloster hominis ($\rho \approx 0.17$), followed by members of the Oscillospiraceae family and Ruthenibacterium lactatiformans. Conversely, several taxa exhibited negative correlations with BMI, including Lachnospira eligens, Ruminococcus sp. AF13-28, and multiple representatives of the Clostridiaceae family.

Several of the taxa identified among the top correlates have previously been linked to host metabolic processes and gut health in the literature. For example, genera such as Ruminococcus, Lachnospira, and Clostridium include species known to participate in the fermentation of dietary fibers and the production of short-chain fatty acids, which play an important role in host energy metabolism [43]. The presence of these taxa among the most strongly associated species suggests that the

observed statistical associations may reflect biologically meaningful links between gut microbial composition and host metabolic state.

Nevertheless, the relatively small effect sizes indicate that no single microbial species exhibits a strong linear association with BMI. This is consistent with the multivariate and highly complex nature of the gut microbiome, where host phenotypes are likely influenced by collective microbial community structure and functional interactions rather than by individual taxa in isolation. Moreover, Spearman correlation captures only monotonic univariate relationships and does not account for non-linear effects or interactions between microbial species.

Taken together, the descriptive analyses suggest that BMI is associated with subtle but statistically detectable shifts in gut microbiome composition at both the community and species levels. While low-dimensional ordination methods (such as PCoA) revealed only weak separation between BMI categories, species-level correlation analysis highlights that meaningful structure exists within the high-dimensional feature space. These findings motivate the use of multivariate machine learning models capable of capturing complex, non-linear patterns and interactions across large sets of microbial features. The following sections therefore focus on evaluating how effectively different modeling approaches can leverage these subtle microbiome signals to predict BMI.

4.2 Predictive model performance

This section presents the predictive performance of the evaluated machine learning models. The experiments assess how well different models predict body mass index (BMI) from gut microbiome data under several experimental conditions. First, model performance is evaluated using the full dataset. Next, the effect of dataset size on predictive performance is examined. Finally, the influence of feature set composition on model performance is analyzed.

4.2.1 Full dataset

All models were evaluated on the held-out test set using only microbiome-derived features. Overall predictive performance was modest across all models, reflecting the inherent difficulty of predicting body mass index from gut microbiome composition alone.

Model	n	R^2	RMSE	MAE
Lasso	7767	0.151	4.892	3.525
XGBoost	7767	0.159	4.869	3.495
CatBoost	7767	0.147	4.906	3.512
TabPFN	7767	**0.228**	**4.752**	**3.428**

Table 4.1: Predictive performance of the evaluated models on the full dataset using microbiome-derived features. Higher R^2 values indicate better predictive performance, while lower RMSE and MAE values indicate smaller prediction errors.

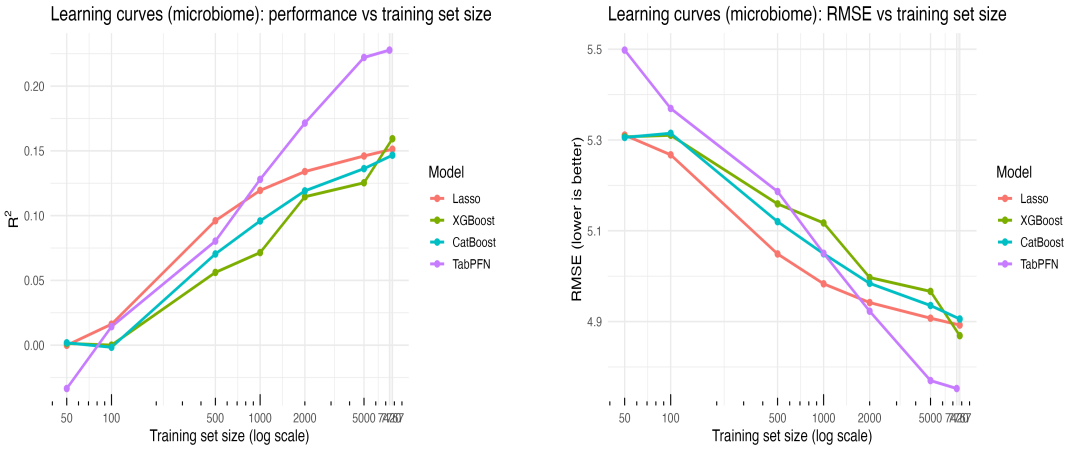
Among the evaluated methods, TabPFN achieved the highest predictive performance with an R^2 of 0.228, outperforming all tree-based and linear baselines. The corresponding RMSE of 4.75 and MAE of 3.43 indicate a moderate reduction in prediction error compared to traditional machine learning approaches. In contrast, Lasso regression, XGBoost, and CatBoost achieved similar performance levels, with R^2 values in the range of approximately 0.14–0.16.

These results suggest that while classical models are able to extract some predictive signal from microbiome composition, the transformer-based TabPFN model benefits from its strong inductive priors and ability to model complex feature interactions, even in high-dimensional biological data.

4.2.2 Effect of training set size

To investigate how model performance scales with the size of the training dataset, learning curves were constructed by training each model on increasingly larger subsets of the training data while evaluating performance on a fixed test set. Dataset sizes ranged from 50 samples to the full training set.

For very small training sets (50–100 samples), all models exhibited near-zero or even negative R^2 values, indicating that meaningful predictive structure could not be learned from such limited data. In this regime, TabPFN performed slightly worse than the tree-based models, reflecting the fact that foundation models for tabular data require a minimal amount of data to adapt effectively to a new task.



(a) R^2 (b) RMSE

Figure 4.5: Effect of training set size on R^2 and RMSE

As the training set size increased, predictive performance improved steadily for all methods, indicating that BMI prediction from microbiome data benefits from larger sample sizes. Overall, tree-based ensemble methods consistently outperformed Lasso regression across all dataset sizes, suggesting that non-linear relationships and feature interactions play an important role in linking gut microbiome composition to BMI.

The performance gap between XGBoost and CatBoost remained relatively small

throughout the scaling experiment, indicating that both gradient boosting frameworks are similarly capable of modeling the underlying structure of the data. In contrast, TabPFN exhibited the steepest performance gains as the dataset size increased. While its performance in very small-sample regimes was comparable to or weaker than that of the tree-based models, TabPFN eventually surpassed all other methods when trained on more than approximately 2,000 samples.

At the largest dataset size, TabPFN achieved an R^2 of approximately 0.23, whereas the best-performing tree-based model achieved an R^2 of approximately 0.16. These results highlight a key trade-off: while classical machine learning models may be more data-efficient in low-sample regimes, transformer-based foundation models appear to benefit from larger datasets. This suggests that models such as TabPFN may be particularly attractive for large-scale microbiome studies, where sufficiently large cohorts are available to fully exploit the model's representational capacity.

4.2.3 Feature set comparison

To disentangle the relative contributions of microbiome composition and basic demographic variables, additional experiments were conducted using (i) microbiome features only, (ii) metadata only (age and sex), and (iii) the combined feature set.

Using metadata alone resulted in weak predictive performance ($R^2 \approx 0.04$), indicating that age and sex explain only a small fraction of BMI variance in the studied cohort. In contrast, microbiome features alone provided substantially higher predictive power ($R^2 \approx 0.23$). The highest performance was achieved when combining microbiome and metadata features ($R^2 \approx 0.31$), suggesting that demographic variables provide complementary information to microbial composition.

This result supports the hypothesis that gut microbiome profiles capture biologically meaningful information related to metabolic status, while also highlighting the importance of integrating host-level covariates for improved prediction accuracy.

4.2.4 Feature importance agreement across models

To assess the consistency of learned feature importance across models, Spearman rank correlations were computed between the SHAP-based feature importance rankings of XGBoost, CatBoost, and TabPFN. The highest agreement was observed between XGBoost and TabPFN ($\rho \approx 0.68$), whereas CatBoost exhibited lower concordance with both XGBoost and TabPFN.

Model	XGBoost	CatBoost	TabPFN
XGBoost	1.000	0.375	**0.675**
CatBoost	0.375	1.000	0.395
TabPFN	**0.675**	0.395	1.000

Table 4.2: Spearman rank correlations between SHAP feature importance profiles of different models (microbiome-only dataset).

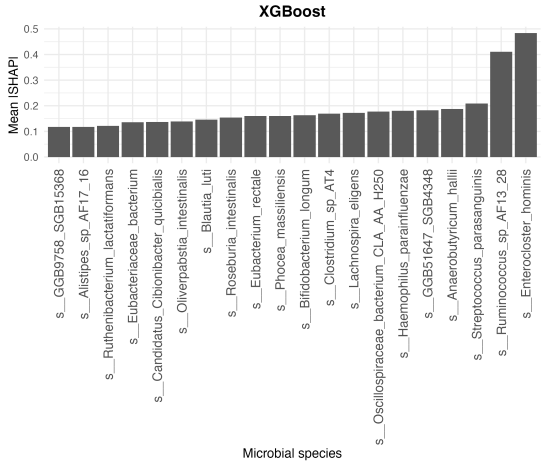
These results suggest that while different models partially converge on similar predictive microbial features, substantial variability remains in how models attribute importance to individual taxa. This highlights the sensitivity of interpretability analyses to model choice and underlines the importance of cautious biological interpretation when relying on feature importance from a single modeling approach.

4.3 Model interpretability

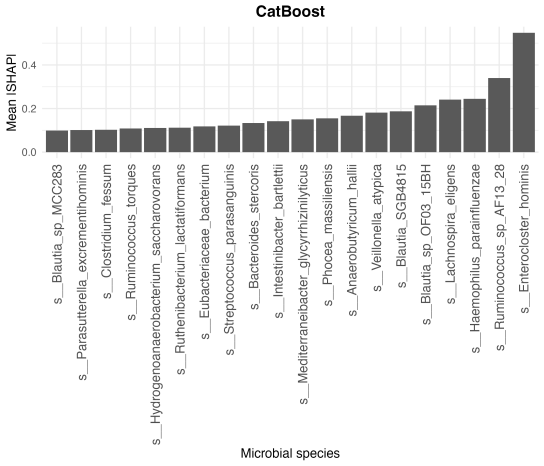
To investigate the biological interpretability of the trained machine learning models, SHAP (SHapley Additive exPlanations) values were computed for the microbiome-only models. SHAP provides a unified, model-agnostic framework for quantifying the contribution of individual features to model predictions, enabling direct comparison of feature importance across different modeling approaches.

Figure 4.6 presents the top 20 most influential microbial species according to mean absolute SHAP values for XGBoost, CatBoost, and TabPFN. Across all three models, a substantial overlap in highly ranked taxa was observed, suggesting that

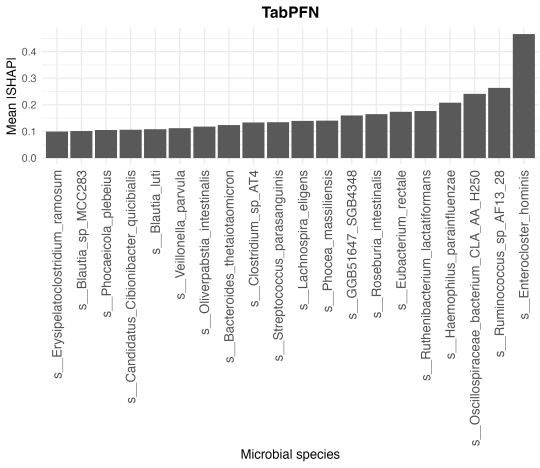
the models converged on similar biological signals despite differences in underlying learning mechanisms. Notably, several taxa consistently appeared among the most important features, including Enterocloster hominis, Ruminococcus sp. AF13_28, Haemophilus parainfluenzae, Lachnospira eligens, Anaerobutyricum hallii, and Oscillospiraceae bacterium CLA_AA_H250.



(a) XGBoost



(b) CatBoost



(c) TabPFN

Figure 4.6: Top 20 microbial species ranked by mean absolute SHAP values for XGBoost, CatBoost, and TabPFN. Higher SHAP values indicate stronger contributions of microbial taxa to BMI prediction.

The agreement between models is further supported by the moderate-to-high rank correlations observed between SHAP feature importance profiles, particularly between XGBoost and TabPFN. This indicates that although TabPFN is a transformer-

based foundation model trained on synthetic tabular datasets, it identifies many of the same microbiome features as being predictive of BMI as the tree-based ensemble methods. This convergence strengthens confidence that the observed feature importance patterns reflect genuine structure in the data rather than model-specific artifacts.

Several of the top-ranked taxa have previously been implicated in host metabolic processes and gut microbial ecology. For example, Anaerobutyricum hallii, Roseburia intestinalis, and Eubacterium rectale are known short-chain fatty acid (SCFA) producers and have been associated with host energy metabolism and insulin sensitivity [44] [45]. Their prominence among the most influential features suggests that metabolic functions of the gut microbiome contribute meaningfully to BMI-related variation captured by the models. Conversely, taxa such as Haemophilus parainfluenzae and Streptococcus parasanguinis, which are more commonly associated with oral or upper respiratory [46], may reflect indirect or lifestyle-related associations with BMI rather than direct metabolic effects.

Despite these biologically interpretable patterns, individual SHAP effect sizes were generally modest, being consistent with the descriptive analyses indicating that BMI is associated with subtle shifts in microbiome composition rather than large, easily separable community-level differences. This highlights an important challenge in microbiome-based prediction that relevant biological signals are distributed across many features with small individual effects, underscoring the need for multivariate modeling approaches.

Overall, the SHAP-based interpretability analysis demonstrates that different machine learning models, including the transformer-based TabPFN, learn broadly similar feature importance structures from microbiome data. This convergence enhances the credibility of the predictive results and supports the biological plausibility of the learned models. At the same time, the diffuse nature of feature contributions

emphasizes that BMI prediction from microbiome composition remains a complex, high-dimensional task in which no single microbial species dominates the predictive signal.

5 Conclusion

5.1 Summary of key findings

This thesis investigated the feasibility of predicting body mass index (BMI) from gut microbiome composition using a range of machine learning models, including linear regression with regularization (Lasso), tree-based ensemble methods (XGBoost and CatBoost), and a transformer-based foundation model for tabular data (TabPFN). Across all experiments, predictive performance remained moderate, highlighting the inherent difficulty of inferring a complex host phenotype such as BMI solely from microbial composition.

Among the evaluated models, TabPFN achieved the highest predictive accuracy on the full microbiome dataset, outperforming classical machine learning baselines in terms of R^2, RMSE, and MAE. Learning curve analyses further demonstrated that TabPFN benefits more strongly from larger training sets, whereas tree-based models were comparatively more data-efficient in small-sample regimes. These findings suggest that foundation models for tabular data may offer performance advantages in large-scale microbiome prediction tasks.

In addition, experiments comparing microbiome-only, metadata-only, and combined feature sets showed that gut microbiome composition provides substantially more predictive signal for BMI than basic demographic variables such as age and sex alone. The best performance was achieved when combining microbiome features

with host metadata, indicating that microbiome-based prediction models benefit from integrating host-level covariates.

Taken together, these results suggest that while microbiome composition contains measurable signal related to BMI, its predictive power remains limited when used in isolation. In practical applications, microbiome-based prediction is therefore unlikely to replace traditional clinical risk factors but may serve as a complementary source of biological information within integrated metabolic health models.

5.2 Methodological implications

The results of this study have several methodological implications for the application of machine learning in microbiome research. First, the modest predictive performance across all models underscores the challenge of predicting complex host phenotypes from compositional microbiome data. While machine learning models can extract non-trivial signal from microbiome profiles, a large proportion of the variance in BMI remains unexplained, suggesting that microbiome composition alone is insufficient to fully capture the multifactorial determinants of body weight.

Second, the comparison between model families highlights the importance of considering both model capacity and data regime. Tree-based ensemble methods such as XGBoost and CatBoost offer strong performance with relatively small training datasets and limited hyperparameter tuning, making them practical choices for typical microbiome studies with limited sample sizes. In contrast, TabPFN exhibited superior performance only when sufficient training data were available, reflecting the need for adequate data volume to leverage the inductive biases encoded in large pre-trained models.

Third, the learning curve analysis demonstrates the importance of dataset size as a key experimental variable. Performance improvements were not linear with respect to sample size, and meaningful gains were observed only beyond several

hundred or thousands of samples. This has practical implications for study design in microbiome research, suggesting that small cohorts may be insufficient for robust predictive modeling and that multi-cohort data integration may be necessary to achieve stable and generalizable results.

Although this study focused specifically on BMI, similar methodological considerations are likely to apply to other microbiome-associated host phenotypes, such as insulin resistance, inflammation markers, or cardiovascular risk factors. However, the predictive strength of microbiome features may vary substantially depending on the biological phenotype being modeled.

5.3 Biological interpretation of microbiome features

Several microbial species consistently emerged among the most predictive features across different models, including taxa belonging to the genera *Roseburia*, *Eubacterium*, *Blautia*, and *Ruminococcus*. Many of these taxa are known short-chain fatty acid (SCFA) producers and have been previously associated with host metabolic health and obesity-related phenotypes in the literature.

The presence of such taxa among the top-ranked predictive features provides biological plausibility for some of the modeling results. SCFAs, such as butyrate and acetate, play key roles in host energy metabolism, gut barrier integrity, and immune regulation, all of which are linked to metabolic health. However, the directionality and causal interpretation of these associations remain unclear. The observed relationships may reflect downstream consequences of obesity-related dietary patterns rather than direct causal effects of microbial taxa on BMI.

Importantly, the moderate predictive performance observed in this study suggests that BMI is influenced by a complex interplay of genetic, environmental, dietary, and lifestyle factors, of which the gut microbiome represents only one component. Therefore, while microbiome-based models may capture biologically mean-

ingful signals, they should be interpreted as complementary rather than standalone predictors of metabolic phenotypes.

5.4 Model interpretability and stability

Model interpretability analyses based on SHAP values revealed partial agreement in feature importance rankings across models, particularly between XGBoost and TabPFN. However, substantial variability was observed in how different models attributed importance to individual microbial features. This variability highlights a general challenge in applying explainable machine learning methods to high-dimensional, correlated biological data: feature importance estimates may be sensitive to model architecture, training dynamics, and feature collinearity.

The limited stability of feature importance rankings across models and dataset sizes cautions against overinterpretation of individual microbial taxa as definitive biomarkers of BMI. Instead, the interpretability results should be viewed as hypothesis-generating rather than confirmatory. From a methodological perspective, these findings underscore the importance of triangulating interpretability analyses across multiple models and validation schemes when drawing biological conclusions from machine learning outputs.

5.5 Limitations

This study has several limitations. First, the analysis was restricted to species-level taxonomic profiles and did not incorporate functional information, such as metabolic pathways or gene families, which may provide more direct links to host metabolism. Second, dietary information and other relevant host-level covariates were not available, limiting the ability to disentangle microbiome effects from confounding lifestyle factors.

Third, although the dataset used in this study was large relative to typical microbiome studies, it remains observational in nature. As such, the models capture associations rather than causal relationships. Furthermore, potential batch effects and cohort-specific biases may influence both microbiome composition and BMI, despite the use of standardized preprocessing pipelines.

Finally, the interpretability analyses relied on SHAP values, which, while widely used, are approximations that may not fully capture the complex, non-linear dependencies present in high-dimensional microbiome data. Additionally, microbiome sequencing data are inherently compositional and subject to technical variability, which may introduce statistical artifacts and complicate the interpretation of predictive models.

5.6 Future directions

Future work could extend this study in several directions. Incorporating functional microbiome features, such as pathway abundances or metabolomic profiles, may improve predictive performance and provide more direct biological interpretability. Additionally, integrating dietary, genetic, and lifestyle data could enable more comprehensive models of host metabolic phenotypes.

From a methodological perspective, further investigation of foundation models for tabular data in microbiome applications is warranted, particularly in multi-task or transfer learning settings where models are trained across related phenotypes. Exploring model ensembling strategies and uncertainty quantification may also yield more robust predictive frameworks.

Another promising direction is the development of foundation models pretrained specifically on biologically realistic microbiome datasets, which may better capture the statistical properties of microbial communities than current synthetic pretraining schemes.

Finally, prospective and longitudinal datasets could enable the development of predictive models that capture temporal dynamics of the microbiome and their relationship to changes in BMI, moving beyond static cross-sectional prediction toward more clinically actionable modeling paradigms.

References

[1] J. A. Gilbert, M. J. Blaser, J. G. Caporaso, J. K. Jansson, S. V. Lynch, and R. Knight, "Current understanding of the human microbiome", *Nature Medicine*, vol. 24, no. 4, pp. 392–400, 2018. DOI: 10.1038/nm.4517.

[2] R. Knight et al., "Best practices for analysing microbiomes", *Nature Reviews Microbiology*, vol. 16, no. 7, pp. 410–422, 2018. DOI: 10.1038/s41579-018-0029-9.

[3] K. Hou et al., "Microbiota in health and diseases", *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, p. 135, Apr. 23, 2022, ISSN: 2059-3635. DOI: 10.1038/s41392-022-00974-4.

[4] Y. Gao, D. Li, and Y.-X. Liu, "Microbiome research outlook: Past, present, and future", *Protein & Cell*, vol. 14, no. 10, pp. 709–712, May 23, 2023, ISSN: 1674-800X. DOI: 10.1093/procel/pwad031.

[5] W. M. de Vos, H. Tilg, M. V. Hul, and P. D. Cani, "Gut microbiome and health: Mechanistic insights", *Gut*, vol. 71, no. 5, pp. 1020–1032, May 1, 2022, Publisher: BMJ Publishing Group Section: Recent advances in basic science, ISSN: 0017-5749, 1468-3288. DOI: 10.1136/gutjnl-2021-326789.

[6] T. Wen et al., "Ggclusternet: An r package for microbiome network analysis and modularity-based multiple network layouts", *iMeta*, vol. 1, no. 3, e32, Sep. 2022, ISSN: 2770-596X. DOI: 10.1002/imt2.32.

[7] T. Chen, H. Zhang, Y. Liu, Y.-X. Liu, and L. Huang, "Evenn: Easy to create repeatable and editable venn diagrams and venn networks online", *Journal of Genetics and Genomics*, vol. 48, no. 9, pp. 863–866, Sep. 2021, ISSN: 16738527. DOI: 10.1016/j.jgg.2021.07.007.

[8] I. Isali, T. R. Wong, and S. Tian, "Best practice guidelines for collecting microbiome samples in research studies", *European Urology Focus*, vol. 10, no. 6, pp. 909–913, Dec. 2024, ISSN: 24054569. DOI: 10.1016/j.euf.2024.12.007.

[9] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis", *Nature Biotechnology*, vol. 35, no. 9, pp. 833–844, 2017. DOI: 10.1038/nbt.3935.

[10] A. Fasolo, S. Deb, P. Stevanato, G. Concheri, and A. Squartini, "Asv vs otus clustering: Effects on alpha, beta, and gamma diversities in microbiome metabarcoding studies", *PLOS ONE*, vol. 19, no. 10, F. Vita, Ed., e0309065, Oct. 3, 2024, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0309065.

[11] B. Kumar, E. Lorusso, B. Fosso, and G. Pesole, "A comprehensive overview of microbiome data in the light of machine learning applications: Categorization, accessibility, and future directions", *Frontiers in Microbiology*, vol. 15, p. 1 343 572, Feb. 13, 2024, ISSN: 1664-302X. DOI: 10.3389/fmicb.2024.1343572.

[12] M. Greenacre, M. Martínez-Álvaro, and A. Blasco, "Compositional data analysis of microbiome and any-omics datasets: A validation of the additive logratio transformation", *Frontiers in Microbiology*, vol. 12, Oct. 11, 2021, Publisher: Frontiers, ISSN: 1664-302X. DOI: 10.3389/fmicb.2021.727398.

[13] C. B. Peterson, S. Saha, and K.-A. Do, "Analysis of microbiome data", *Annual review of statistics and its application*, vol. 11, no. 1, pp. 483–504, Apr. 2024, ISSN: 2326-8298. DOI: 10.1146/annurev-statistics-040522-120734.

[14] D. Vandeputte et al., "Temporal variability in quantitative human gut microbiome profiles and implications for clinical research", *Nature Communications*, vol. 12, no. 1, p. 6740, Nov. 18, 2021, ISSN: 2041-1723. DOI: 10.1038/s41467-021-27098-7.

[15] P. Przymus et al., "Deep learning in microbiome analysis: A comprehensive review of neural network models", *Frontiers in Microbiology*, vol. 15, p. 1 516 667, Jan. 22, 2025, ISSN: 1664-302X. DOI: 10.3389/fmicb.2024.1516667.

[16] M. Afzaal et al., "Human gut microbiota in health and disease: Unveiling the relationship", *Frontiers in Microbiology*, vol. 13, Sep. 26, 2022, Publisher: Frontiers, ISSN: 1664-302X. DOI: 10.3389/fmicb.2022.999001.

[17] I. Creus-Martí, A. Moya, and F. J. Santonja, "Methodology for microbiome data analysis: An overview", *Computers in Biology and Medicine*, vol. 192, p. 110 157, Jun. 2025, ISSN: 00104825. DOI: 10.1016/j.compbiomed.2025.110157.

[18] H. Amit. "Explaining sparse datasets with practical examples", We Talk Data, Accessed: Oct. 22, 2025. [Online]. Available: https://medium.com/we-talk-data/explaining-sparse-datasets-with-practical-examples-dead60c2c3b7.

[19] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: And this is not optional", *Frontiers in microbiology*, vol. 8, p. 2224, 2017. DOI: 10.3389/fmicb.2017.02224.

[20] A. Hinton. "A guide for data scientists: Log-ratio transformations in machine learning", Medium, Accessed: Oct. 23, 2025. [Online]. Available: https://medium.com/@nextgendatascientist/a-guide-for-data-scientists-log-ratio-transformations-in-machine-learning-a2db44e2a455.

[21] C. Martino, L. Shenhav, C. Marotz, L. R. Thompson, and R. Knight, "High-dimensional microbiome data analysis: Challenges and perspectives", *BMC Bioinformatics*, vol. 21, no. 1, p. 343, 2020. DOI: 10.1186/s12859-020-3530-x.

[22] H. Xu, Y. Chen, and J. Li, "Sparse compositional data modeling for microbiome analysis: A unified zero-inflated framework", *BMC Bioinformatics*, vol. 26, no. 1, p. 78, 2025. DOI: 10.1186/s12859-025-06078-4.

[23] B. D. Topçuoğlu, N. A. Lesniak, M. T. Ruffin, J. Wiens, and P. D. Schloss, "A framework for effective application of machine learning to microbiome-based classification problems", *mBio*, vol. 11, no. 3, Jun. 30, 2020, ISSN: 2161-2129, 2150-7511. DOI: 10.1128/mBio.00434-20.

[24] J. Wirbel et al., "Microbiome meta-analysis and cross-disease comparison enabled by the siamcat machine learning toolbox", *Genome Biology*, vol. 22, no. 1, p. 93, Dec. 2021, ISSN: 1474-760X. DOI: 10.1186/s13059-021-02306-1.

[25] D. Rothschild et al., "Environment dominates over host genetics in shaping human gut microbiota", *Nature*, vol. 555, no. 7695, pp. 210–215, Mar. 2018, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature25973.

[26] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey", 2021, Version Number: 3. DOI: 10.48550/ARXIV.2110.01889.

[27] H.-J. Ye, S.-Y. Liu, and W.-L. Chao, *A closer look at tabpfn v2: Understanding its strengths and extending its capabilities*, Jun. 11, 2025. DOI: 10.48550/arXiv.2502.17361. arXiv: 2502.17361[cs].

[28] Z. Bobbitt. "What is tabular data? (definition & example)", Statology, Accessed: Oct. 28, 2025. [Online]. Available: https://www.statology.org/tabular-data/.

[29] A. X. Wang, S. S. Chukova, C. R. Simpson, and B. P. Nguyen, "Challenges and opportunities of generative models on tabular data", *Applied Soft Computing*, vol. 166, p. 112 223, 2024, ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2024.112223.

[30] N. Hollmann et al., "Accurate predictions on small data with a tabular foundation model", *Nature*, vol. 637, no. 8045, pp. 319–326, Jan. 9, 2025, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-024-08328-6.

[31] R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[32] F. Emmert-Streib and M. Dehmer, "High-dimensional lasso-based regression models: Regularization, feature selection and applications in machine learning", *Machine Learning and Knowledge Extraction*, vol. 4, no. 1, pp. 1–21, 2022. DOI: 10.3390/make4010001.

[33] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 13, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754[cs].

[34] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features", in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[35] J. T. Hancock and T. M. Khoshgoftaar, "Catboost for big data: An interdisciplinary review", *Journal of Big Data*, vol. 7, no. 1, p. 94, 2020, ISSN: 2196-1115. DOI: 10.1186/s40537-020-00369-8.

[36] Z. Qiang-long, L. Shi, G. Peng, and L. Fei-shi, "High-throughput sequencing technology and its application", *Journal of Northeast Agricultural University*

(English Edition), vol. 21, no. 3, pp. 84–96, Sep. 2014, ISSN: 10068104. DOI: 10.1016/S1006-8104(14)60073-8.

[37] Y.-X. Liu et al., "A practical guide to amplicon and metagenomic analysis of microbiome data", *Protein & Cell*, vol. 12, no. 5, pp. 315–330, May 2021, ISSN: 1674-800X, 1674-8018. DOI: 10.1007/s13238-020-00724-8.

[38] M. Kuhn et al., "Metalog: Curated and harmonised contextual data for global metagenomics samples", *Nucleic Acids Research*, vol. 54, no. D1, pp. D826–D834, Oct. 2025, ISSN: 1362-4962. DOI: 10.1093/nar/gkaf1118. [Online]. Available: http://dx.doi.org/10.1093/nar/gkaf1118.

[39] A. Blanco-Miguez et al., "Extending and improving metagenomic taxonomic profiling with uncharacterized species with metaphlan 4", Aug. 2022. DOI: 10.1101/2022.08.22.504593.

[40] S. Weiss et al., "Normalization and microbial differential abundance strategies depend upon data characteristics", *Microbiome*, vol. 5, p. 27, 2017. DOI: 10.1186/s40168-017-0237-y.

[41] "Moderate and severe thinness, underweight, overweight, obesity", Accessed: Feb. 26, 2026. [Online]. Available: https://apps.who.int/nutrition/landscape/help.aspx?menu=0&helpid=420.

[42] O. Manor, A. Zhernakova, and et al., "Health and disease markers correlate with gut microbiome composition across thousands of people", *Nature Communications*, vol. 11, 2020. DOI: 10.1038/s41467-020-18871-1.

[43] E. Ma, W. Xu, and et al., "Dietary fiber-fermenting bacteria of the phylum firmicutes such as lachnospira and ruminococcus metabolize dietary plant polysaccharides to short-chain fatty acids that serve as energy for intestinal cells", *Journal of Nutritional Science and Metabolism*, vol. 9, no. 6, pp. 162–172, 2021. DOI: 10.11648/j.ajls.20210906.12.

[44] S. D. Udayappan et al., "Oral treatment with eubacterium hallii improves insulin sensitivity in db/db mice", *NPJ Biofilms and Microbiomes*, vol. 2, p. 16 009, 2016. DOI: 10.1038/npjbiofilms.2016.9.

[45] P. Louis and H. J. Flint, "Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine", *FEMS Microbiology Letters*, vol. 294, no. 1, pp. 1–8, 2009. DOI: 10.1111/j.1574-6968.2009.01514.x.

[46] P. E. Kolenbrander, R. J. Palmer, S. Periasamy, and N. S. Jakubovics, "Oral multispecies biofilm development and the key role of cell-cell distance", *Nature Reviews Microbiology*, vol. 8, no. 7, pp. 471–480, 2010. DOI: 10.1038/nrmicro2381.

Appendix A Appendix

The code for this thesis can be found on GitHub.