Generalizing Microbiome Disease Classifiers Across Studies and Contexts

Ayoub Acharki • Supervisor: Bianca Cosma • Responsible Professor: Thomas Abeel

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), Delft University of Technology

TU Delft

01. INTRODUCTION

- The human gut microbiome contains disease-associated patterns that support classification and prediction tasks across multiple diseases [1].
- Microbiome data are high-dimensional, sparse, compositional, and strongly affected by study heterogeneity and technical variation [2].
- Within-study train--test splits can overestimate performance when the real goal is transfer to unseen cohorts or disease contexts.
- This benchmark evaluates how well selected models transfer across studies, disease contexts, and feature-space choices.

02. RESEARCH OBJECTIVES

Main RQ: Which selected deep learning models generalize best across microbiome studies and contexts, and to what extent does feature-space alignment influence their external performance?

Focus:

- Compare study-to-study, leave-one-study-out, disease-context, and within-study evaluation.
- Test four feature-space strategies: intersection, union, 5% prevalence filtering, and top-100 prevalence filtering.
- Benchmark Random Forest against the foundation-style models TabPFN and MGM.

03. FEATURE ALIGNMENT

Intersection

Keep only taxa shared across selected studies or disease contexts. This reduces feature-space mismatch, but may remove informative taxa that appear in only part of the data [3].

Union-based strategies

Union all: retain all observed taxa and fill missing features with zero.

Union + prevalence 5%: keep taxa present in at least 5% of the samples.

Union + top-100 prevalence: keep the 100 most prevalent taxa in the samples.

Contact: A.Acharki-1@student.tudelft.nl   Delft University of Technology

04. METHODOLOGY

GM Common Diseases dataset [4]

6309 gut metagenome samples from 36 studies and 11 disease contexts.

Species-level input: 2557 features. Genus-level input: 1179 features.

↓

Feature-Space Strategies:

Intersection • Union • 5% prevalence • Top-100 prevalence

↓

Validation Tasks

- **Task A:** train on one study, test on another
- **Task B:** leave-one-study-out validation
- **Task C:** train on one disease context, test on another
- **Task D:** within-study 80/20 train-test split

Compared Models

- **Dummy Classifier:** baseline that predicts from the training class distribution.
- **Random Forest:** classical baseline for tabular microbiome data.
- **TabPFN:** pretrained tabular foundation model applied to species-level abundance profiles.
- **MGM:** pretrained microbiome foundation model evaluated on genus-level input.

↓

Evaluation summary

ROC-AUC | Task comparison | Model comparison

REFERENCES

[1] Ibrahimi et al. Overview of data preprocessing for machine learning applications in human microbiome research. *Frontiers in Microbiology*, 2023. DOI: 10.3389/fmicb.2023.1250909.

[2] Kubinski et al. Benchmark of data processing methods and machine learning models for gut microbiome-based diagnosis. *Frontiers in Genetics*, 2022.

[3] Su et al. Method development for cross-study microbiome data mining: Challenges and opportunities. *Computational and Structural Biotechnology Journal*, 2020.

[4] GM Common Diseases dataset: github.com/yexianingyue/GM_common_diseases.

05. RESULTS

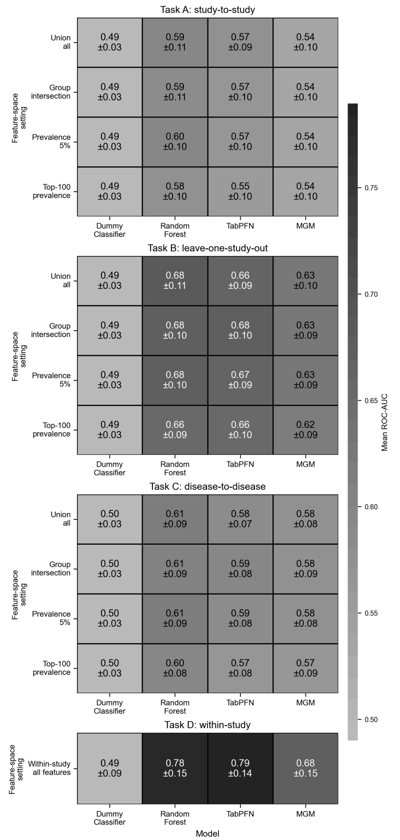

Figure 1: Mean ROC-AUC (± standard deviation) across tasks, models, and feature-space strategies. Within-study prediction performs best, while external transfer remains more difficult.

06. CONCLUSIONS

1. External validation is harder than within-study prediction, so within-study scores alone overestimate real-world usefulness.

2. Random Forest is the strongest external-transfer model overall. TabPFN is only numerically best within-study, and its advantage there is small.

3. Feature-space alignment matters, but much less than task design itself; no single strategy dominates across all settings, and the restrictive top-100 prevalence setting is often weaker.