XÂY DỰNG MÔ HÌNH DỰ ĐOÁN BỆNH VIÊM RUỘT (IBD) DỰA VÀO DỮ LIỆU 16S rDNA CỦA HỆ VI SINH ĐƯỜNG RUỘT

HỌC VIÊN BÁO CÁO: VÕ THỊ HOÀNG ANH
NGÀNH: CÔNG NGHỆ SINH HỌC

GIẢNG VIÊN HƯỚNG DẪN: TS. LƯU PHÚC LỢI

NỘI DUNG

MỞ ĐẦU

- *Tầm quan trọng của hệ vi sinh vật đối với sức khỏe*
- *Bệnh viêm ruột IBD*

MỤC TIÊU NGHIÊN CỨU

- *Xây dựng dữ liệu IBD và xây dựng mô hình dự đoán IBD*

TỔNG QUAN QUY TRÌNH

THIẾT KẾ THÍ NGHIỆM

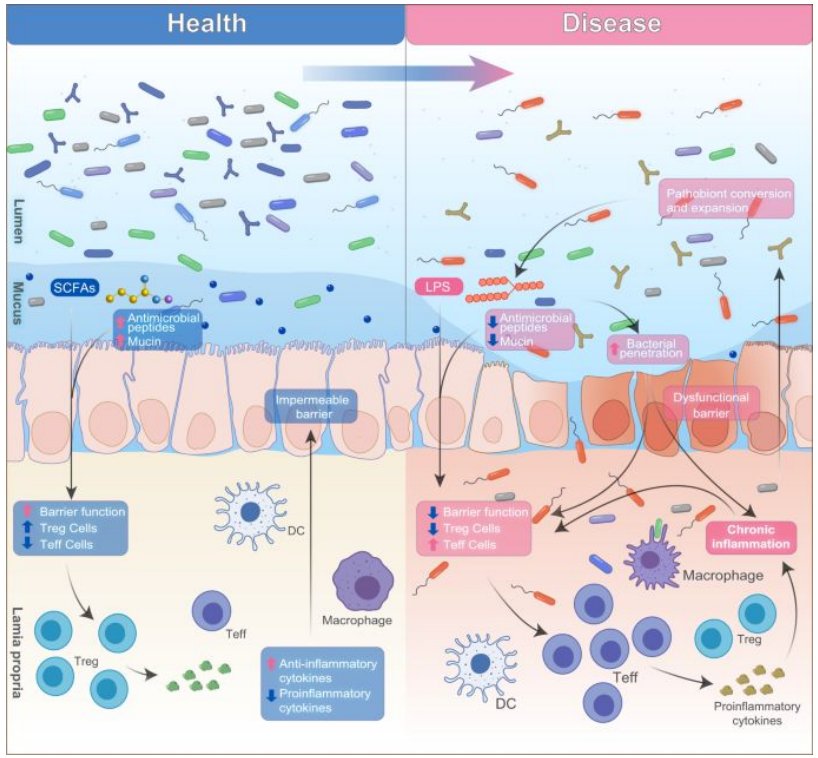
KẾT QUẢ

- *Thu thập dữ liệu microbiome*
- *Phân tích dữ liệu bằng QIIME2*
- *Mô hình học máy phân lớp IBD/HC*

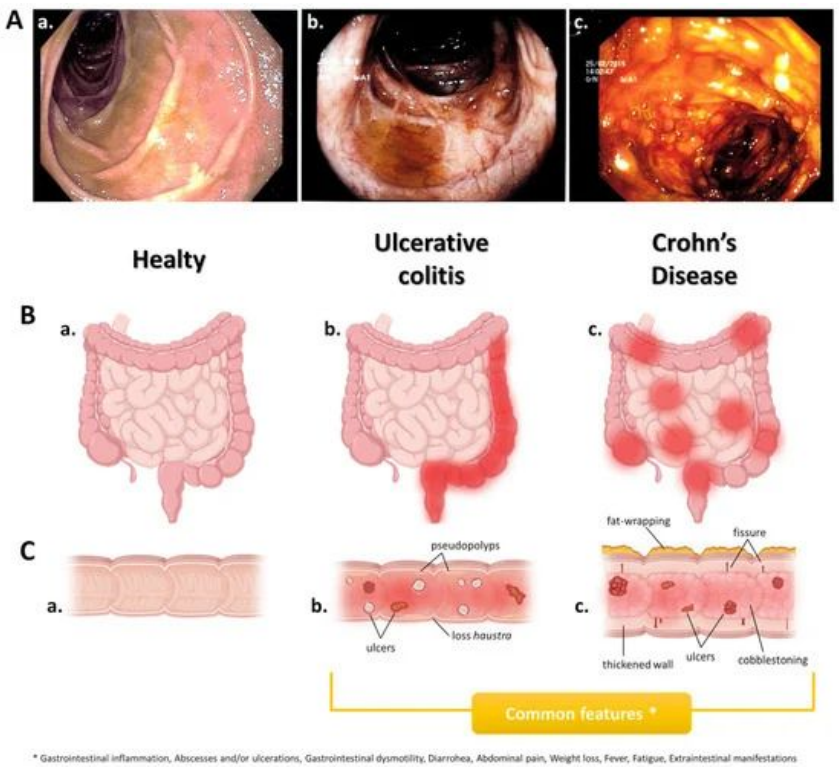
THẢO LUẬN

KẾT LUẬN

MỞ ĐẦU - BỆNH VIÊM RUỘT IBD



(Hou et al., *Sig. Transduct. Target Ther.*, 2022)



(Amodeo et al., *Biomedicines*, 2023)

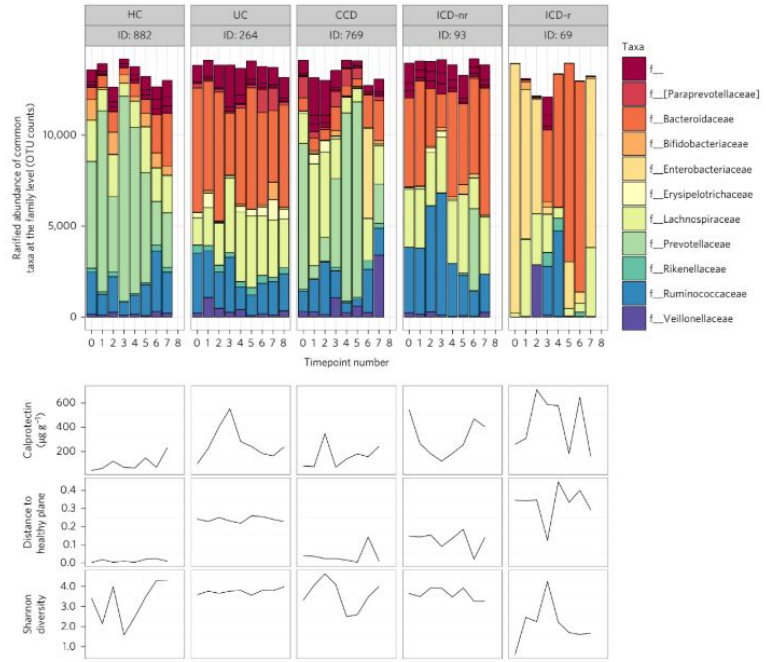


Figure 4 | Microbiome dynamics of selected individuals from each IBD subtype and a healthy control. From each IBD subtype and healthy control (HC) group, representative individuals sampled over the most time points and with complete clinical and sequence data were selected. Data represent f-calprotectin values, distance to the HP and Shannon diversity and rarified abundances of most common taxa at the family level. Note that taxa unclassified at the family level are represented in the 'f__' category.

(Halfvarson et al., *Nat. Microbiol.*, 2017)

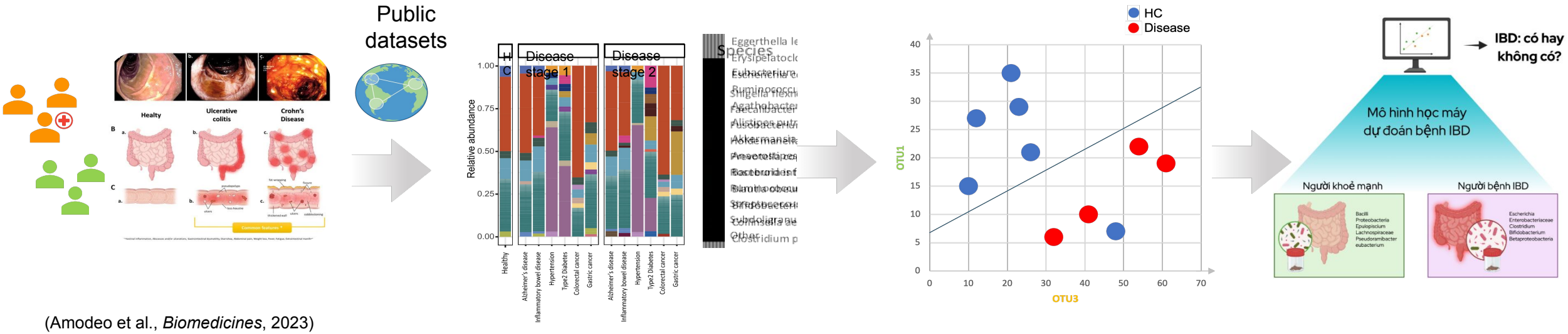
MỤC TIÊU NGHIÊN CỨU

Mục tiêu 1: Xây dựng dữ liệu hệ vi sinh đường ruột ở người khỏe mạnh và người mắc bệnh viêm ruột (IBD) từ các cơ sở dữ liệu cộng đồng.

Mục tiêu 2: So sánh đặc điểm thành phần vi sinh vật đường ruột giữa người khỏe mạnh và người mắc bệnh viêm ruột (IBD) dựa trên kết quả giải trình tự hệ vi sinh đường ruột.

Mục tiêu 3: Xây dựng mô hình AI phân lớp hệ vi sinh của người bệnh IBD và người khỏe mạnh.

TỔNG QUAN QUY TRÌNH



(Amodeo et al., *Biomedicines*, 2023)

TÌNH HÌNH NGHIÊN CỨU CỦA MÔ HÌNH ĐỀ XUẤT

✅ Đã có mô hình học máy dự đoán bệnh từ microbiome (CRC, tiểu đường, IBD...) dùng Random Forest, SVM, deep learning với dữ liệu 16S/shotgun.

⚠️ Mô hình IBD trước đây thường chỉ dùng 1 nguồn dữ liệu, thiếu kiểm tra tổng quát trên tập ngoài.

✅ Đề án này khác biệt: gộp đa dữ liệu, lọc đặc trưng bằng ANCOM-BC, cân bằng bằng SMOTE, và đánh giá trên tập xác thực độc lập.

✅ Dữ liệu 16S vùng V3–V4 phổ biến, phân loại vi sinh đến cấp chi tốt, dễ chuẩn hóa và gộp từ nhiều nguồn.

✅ Machine learning phù hợp với dữ liệu vi sinh phức tạp, giúp phát hiện mẫu ẩn và phân loại hiệu quả.

THIẾT KẾ THÍ NGHIỆM

Public datasets

📥 Danh sách sample ID (IBD / HC)

📂 Dữ liệu thô dạng .fastq từ NCBI SRA

📝Metadata (SraRunTable)

Xử lý dữ liệu bằng QIIME2

✅ Bảng đếm vi sinh vật (table.qza)]
🧬 Trình tự đại diện (rep-seqs.qza)
📈 File trực quan hóa (table.qzv, rep-seqs.qzv)

Phân tích DAA bằng ANCOM-BC

✅ Các chi vi sinh vật có mức độ khác biệt ý nghĩa
📊 Biểu đồ phân bố thành phần vi sinh vật

Gộp mẫu và xử lý tiền dữ liệu

👉 Gộp & chuẩn hóa training

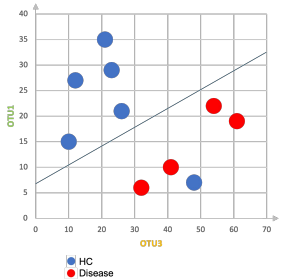
👉 Phân loại & DAA (training)

👉 Tiền xử lý test set

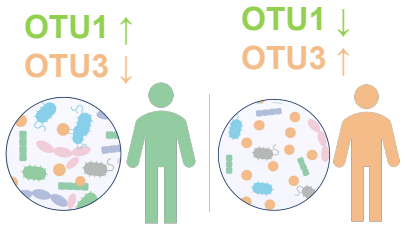
✅ Các vi sinh vật phân biệt rõ giữa nhóm bệnh / khỏe mạnh
📂training_level-6_genus.csv
(Dữ liệu vi sinh vật đã xử lý)
📂Test_level-6.csv
(Dữ liệu xác thực)

Xây dựng mô hình dự đoán (machine learning)

👉 Làm sạch dữ liệu, chuyển đổi tỉ lệ
👉 Huấn luyện mô hình Random Forest + tối ưu
👉 Đánh giá mô hình bằng tập kiểm tra và tập xác thực



MÔ HÌNH DỰ ĐOÁN

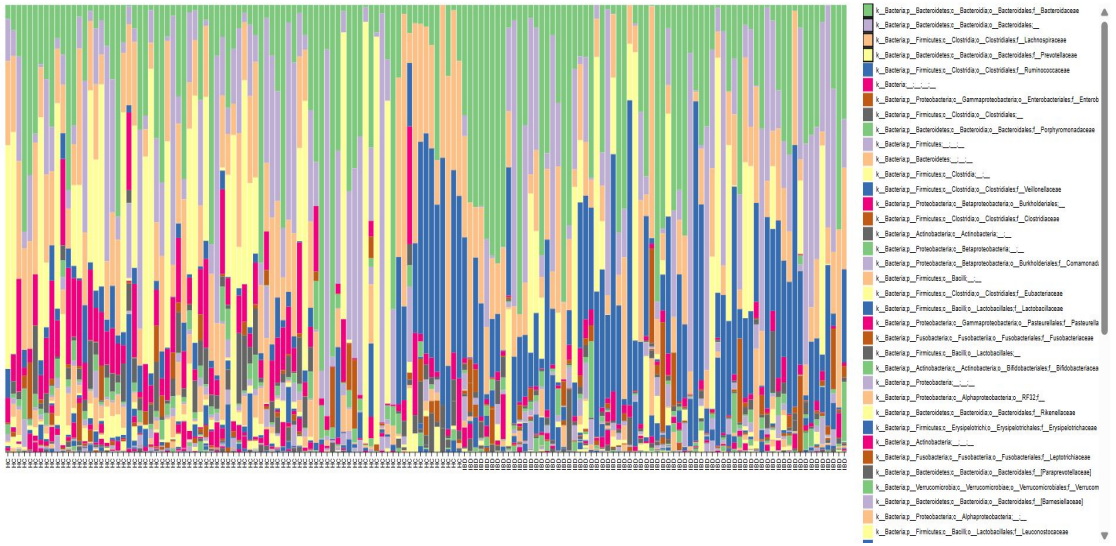


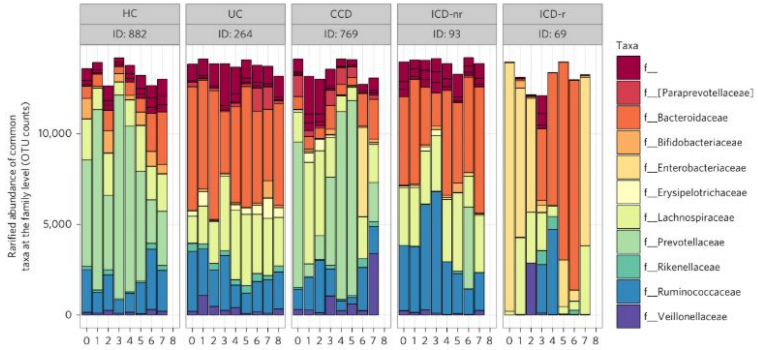
KẾT QUẢ

Bảng tổng kết mẫu và kết quả xử lý Qiime2 tới Feature Tables

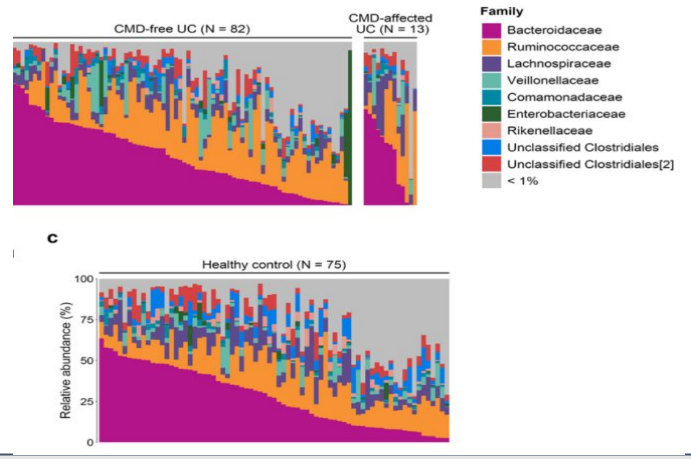
Mã dữ liệu trên SRA NCBI	Nguồn mẫu	Nguồn công bố	Số lượng mẫu		Lưu trữ dữ liệu	Bảng đặc trưng IBD
			IBD	HC		
PRJNA1101759	mẫu phân	Lee et al., *NPJ Genomic Medicine* 2024	507	75	/home/hoanganh/Gut Microbiome - IBD - Thesis/PRJNA1101759	PRJNA1101759
PRJNA855620	mẫu phân	Abdelbary et al., *Front. Cell. Infect. Microbiol.*, 2022	20	12	/home/hoanganh/Gut Microbiome - IBD - Thesis/PRJNA855620	PRJNA855620
PRJNA909073	mẫu phân	Liu et al., *Inflammatory Bowel Diseases*, 2024	87	21	/home/hoanganh/Gut Microbiome - IBD - Thesis/PRJNA909073	PRJNA909073
PRJNA978516	mẫu phân	Kim et al., *Scientific Reports* 2024	523	117	/home/hoanganh/Gut Microbiome - IBD - Thesis/PRJNA978516	PRJNA978516
Tổng mẫu = 1362			**1137**	**225**		

KẾT QUẢ





(Halfvarson et al., Nature Microbiology 2017)

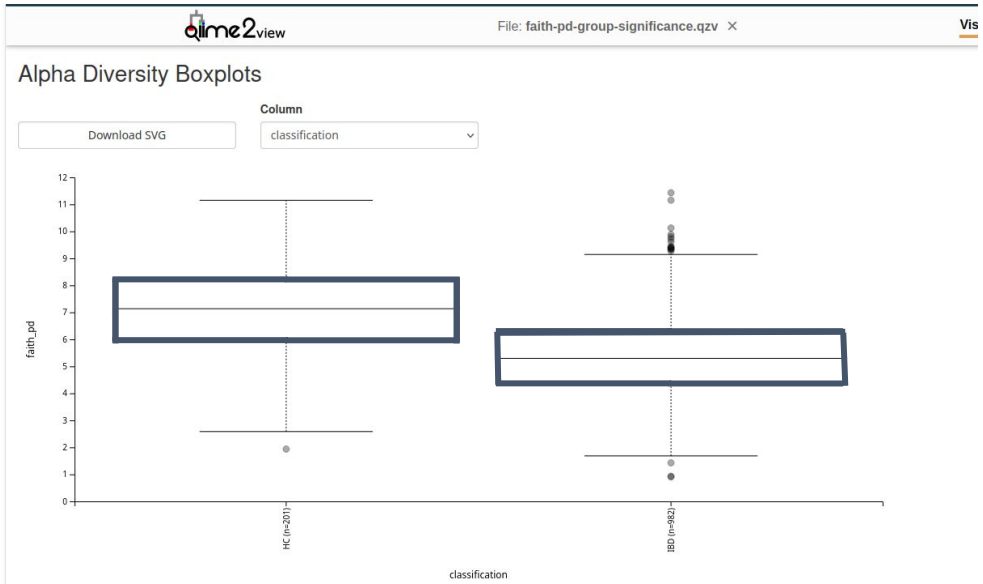


(Lee et al., NPJ Genomic Medicine 2024)

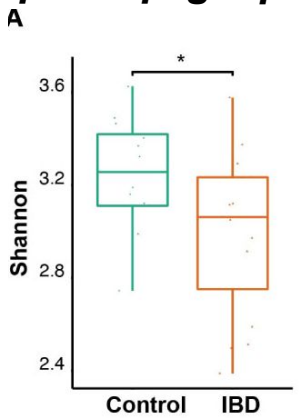
✅ Các vi sinh vật ở cấp họ (family) trong đề tài phù hợp với các nghiên cứu trước (vd. *Bacteroidaceae*, *Lachnospiraceae*...).

🔍 Để tăng độ chi tiết cho mô hình học máy, đề tài phân tích sâu đến cấp chi (genus).

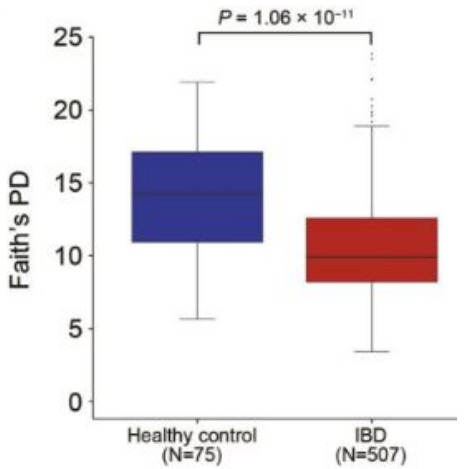
KẾT QUẢ



Biểu đồ hộp thể hiện độ đa dạng alpha



(Abdelbary et al., Front. Cell. Infect. Microbiol., 2022)



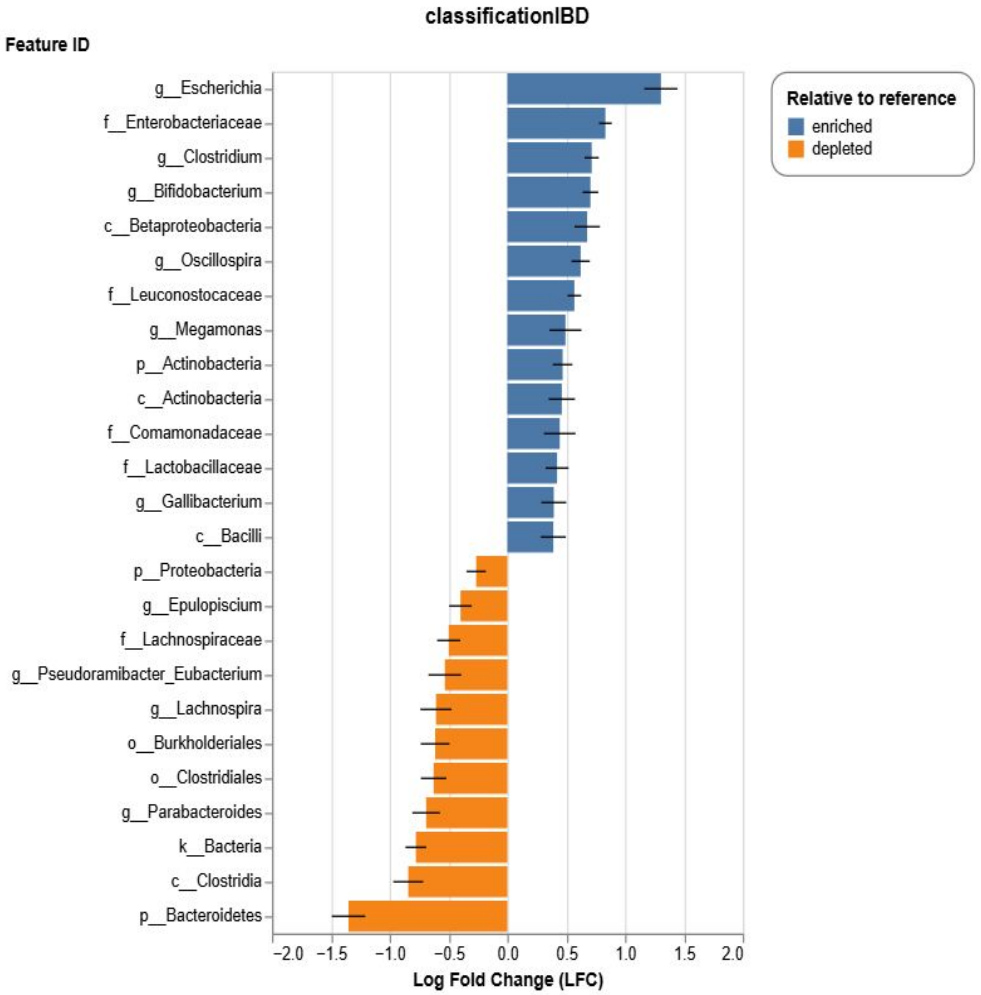
(Lee et al., NPJ Genomic Medicine 2024)

🔍 Người khỏe mạnh có hệ vi sinh vật đường ruột đa dạng hơn so với bệnh nhân IBD.

✅ Kết quả này phù hợp với nhiều nghiên cứu trước về microbiome ở bệnh IBD và nhóm chứng.

💡 Giả thuyết: Môi trường đường ruột của người IBD tăng viêm
→ Tăng vi sinh vật ưa viêm
→ Giảm vi sinh vật có lợi
→ Giảm đa dạng hệ vi sinh vật so với người khỏe mạnh.

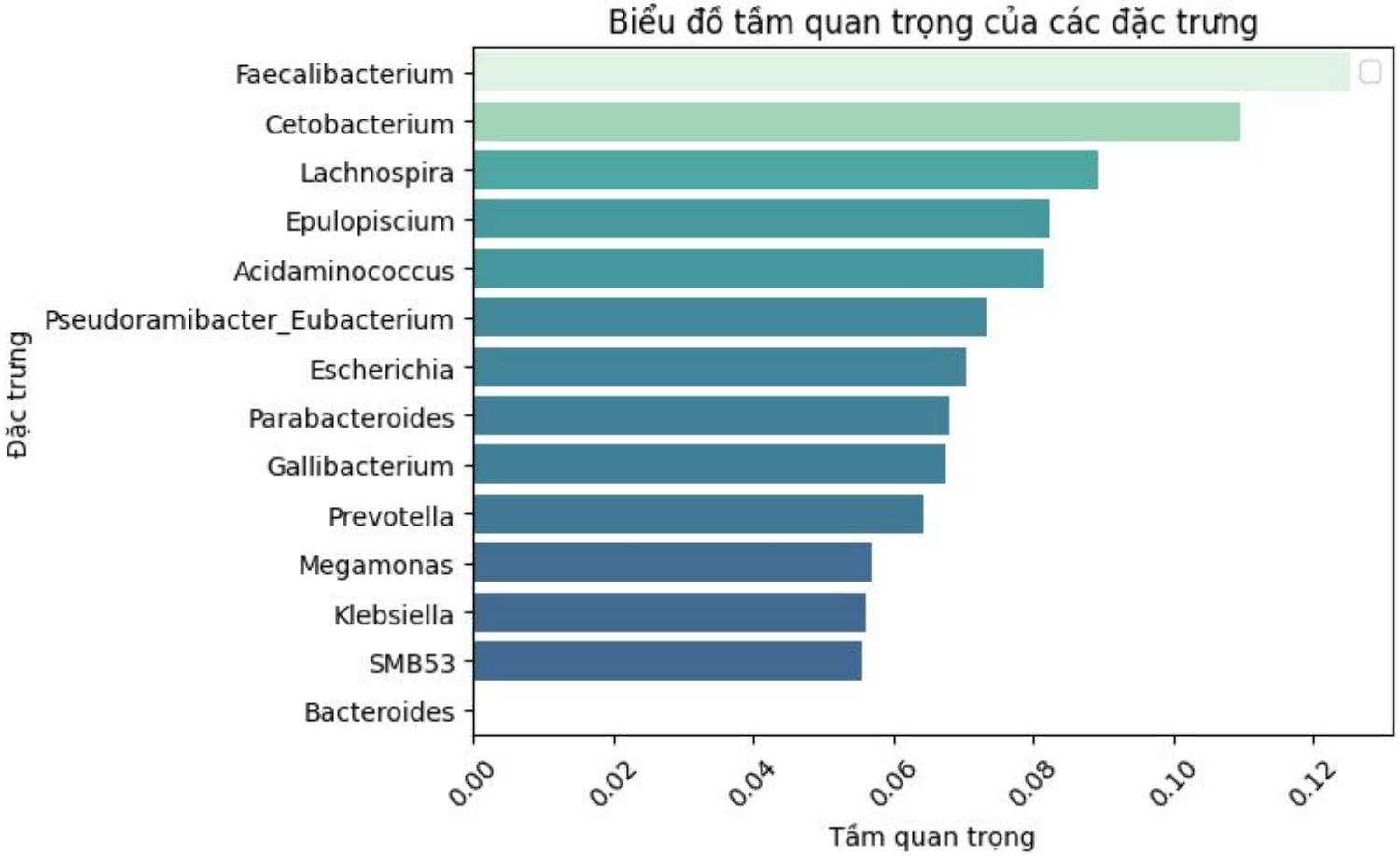
KẾT QUẢ



🔺 **Tăng vi khuẩn ưa viêm, nổi bật là *Escherichia*.**

🔻 **Giảm lợi khuẩn, đặc biệt là họ *Bacteroidetes*.**

KẾT QUẢ



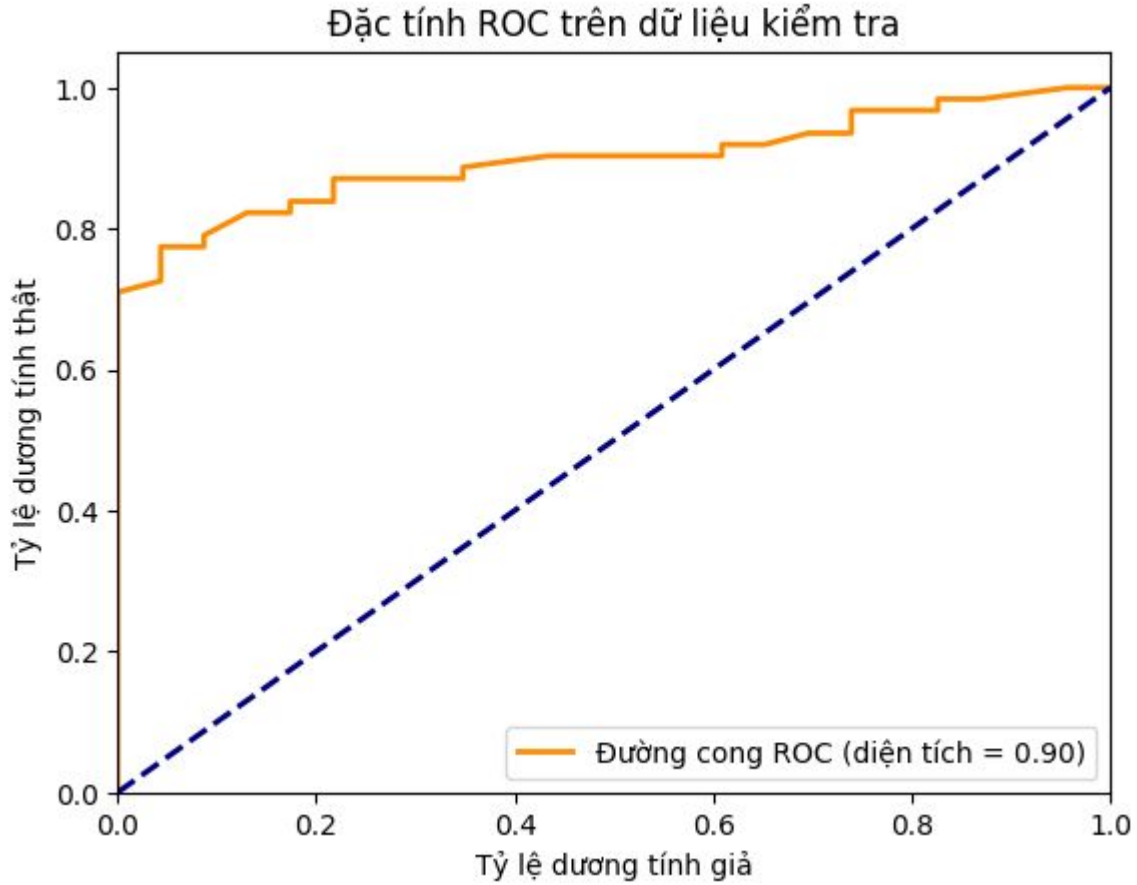
Biểu đồ thể hiện mức độ quan trọng của 14 chi vi sinh vật hàng đầu trong phân loại bệnh IBD. (Feature importance)

🔍 **Tỉ lệ Faecalibacterium/Bacteroides là đặc trưng quan trọng nhất trong mô hình.**

📉 **Bacteroides do được dùng làm chuẩn để tính tỉ lệ → luôn bằng 1 ở mọi mẫu.**

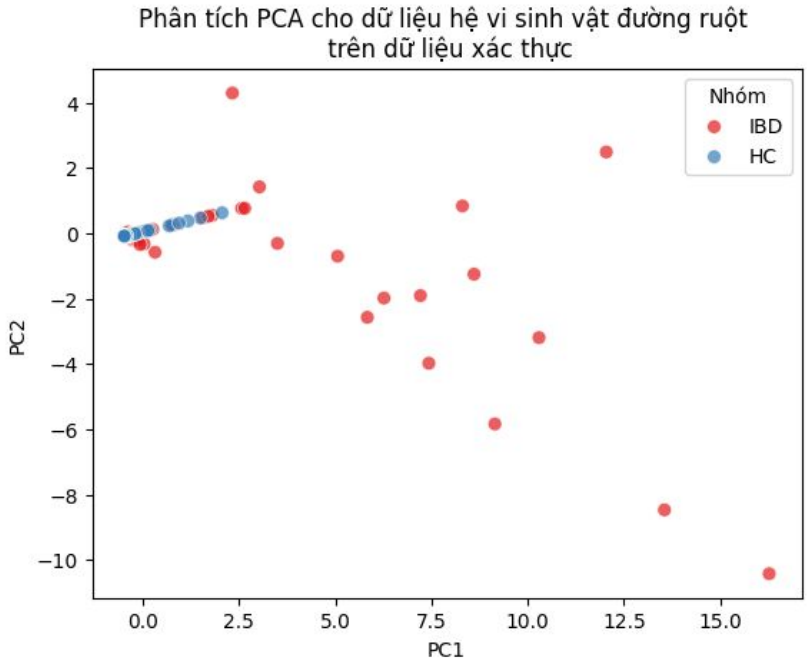
KẾT QUẢ

Sử dụng python và thư viện sklearn để huấn luyện mô hình random forest

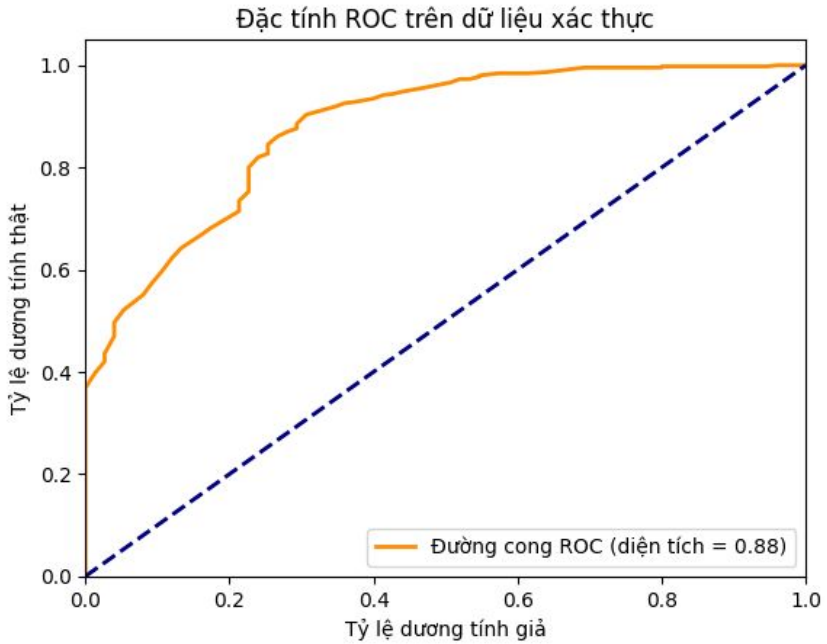


Đường cong ROC (Receiver Operating Characteristic) của mô hình Random Forest trong dự đoán bệnh viêm ruột (IBD).

KẾT QUẢ



Phân tích PCA cho dữ liệu hệ vi sinh vật đường ruột trên dữ liệu xác thực.



Đường cong ROC (Receiver Operating Characteristic) của mô hình Random Forest trong dự đoán bệnh viêm ruột (IBD).

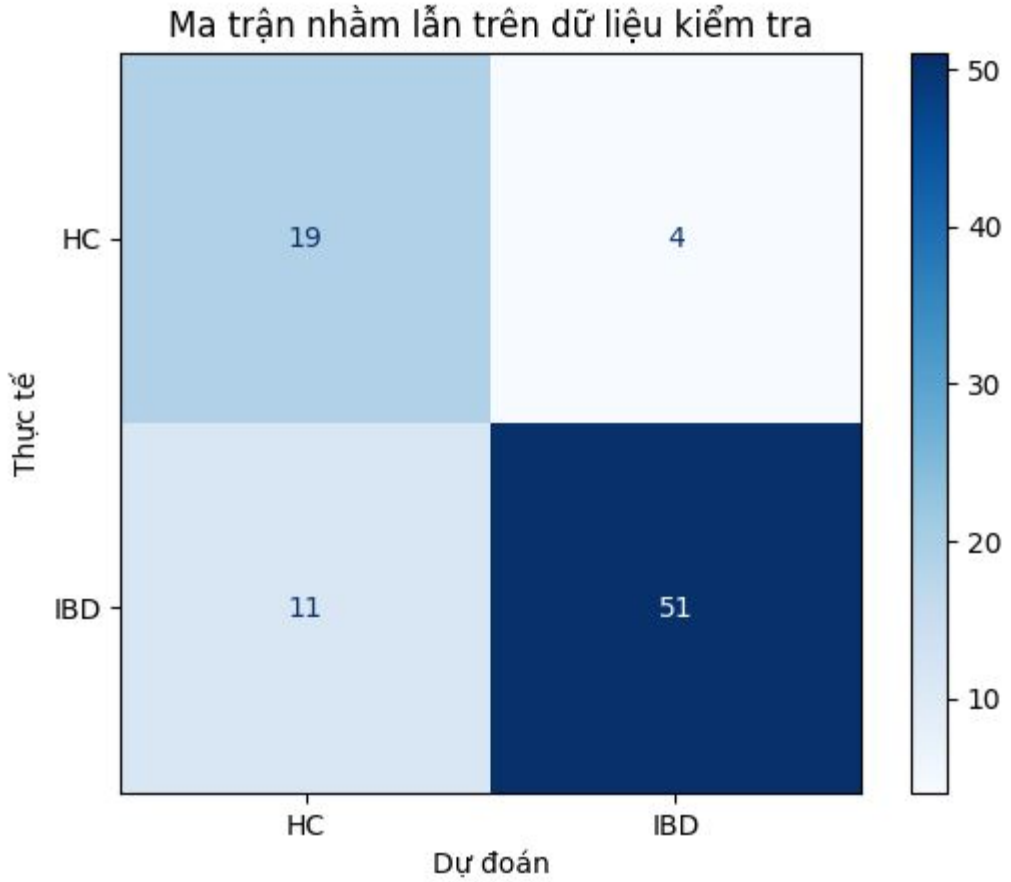
📈 PC1 và PC2 phân tách được nhóm IBD và khỏe mạnh, nhưng vẫn có một số mẫu IBD chồng lắp với nhóm khỏe.

💡 Giả thuyết: Do khác biệt về thời gian mắc bệnh, mẫu IBD mới phát hiện vẫn giữ đa dạng vi sinh tương đương người khỏe mạnh.

THẢO LUẬN

Độ chính xác: 0.8235294117647058
Báo cáo phân loại:

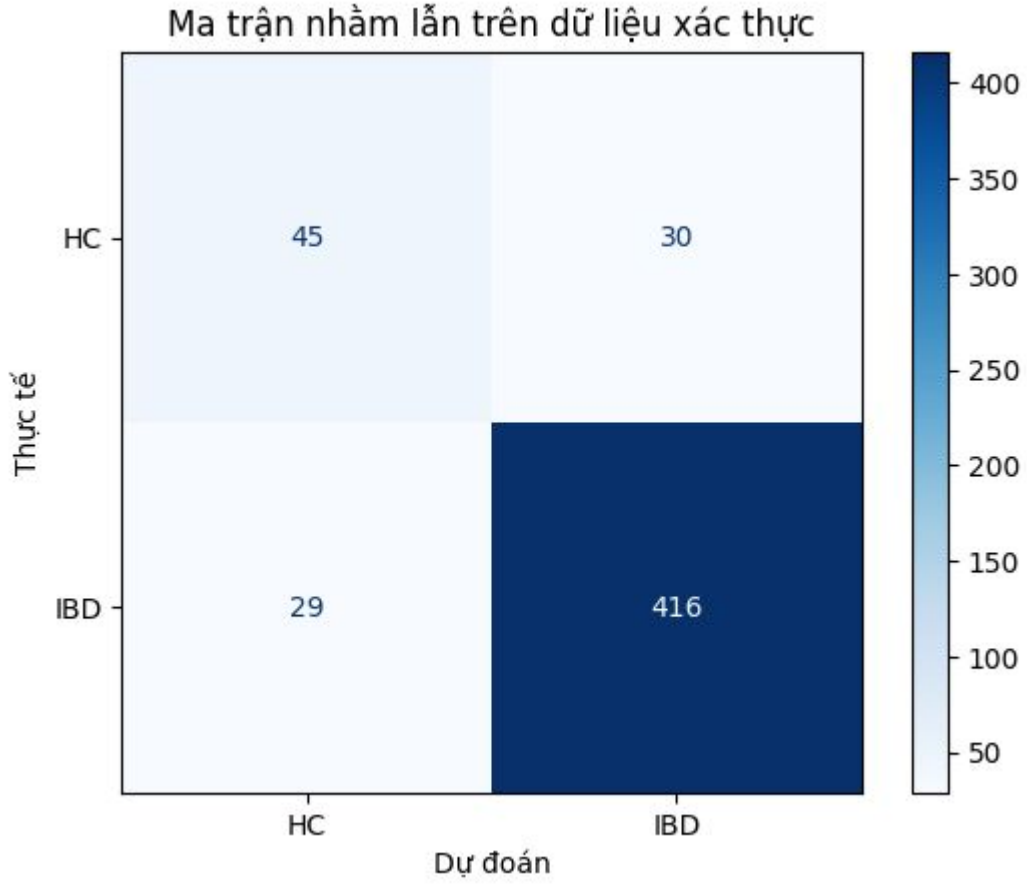
	precision	recall	f1-score	support
0	0.63	0.83	0.72	23
1	0.93	0.82	0.87	62
accuracy			0.82	85
macro avg	0.78	0.82	0.79	85
weighted avg	0.85	0.82	0.83	85



Ma trận nhầm lẫn (Confusion matrix) trên dữ liệu kiểm tra hiển thị số lượng mẫu được mô hình phân loại đúng và sai giữa hai nhóm IBD và HC.

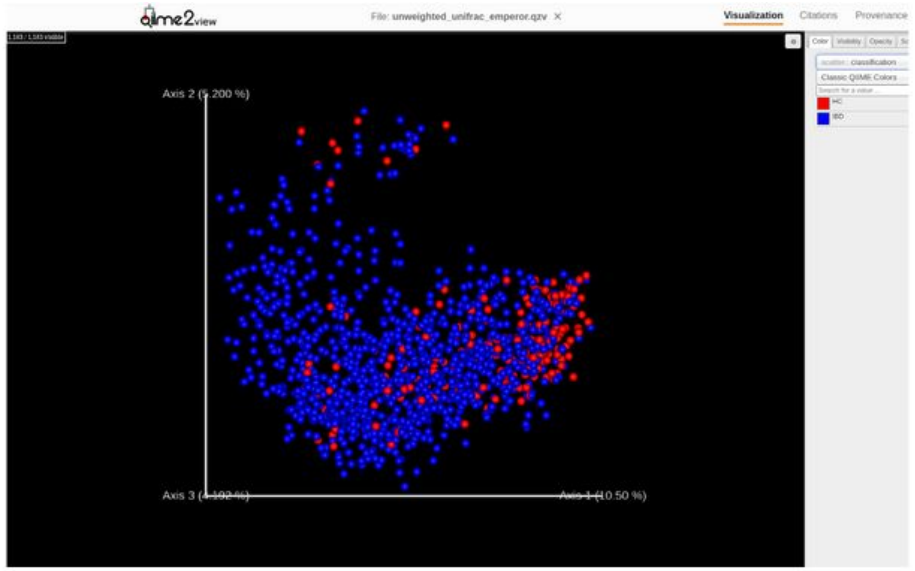
Độ chính xác: 0.8865384615384615
Báo cáo phân loại:

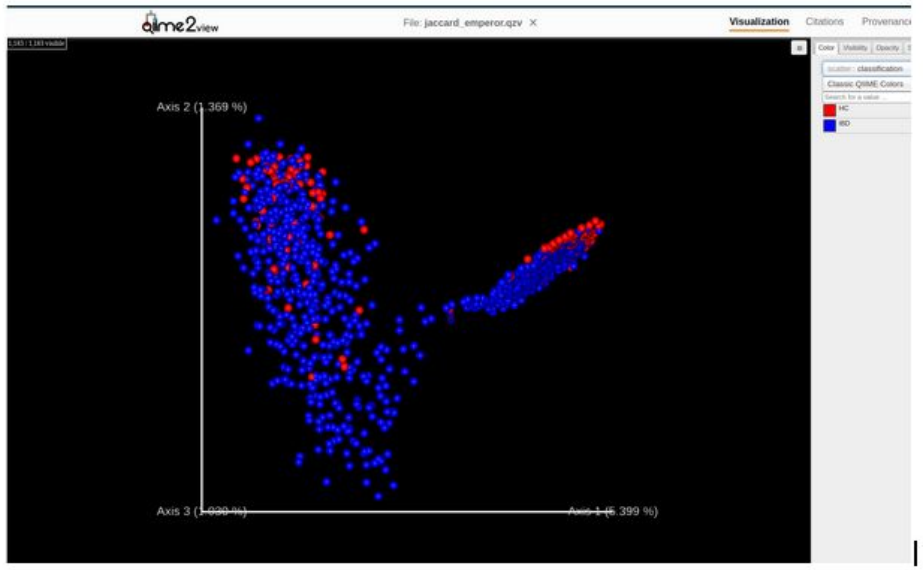
	precision	recall	f1-score	support
HC	0.61	0.60	0.60	75
IBD	0.93	0.93	0.93	445
accuracy			0.89	520
macro avg	0.77	0.77	0.77	520
weighted avg	0.89	0.89	0.89	520



Ma trận nhầm lẫn (Confusion matrix) hiển thị số lượng mẫu được mô hình phân loại đúng và sai giữa hai nhóm IBD và HC

THẢO LUẬN





📊 PCoA cho thấy một số mẫu IBD chồng lắp với nhóm khỏe mạnh, cho thấy hệ vi sinh tương đồng

→ gây khó khăn cho mô hình phân loại và dễ gây nhầm lẫn.

⚠️ Đây là giới hạn của mô hình trong vùng dữ liệu không phân tách rõ.

💡 Đề xuất:

→ Vẽ vị trí mẫu mới trên đồ thị PCoA để hỗ trợ phân loại. Nếu mẫu nằm xa cụm HC → nghi IBD, còn nằm gần HC → cần xét nghiệm bổ sung.

→ Tìm thêm dữ liệu cho nhóm HC để có sự cân bằng về mẫu khi huấn luyện mô hình.

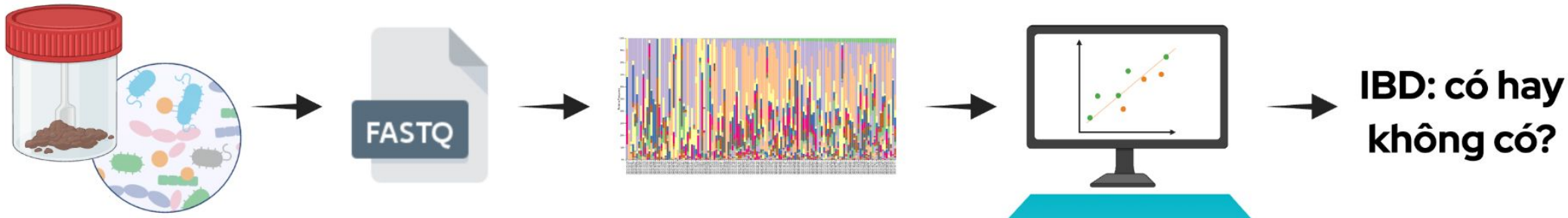
KẾT QUẢ NGHIÊN CỨU

1. Xây dựng dữ liệu hệ vi sinh đường ruột ở người khỏe mạnh và người mắc bệnh viêm ruột (IBD) từ các cơ sở dữ liệu cộng đồng (Tổng mẫu: 1362, IBD: 1137, HC: 225)

2. So sánh đặc điểm thành phần vi sinh vật đường ruột giữa HC và người mắc bệnh IBD dựa trên kết quả giải trình tự hệ vi sinh đường ruột cho thấy các vi sinh vật tăng hoặc giảm trong hệ vi sinh vật đường ruột người bệnh IBD và biểu đồ tầm quan trọng của các vi sinh vật.

3. Xây dựng mô hình AI phân lớp hệ vi sinh của người bệnh IBD và người khỏe mạnh (Đạt được mô hình phân loại Random Forest khá chính xác, với AUC = 0.9 (Area under Curve))

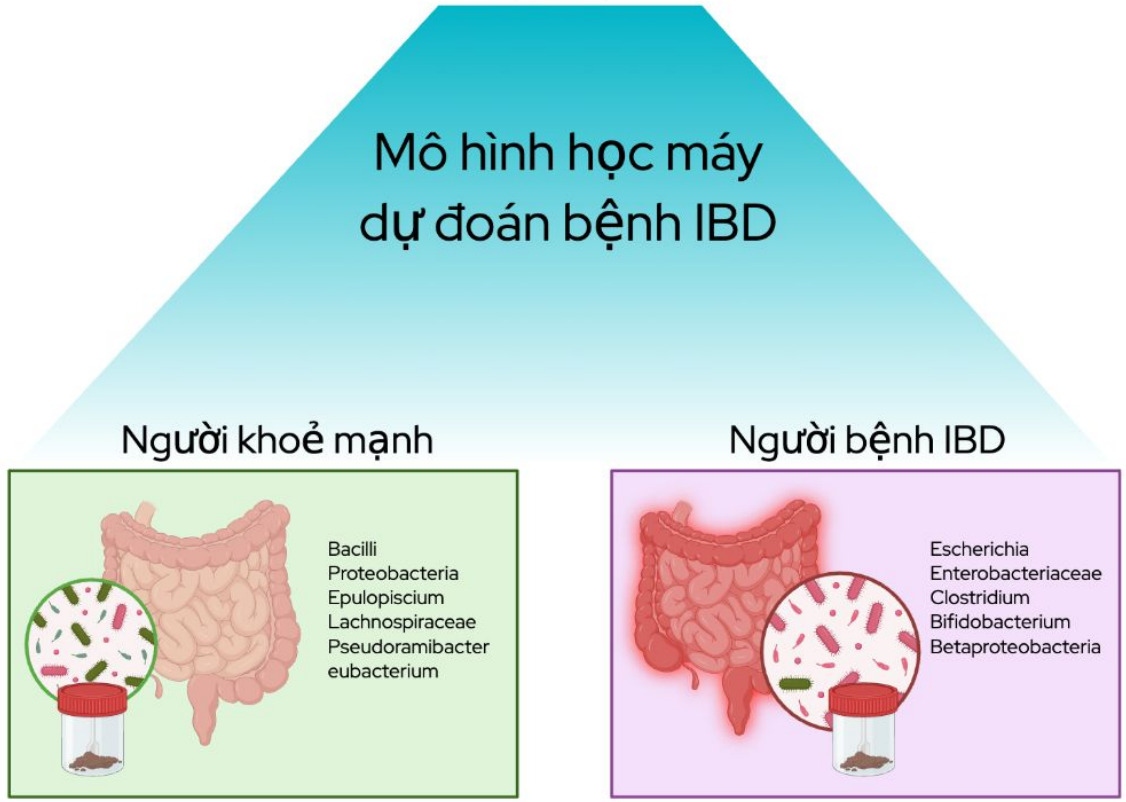
TỔNG KẾT



✅ Xây dựng được mô hình dự đoán bệnh IBD từ dữ liệu.

✅ Cần phân loại bệnh IBD sớm hay muộn để tăng độ chính xác và giảm độ nhầm lẫn của mô hình dự đoán.

✅ Tăng dữ liệu HC để cân bằng cỡ mẫu thực tế.

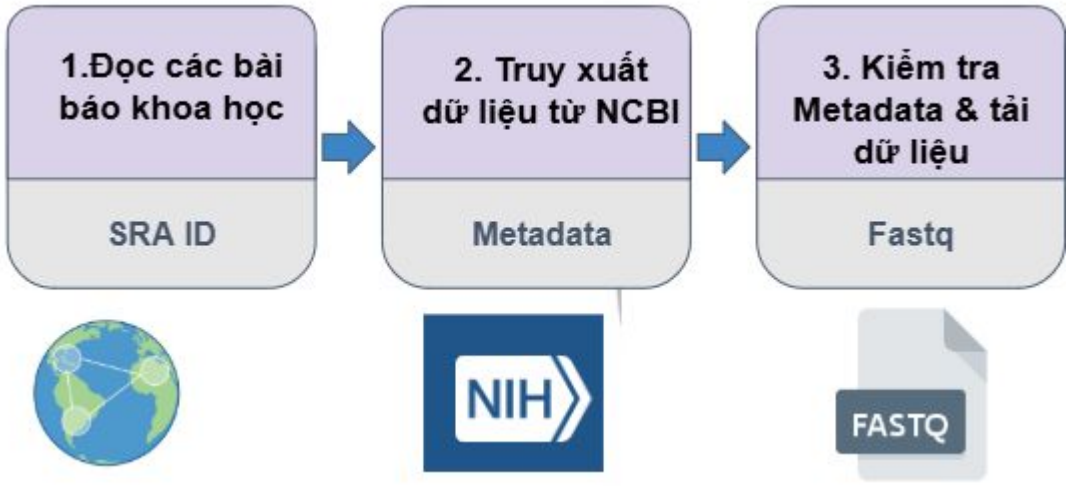


Xin chân thành cảm ơn!

CÁC SLIDE PHỤ
BỔ SUNG QUY TRÌNH CHI TIẾT
(ĐỂ DIỄN GIẢI THÊM NẾU CÓ CÂU HỎI)

III. QUY TRÌNH THỰC HIỆN ĐỀ TÀI

1. Xây dựng dữ liệu 16S rDNA hệ vi sinh vật đường ruột



BẢNG TIÊU CHÍ CHỌN MẪU BÀI BÁO VÀ DỮ LIỆU

STT	Tiêu chí	Mô tả cụ thể	Lý do / Ghi chú
1	Loại dữ liệu	16S rDNA (chấp nhận V3-V4 hoặc V4)	Phù hợp với mục tiêu phân loại vi sinh vật; V3-V4/V4 giúp xác định taxonomy tốt
2	Phân vùng 16S	V3-V4 hoặc V4	Để xây dựng bảng đặc trưng (feature table) cho mô hình học máy
3	Dạng dữ liệu	Dữ liệu thô (raw reads, fastq)	Phải có raw data để xử lý bằng QIIME2 và xây dựng pipeline
4	Kiểu dữ liệu đọc	Pair-end	
5	Đối tượng nghiên cứu	Human gut microbiome, có phân nhóm IBD vs Healthy	Phải có nhóm đối chứng và bệnh để xây dựng mô hình phân loại
6	Thời gian xuất bản bài báo	Trong vòng 5 năm gần đây (2020–2024)	Dữ liệu mới hơn, ít lỗi kỹ thuật hơn, chuẩn phân tích cập nhật hơn
7	Nguồn dữ liệu công khai	Có sẵn link tải (NCBI, EBI, Qiita...)	Cần đảm bảo tải được dữ liệu thực sự, không bị lỗi hoặc mất link
8	Loại nghiên cứu	Nghiên cứu thực nghiệm hoặc meta-analysis có cung cấp dữ liệu	Meta-analysis cũng được nếu có dataset chi tiết để tải
9	Chất lượng dữ liệu	Có kiểm tra sơ bộ: không bị lỗi định dạng, đủ metadata, phân bố nhóm đều	Đảm bảo pipeline xử lý không bị gián đoạn, dữ liệu có ý nghĩa
10	Nhóm quần thể nghiên cứu	**Tiêu chuẩn đưa vào chung:** + Đối với nhóm bệnh: được chẩn đoán một trong các bệnh lý khảo sát trong thời gian nghiên cứu. + Đối với nhóm đối chứng: sức khỏe bình thường được kết luận. **Tiêu chuẩn loại trừ chung:** + Bệnh nhân có các tổn thương cấp tính hoặc đang mắc các bệnh mãn tính khác. + Có các vấn đề về đường tiêu hóa như nôn mửa, táo bón trong vòng 1 tuần gần đây + Đã sử dụng kháng sinh, thuốc ức chế miễn dịch hoặc men vi sinh trong vòng 2 tháng gần đây + Đã sử dụng thuốc chống viêm không steroid trong thời gian dài + Tiền sử IBD + Tiền sử phẫu thuật đường ruột dẫn đến mất đoạn ruột hoặc thay đổi giải phẫu đường tiêu hóa. + Bệnh nhân từ chối tham gia nghiên cứu + Loại trừ bệnh nhân từng phẫu thuật cắt bỏ phần ruột hay phần dạ dày trước đó.	
11	Sản phẩm đầu ra có thể tạo	Có thể tạo được feature table	Để phục vụ xây dựng mô hình AI dự đoán IBD

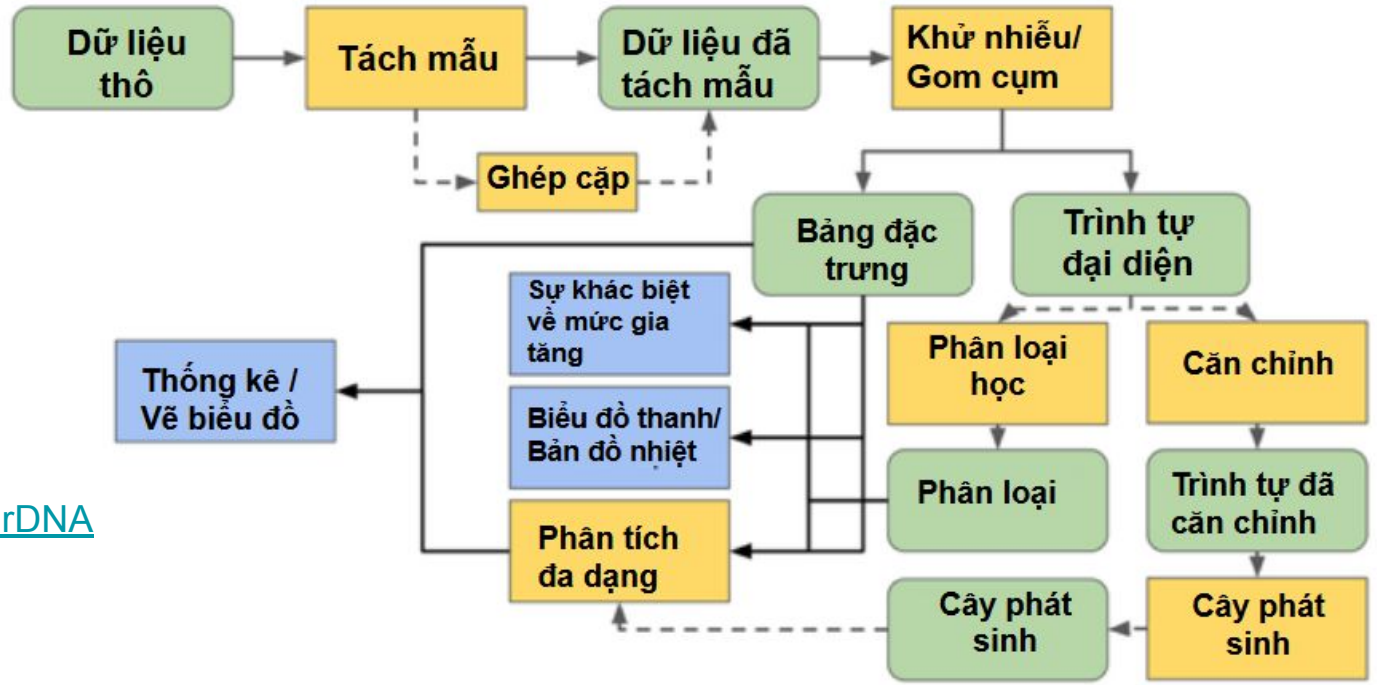
CHI TIẾT

2. Phân tích dữ liệu 16S rDNA hệ vi sinh vật đường ruột

Phân tích dữ liệu với QIIME2

table.qzv, table.qza, manifest.tsv

Workflow QIIME2



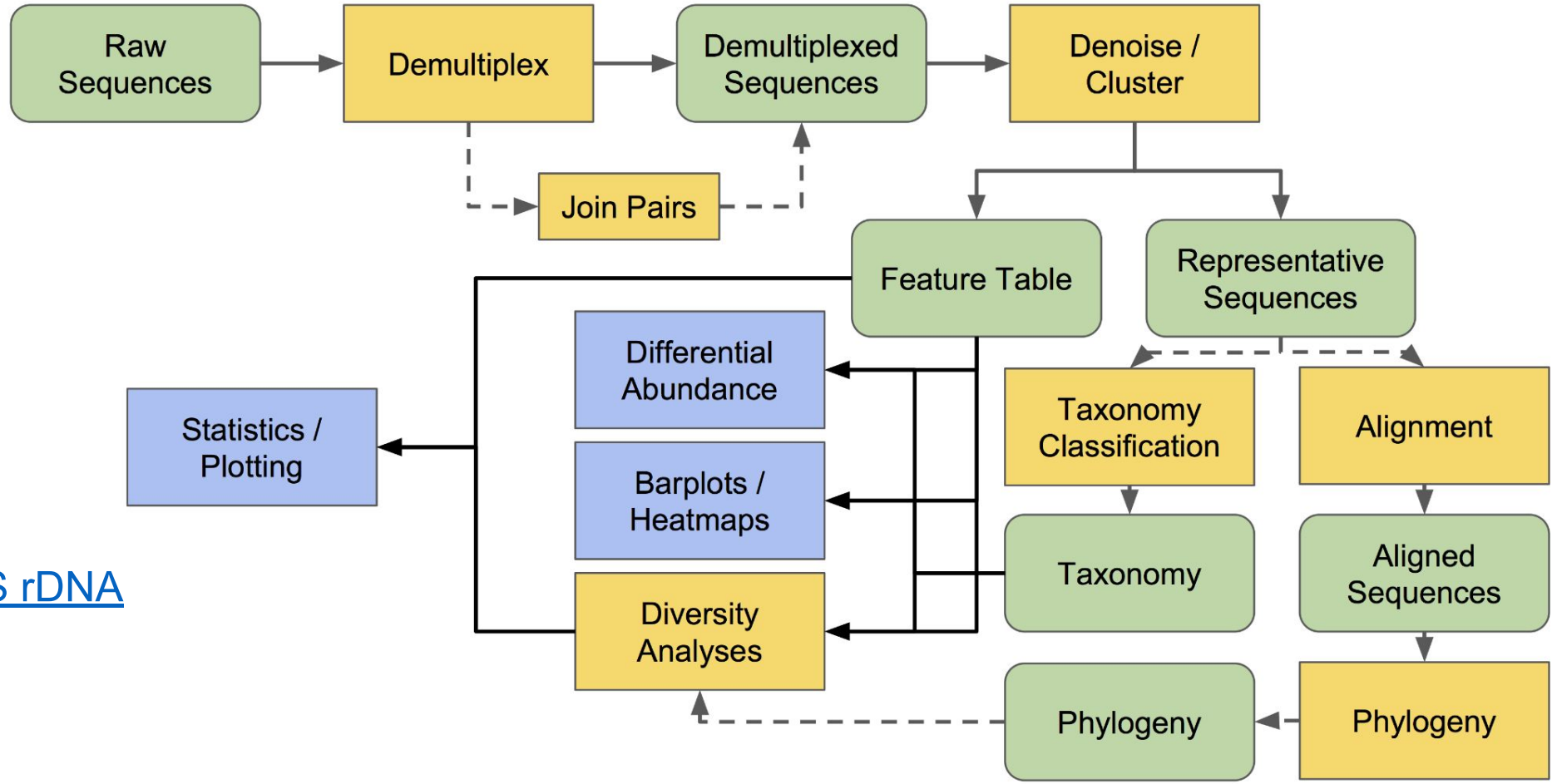
Data Analysis for 16S rDNA

Code QIIME2

Nguồn: https://otagoedna.github.io/edna_workshop_june2021/chapters/04_intro_to_qiime.html.

CHI TIẾT

Workflow QIIME2



[Data Analysis for 16S rDNA]

[Code QIIME2]

Nguồn: https://otagoedna.github.io/edna_workshop_june2021/chapters/04_intro_to_qiime.html.

2. Phân tích dữ liệu 16S rDNA hệ vi sinh vật đường ruột

Phân tích sự khác biệt về độ phong phú - DAA

Feature tables

CHI TIẾT

Tiền xử lý dữ liệu thô

Kích hoạt môi trường QIIME2 (`conda activate qiime2-amplicon-2024.10`).
Nhập dữ liệu từ file manifest vào định dạng `.qza`.
Ghép cặp reads (merge-pairs), lọc chất lượng và làm sạch bằng `deblur`.
Tạo các tệp: `rep-seqs.qza` (chuỗi đại diện), `table.qza` (bảng đặc trưng)

Phân loại trình tự và tạo bảng thành phần vi sinh vật

Tải classifier từ Greengenes.

Dùng `classify-sklearn` để gán phân loại taxanomy cho các ASVs.

Dùng `taxa barplot` để trực quan hóa thành phần vi sinh vật.

Phân tích DAA với ANCOM-BC ở cấp độ ASV

Chạy `qiime composition ancombc` với bảng `table.qza` và metadata.

Tạo biểu đồ `da-barplot` với ngưỡng ý nghĩa thống kê `p < 0.001`.

Gom nhóm vi sinh vật về cấp độ genus

Dùng `taxa collapse` để gom ASVs về cấp độ genus (`--p-level 6`).

Tiến hành lại phân tích ANCOM-BC ở cấp độ genus.

Trực quan hóa kết quả bằng `da-barplot` (cấp độ genus, p < 0.05).

Kiểm tra và xuất kết quả

Dùng `qiime metadata tabulate` để kiểm tra metadata, taxonomy.
Xuất bảng kết quả bằng `qiime tools export`.
Gộp dữ liệu từ nhiều dự án nếu cần (`feature-table merge`, `merge-seqs`, `metadata combine`).

3. Chuẩn bị dữ liệu đầu vào hoàn chỉnh để huấn luyện mô hình dự đoán

📂 **Bắt đầu: Dữ liệu từ nhiều dự án (multiple datasets)**

Gộp dữ liệu từ nhiều nguồn	Rarefying dữ liệu để chuẩn hóa độ sâu	Gán định danh phân loại vi sinh vật	Trực quan hóa phân bố vi sinh vật	Phân tích độ phong phú khác biệt (DAA) với ANCOM-BC	Gom cấp phân loại về genus	Xuất dữ liệu & tạo cây phát sinh loài	Phân tích đa dạng vi sinh vật (diversity)
- Gộp feature table → `qiime feature-table merge` - Gộp rep-seqs → `qiime feature-table merge-seqs` - Gộp metadata → gộp .tsv thủ công	- Giảm xuống cùng mức 2000 reads/mẫu → `qiime feature-table rarefy`	- Sử dụng sklearn + Greengenes → `qiime feature-classifier classify-sklearn`	- Tạo biểu đồ thanh độ phong phú → `qiime taxa barplot`	- Chạy ANCOM-BC → `qiime composition ancombc` - Tạo biểu đồ barplot → `qiime composition da-barplot`	- Collapse taxa → `qiime taxa collapse --p-level 7` - Lặp lại ANCOM-BC và barplot → như trên (thêm `--p-level-delimiter ';'`)	- Export kết quả → `qiime tools export` - Tạo cây → `qiime phylogeny align-to-tree-mafft-fasttree`	- Alpha/beta metrics → `qiime diversity core-metrics-phylogenetic` - So sánh giữa nhóm → `qiime diversity alpha-group-significance` `qiime diversity beta-group-significance`

✅ **Kết thúc: Dữ liệu đã sẵn sàng cho học máy**

GỘP MẪU VÀ XỬ LÝ TIỀN DỮ LIỆU

CHI TIẾT

4. Xây dựng mô hình dự đoán

📂 **Bắt đầu: Dữ liệu microbiome đầu vào (CSV: đặc trưng genus + nhãn IBD/HC)]**

Chuẩn bị dữ liệu & thư viện	Trực quan hóa dữ liệu	Tiền xử lý cho học máy	Huấn luyện mô hình Random Forest	Đánh giá mô hình (Tập test)	Phân tích đặc trưng	Kiểm thử với dữ liệu xác thực
- Nhập thư viện: pandas, numpy, sklearn, seaborn... - Xử lý thiếu dữ liệu (NaN → 0) - Chuyển dữ liệu về tỷ lệ tương đối theo Bacteroides	- PCA & t-SNE: quan sát phân bố IBD vs HC	- Tách X (features), y (label) - Train/Test split (80/20) - SMOTE cân bằng dữ liệu train	- GridSearchCV (cv=5) → chọn tham số tốt - Train mô hình với tập huấn luyện	- Dự đoán nhãn & xác suất - Metrics: Accuracy, Precision, Recall, F1-score - ROC-AUC, PR-curve, Confusion matrix	- Lấy feature_importances_ - Trực quan Top genus ảnh hưởng đến IBD	- Chuẩn hóa dữ liệu từ study khác - Mapping lại format tương tự train - PCA xác thực → dự đoán → đánh giá lại

[✅ Đầu ra cuối cùng: Mô hình Random Forest dự đoán IBD + Các genus liên quan chính]

CHI TIẾT

Cỡ mẫu: Áp dụng công thức tính cỡ mẫu xác định một tỷ lệ cho mục tiêu khảo sát:

$$n = Z^2_{(1-\alpha/2)} \frac{p \times (1-p)}{d^2}$$

Trong đó:

Z: Trị số từ phân phối chuẩn, với khoảng tin cậy 95%, $Z_{1-alpha/2} = 1,96$

d: Độ chính xác mong muốn, với:

$p < 0,1$ sử dụng $d = p/2$

$0,1 \leq p < 0,3$ sử dụng $d = 0,05$

p: Tỷ lệ bệnh nhân cao tuổi được chẩn đoán một trong các bệnh khảo sát

Nguồn: https://cphs.huph.edu.vn/uploads/tintuc/2020/Phuongphapchonmauvatinhtoancomau.pdf

28