Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2025.076

GIẢI PHÁP HỖ TRỢ CHẨN ĐOÁN UNG THƯ ĐẠI TRỰC TRÀNG DỰA TRÊN DỮ LIỆU VI KHUẨN ĐƯỜNG RUỘT ĐƯỢC CHỌN LỌC DÙNG CÁC GIẢI THÍCH PHẢN THỰC ĐA DẠNG

Nguyễn Thanh Hải[1], Đặng Ngọc Xuân Đài[1], Bùi Đăng Hà Phương[1], Nguyễn Thị Thanh Hiền[1] và Trần Thanh Điện[2*]

[1]Trường Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ, Việt Nam
[2]Nhà xuất bản Đại học Cần Thơ, Trường Đại học Cần Thơ, Việt Nam
**Tác giả liên hệ (Corresponding author): thanhdien@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 04/03/2025
Sửa bài (Revised): 12/05/2025
Duyệt đăng (Accepted): 20/05/2025

Title: *An approach for supporting colorectal cancer diagnosis using intestinal bacteria selected with diverse counterfactual explanations*

Author: *Nguyen Thanh Hai[1], Dang Ngoc Xuan Dai[1], Bui Dang Ha Phuong[1], Nguyen Thi Thanh Hien[1] and Tran Thanh Dien[2*]*

Affiliation(s): *[1]College of Information and Communication Technology, Can Tho University, Viet Nam; [2]Can Tho University Publishing House, Can Tho University, Viet Nam*

TÓM TẮT

Bệnh ung thư đại trực tràng là căn bệnh nguy hiểm đến sức khỏe con người nếu không phát hiện và điều trị sớm. Việc phân tích dữ liệu vi sinh vật trong môi trường đường ruột có thể hỗ trợ cho chẩn đoán bệnh này. Cách tiếp cận chọn lọc vi sinh vật bằng phương pháp giải thích kết quả của thuật toán trí tuệ nhân tạo bằng các giải thích phản thực đa dạng (Diverse Counterfactual Explanations-DCE) được đề xuất trong bài viết. Kết quả phân lớp với giải thuật máy học cổ điển như Rừng ngẫu nhiên và Gradient Boosting trên dữ liệu chỉ dưới 3% tổng số đặc trưng ban đầu, đã cho kết quả 0,7759, 0,8055, 0,8093 và 0,7923 với độ đo AUC trên các bộ dữ liệu thu thập từ nhóm người Áo, Mỹ, Trung Quốc, và Đức-Pháp. Kết quả này tốt hơn so với trên tập dữ liệu ban đầu với hơn 1900 loài vi sinh vật.

Từ khóa: *Máy học cổ điển, lựa chọn đặc trưng, ung thư đại trực tràng, vi sinh đường ruột*

ABSTRACT

Colorectal cancer is a dangerous disease that can endanger human health if not detected and treated early. Analysis of microbial data in the intestinal environment can support the diagnosis of this disease. This article proposes a microbial selection approach by Diverse Counterfactual Explanations (DCE). The classification results with classical machine learning algorithms such as Random Forest and Gradient Boosting on data with less than 3% of the total original features are 0.7759, 0.8055, 0.8093, 0.7923, on Austria, American, Chinese, and German-French cohorts, respectively. These results are better than on the original datasets with more than 1900 microbial species.

Keywords: *Colorectal cancer, classical machine learning, feature selection, intestinal microbiota*

1. GIỚI THIỆU

Metagenomics là thuật ngữ được giới thiệu trong Handelsman et al. (1998) để chỉ các nghiên cứu về khai thác thông tin di truyền từ vi sinh vật trong môi trường, từ đó góp phần thúc đẩy việc nghiên cứu những ảnh hưởng của vi sinh vật này trong môi trường mà chúng đang sống. Trong hầu hết trường hợp, sức khỏe của con người và các sinh vật khác đều bị ảnh hưởng bởi vi khuẩn. Vì thế metagenomics cũng được nghiên cứu để phục vụ trong hoạt động chăm sóc sức khỏe con người khi cơ thể con người cũng có rất nhiều hệ vi sinh vật đang sinh sống như: trên da, trong ruột, khoang miệng, khoang mũi, mắt, và các bộ phận khác. Theo kết quả nghiên cứu của Moges and Mengistu (2024) và Martín et al. (2014), các vi sinh vật này có ảnh hưởng lớn đến sức khỏe cả hai mặt tốt lẫn xấu, ảnh hưởng trên hiệu quả sử dụng thuốc, khả năng hấp thu dinh dưỡng của con người. Cùng với việc xác định rõ đóng góp của hệ vi sinh vật cho sức khỏe, phát hiện, điều trị bệnh cho các cá thể, không chỉ với người mà động vật, thực vật cũng được phát triển các phương pháp điều trị mới trên cơ sở các kiến thức metagenomics.

Metagenomics được áp dụng hỗ trợ cho hoạt động y tế. Nguyen et al. (2021) đã phân tích metagenomic virus trong dịch não tủy từ bệnh nhân bị nhiễm trùng hệ thần kinh trung ương, Central Nervous System (CNS) cấp tính không rõ nguyên nhân tại Việt Nam đã sử dụng công nghệ giải trình tự thế hệ mới. Nghiên cứu đã phát hiện 8 loài virus trong 52,4 mẫu dịch não tủy, nhưng sau khi xác nhận bằng Polymerase Chain Reaction (PCR), tỷ lệ phát hiện giảm xuống còn 14,7 với enterovirus là loại virus phổ biến nhất. Kết quả cho thấy gánh nặng lâm sàng của enterovirus tại Việt Nam và nhấn mạnh những thách thức trong việc xác định tác nhân virus hợp lý trong dịch não tủy của bệnh nhân nhiễm trùng CNS. Tang et al. (2025) đã cung cấp những đóng góp quan trọng trong việc đánh giá hiệu quả thực tế của kỹ thuật giải trình tự metagenomics (metagenomic Next-Generation Sequencing - mNGS), từ đó tiềm năng lớn của kỹ thuật này trong hỗ trợ chẩn đoán lâm sàng được chứng minh. Thêm vào đó, tầm quan trọng của việc sử dụng mNGS một cách chọn lọc và hợp lý, để tối ưu chi phí và độ chính xác chẩn đoán được nhấn mạnh. Những phát hiện này mang ý nghĩa thực tiễn cao, giúp các cơ sở y tế xây dựng chiến lược chẩn đoán phù hợp và nâng cao hiệu quả chăm sóc bệnh nhân dựa trên phân tích dữ liệu metagenomic. Với kết quả nghiên cứu trong Kim et al. (2020), tính cấp thiết của việc ứng dụng metagenomics với các thành phần vi sinh vật trong chẩn đoán ung thư đại trực tràng thông qua phân tích hệ vi sinh vật đường ruột được nhấn mạnh cho thấy dữ liệu này là cơ sở quan trọng để chẩn đoán bệnh ung thư đại trực tràng.

Lựa chọn đặc trưng được chú trọng nghiên cứu trên dữ liệu Metagenomic, như Karwowska et al. (2025) đã chỉ rõ tầm quan trọng của hành động này trên dữ liệu metagenomic. Họ đã cho thấy chỉ cần trích lọc 1 bộ nhỏ có thể tiến hành phân lớp chính xác. Trong nghiên cứu Gungor et al. (2023), nhóm tác giả đã cho rằng quá trình lựa chọn đặc trưng trên metagenome khi thực hiện chẩn đoán bệnh dựa có thể giúp chúng ta hiểu rõ hơn về các cơ chế phát triển bệnh, từ đây có thể làm cơ sở cho việc điều trị hiệu quả. Theo LaPierre et al. (2019), nhánh nghiên cứu về lựa chọn đặc trưng để hỗ trợ chẩn đoán bệnh dựa trên metagenomics vẫn chưa được nghiên cứu nhiều nhưng đây là bước quan trọng và có thể được dùng để phát triển, cải thiện khả năng giải thích cho kết quả chẩn đoán trong y học cá thể hóa.

Explainable Artificial Intelligence (xAI) là một lĩnh vực quan trọng trong khoa học dữ liệu. xAI cung cấp các phương pháp và công cụ giúp con người hiểu và giải thích được các quyết định của mô hình học máy. Trong phân tích dữ liệu metagenomic, Machine Learning (ML)- máy học đóng vai trò quan trọng không chỉ trong các ứng dụng dữ liệu DeoxyriboNucleic Acid (DNA) mà còn trong các loại dữ liệu khác. Các thuật toán ML có khả năng học và dự đoán từ các mẫu dữ liệu dựa trên đặc trưng. Tuy nhiên, theo Telenti et al. (2018), các mô hình ML thường hoạt động như "hộp đen", và thường không rõ sự hiệu quả do gặp nhiều thử thách khi phân tích các diễn biến bên trong để nhận biết mô hình kết luận do đâu. Các mô hình học sâu có thể đạt độ chính xác cao trên tập dữ liệu lớn nhưng gây khó hiểu và khó giải thích lý do tại sao một mẫu dữ liệu lại được phân vào một lớp hay các nhóm nào đó.

Các phương pháp xAI không chỉ giúp giải thích cách thức các thuật toán phân tích và xử lý dữ liệu mà còn cung cấp thông tin về tầm quan trọng của từng đặc trưng trong việc tạo ra dự đoán hoặc kết quả. Từ đó, cách tiếp cận này có thể được tận dụng để lựa chọn đặc trưng. Trong bối cảnh nghiên cứu metagenomics, xAI có vai trò rất quan trọng. Bởi vì dữ liệu metagenomics có tính đa dạng và phức tạp cao, việc áp dụng các phương pháp xAI có thể giúp các nhà khoa học hiểu rõ hơn về mối quan hệ giữa các vi sinh vật và các biến đổi trong hệ sinh thái. Trong các nghiên cứu metagenomics về việc xác định phương pháp điều trị, xAI có thể xác định các

yếu tố hoặc nhóm gen tiềm năng có liên quan đến kháng thuốc, giúp cải thiện các phương pháp chẩn đoán và điều trị. Với việc hỗ trợ để giải thích tầm quan trọng đặc trưng, khi ứng dụng xAI ta có thể đánh giá mức độ ảnh hưởng của từng đặc trưng đến kết quả đầu ra, qua đó hỗ trợ loại bỏ các đặc trưng kém quan trọng và lựa chọn ưu tiên những đặc trưng có ảnh hưởng mạnh đến đầu ra kỳ vọng có thể đạt hiệu quả tốt cho bài toán phân lớp. Những giải thuật phổ biến hiện nay như LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017) đã được sử dụng rộng rãi. Riêng với lĩnh vực metagenomics, một số nghiên cứu đã thử áp dụng xAI vào lĩnh vực này. Sekaran et al. (2023) đã khám phá sự mất cân bằng trong hệ vi sinh vật của âm đạo liên quan đến nguyên nhân gây bệnh ung thư cổ tử cung, cho thấy sự thống trị của các vi khuẩn Firmicutes, Actinobacteria và Proteobacteria. Nghiên cứu đã cho thấy sự gia tăng đáng kể của Lactobacillus iners và Prevotella timonensis được xác định có ảnh hưởng đến sự tiến triển của bệnh. Nghiên cứu đã sử dụng SHAP để phân tích kết quả và phát hiện rằng sự gia tăng của Ralstonia có khả năng cao trong việc dự đoán ung thư cổ tử cung, xác nhận sự hiện diện của các vi sinh vật gây bệnh trong mẫu âm đạo của bệnh nhân. Một ứng dụng của mô hình xAI nhằm xác định các dấu ấn sinh học về gen liên quan đến COVID-19 thông qua dữ liệu mNGS trên 234 bệnh nhân được trình bày trong bài báo của Yagin et al. (2023).

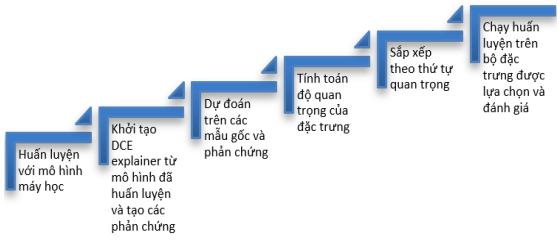
Thực hiện lựa chọn đặc trưng để chẩn đoán bệnh ung thư đại trực tràng, Jabeer et al. (2022) nhận xét rằng kỹ thuật *Select K Best*, *Information Gain* và *XGBoost* đóng góp đáng kể trong việc giảm số lượng vi sinh vật được sử dụng cho chẩn đoán ung thư đại trực tràng, qua đó giúp tiết kiệm chi phí và thời gian; Rừng ngẫu nhiên - Random Forest (RF) có hiệu suất phân loại ung thư đại trực tràng vượt trội hơn so với Adaboost, Máy Vector Hỗ trợ (SVM), Cây Quyết định, Logitboost và các mô hình tổng hợp. Gungor et al. (2025) đã đề xuất phương pháp lựa chọn đặc trưng có tên G-S-M (Grouping-Scoring-Modeling) nhằm phát hiện các enzyme vi sinh vật liên quan đến ung thư đại trực tràng từ dữ liệu metagenomics. G-S-M tích hợp kiến thức sinh học bằng cách nhóm các enzyme theo phân loại EC (enzyme commission) trước khi đánh giá và chọn đặc trưng. Kết quả cho thấy, một số enzyme như glycosidase, CoA-transferase và hydro-lyase có liên quan đến bệnh ung thư đại trực tràng. Phương pháp được đề xuất vượt trội hơn so với các phương pháp chọn đặc trưng truyền thống khác trong các thí nghiệm so sánh.

Merrick & Taly (2020) đã đề xuất phương pháp giải thích kết quả của các thuật toán trí tuệ nhân tạo dựa trên các phản chứng đa dạng: Diverse Counterfactuals Explanation (DCE). Không giống như các phương pháp truyền thống tập trung vào việc phân tích tầm quan trọng của từng đặc trưng, các phản chứng đa dạng được tạo ra trong nghiên cứu này, tức là các điểm dữ liệu giả lập gần với mẫu gốc nhưng có kết quả dự đoán khác biệt. Mục tiêu chính là giải thích các dự đoán của mô hình học máy được sử dụng để hỗ trợ ra quyết định trong các lĩnh vực quan trọng đối với xã hội như: tài chính, y tế, giáo dục và tư pháp hình sự. Tuy nhiên, việc ứng dụng DCE để phân tích lựa chọn đặc trưng cho dữ liệu y tế là một khoảng trống nghiên cứu rất có tiềm năng cần được thực hiện để đánh giá mức độ hiệu quả của DCE trên dữ liệu Metagenomic. Do đó, mục tiêu của nghiên cứu này là đề xuất cách tiếp cận lựa chọn đặc trưng với DCE và đánh giá độ hiệu quả của cách tiếp cận này trên bài toán chẩn đoán ung thư đại trực tràng trên dữ liệu vi sinh vật đường ruột. Bên cạnh, mục tiêu của nghiên cứu cũng đánh giá độ biến động của hiệu quả với các tập đặc trưng lần lượt tăng dần từ 1 đến 50 đặc trưng đã được xếp hạng độ quan trọng với giải thuật DCE. Kết quả cho thấy cách tiếp cận này có hiệu quả trong việc giảm được số lượng đặc trưng cần xem xét tuy nhiên vẫn đạt hiệu suất phân loại cao so với trên tập dữ liệu ban đầu chưa lựa chọn đặc trưng.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Hình 1 minh họa các bước chính cho việc triển khai một giải thuật xAI như DCE cho việc trích chọn đặc trưng. Đầu tiên, dữ liệu được huấn luyện để có mô hình đã được huấn luyện và sử dụng DCE Explainer để có thể tạo các phản chứng nhằm đánh giá mức độ quan trọng của các đặc trưng trên việc quyết định của mô hình.

Các đặc trưng được xếp hạng và lần lượt đưa vào các mô hình ML để đánh giá độ hiệu quả của các bộ đặc trưng tìm được.



Hình 1. Các bước chính cho quá trình lựa chọn đặc trưng với DCE

2.1. Các mô hình máy học huấn luyện

Việc lựa chọn các đặc trưng với các phương pháp xAI thông thường cần bắt đầu bằng việc huấn luyện các mô hình máy học để từ đó có thể cung cấp các mô hình đã được huấn luyện đưa vào các thuật toán giải thích để đánh giá mức độ đóng góp của các đặc trưng vào việc ra quyết định của mô hình. Mô hình đã được huấn luyện này kết hợp với kỹ thuật xAI có thể đánh giá mức độ đóng góp của mỗi đặc trưng vào quyết định của mô hình trên mỗi mẫu dữ liệu. Để thực hiện việc huấn luyện, các giải thuật máy học cổ điển nhưng có hiệu quả cao đã được chứng minh qua các nghiên cứu phân tích trên dữ liệu Metagenomic là RF và Gradient Boosting Classifier (GBC) được sử dụng. Cách giải thuật này được đánh giá rất cao về mặt hiệu suất qua độ chính xác cao và thời gian thực hiện nhanh trong các bài toán phân lớp trên dữ liệu Metagenomic.

RF, GBC hoạt động bằng cách xây dựng nhiều cây quyết định và kết hợp chúng để đưa ra kết quả cuối cùng. Mỗi cây quyết định trong rừng được huấn luyện trên một tập con của dữ liệu và được dự đoán kết quả, sau đó kết quả của tất cả các cây được tổng hợp lại để đưa ra dự đoán cuối cùng. Các siêu tham số được sử dụng bao gồm n_estimators là 500 (số lượng cây tương ứng khuyến nghị trong Pasolli et al. (2016), Nguyen and TN (2020) và Karwowska et al. (2025)).

2.2. Lựa chọn đặc trưng với Diverse Counterfactuals explanation

Wachter et al. (2017) và Merrick & Taly (2020) đề xuất giải quyết vấn đề đa dạng trong các phản thực (counterfactuals) bằng cách đề xuất một ràng buộc đa dạng giữa các phản thực được tạo ra ngẫu nhiên.

Như mô tả trong nghiên cứu, *Giải thích phản thực* (counterfactual explanation) là một cách tiếp cận giải thích bằng việc đề xuất một kịch bản giả định ngược với thực tế, cho biết điều gì cần thay đổi để một mô hình máy học cho ra kết quả khác. Ngoại trừ trên các mô hình tuyến tính đơn giản được áp dụng, rất thử thách để tạo ra các giải thích phản thực đạt hiệu quả tốt cho mọi mô hình máy học. Chính vì thế, các giải thích phản thực cho các mô hình được đề xuất. Ý tưởng cốt lõi là thiết lập việc tìm kiếm các giải thích như một vấn đề tối ưu hóa, tương tự như việc tìm các ví dụ đối nghịch.

Trong kết quả thể hiện ở Hình 1, sau quá trình huấn luyện để có mô hình đã huấn luyện trên bộ dữ liệu được tạo, đối tượng DCE Explainer được khởi tạo bởi nghiên cứu để bắt đầu quá trình giải thích các dự đoán của mô hình. Mẫu dữ liệu cụ thể cần giải thích sẽ được chọn, trong nghiên cứu này là 50 mẫu ngẫu nhiên trong mỗi bộ dữ liệu được chọn và nhãn mục tiêu là 0 hoặc 1, thể hiện có hoặc không bị ung thư. Sau khi có mẫu dữ liệu, giải thuật DCE được sử dụng để tạo ra các phản chứng, đây là các mẫu dữ liệu giả định từ việc thay đổi một số đặc trưng sao cho mô hình đưa ra dự đoán khác với mẫu ban đầu. Ở đây, thực nghiệm được lựa chọn với số lượng phản chứng là 5 vì với số lượng này có thể thu thập đủ dữ liệu để phân tích, đồng thời tránh làm tăng độ phức tạp và tăng chi phí tính toán quá mức. Đây cũng là một số lượng hợp lý để đạt được kết quả phù hợp mà không làm quá tải quá trình tính toán. Việc tạo ra 5 phản chứng giúp phân tích, quan sát dễ dàng và so sánh sự thay đổi trong kết quả dự đoán của mô hình khi có sự thay đổi về đặc trưng. Sau khi tạo phản chứng, kết quả cho cả mẫu dữ liệu gốc và các phản chứng để kiểm tra sự thay đổi của mô hình khi có sự thay đổi về các đặc trưng được dự đoán. Kết quả dự đoán giữa mẫu gốc và phản chứng sẽ được so sánh để hiểu rõ sự thay đổi trong mô hình khi thay đổi đầu vào. Một cách chi tiết, quy trình lựa chọn đặc trưng với DCE được liệt kê như sau:

- Lấy mẫu dữ liệu đầu vào: 50 mẫu ngẫu nhiên được chọn từ các tập dữ liệu, với số lượng mẫu này chỉ chiếm từ 50% trở xuống trên tổng số lượng mẫu đã có của mỗi bộ. Các mẫu ngẫu nhiên này được dùng để đánh giá độ quan trọng của đặc trưng. Khi DCE được áp dụng trên mỗi mẫu cho ra các độ đo quan trọng trên mỗi đặc trưng và để lựa chọn các đặc trưng thì 50 mẫu này được tính trung bình để sắp xếp lựa chọn.
- Tạo phản chứng: DCE được sử dụng để tạo ra số lượng phản chứng cho từng mẫu đã chọn. DCE tạo ra các phản chứng là các bản sao của mẫu với các đặc trưng thay đổi nhằm xem liệu dự đoán của mô hình có thay đổi không. Công thức (công thức (1)) là hàm mục tiêu tối ưu hóa, được thiết kế để cân bằng giữa các yếu tố như: độ gần gũi (proximity), độ đa dạng (diversity) và tính khả thi (feasibility) của các phản chứng. Công thức này tối ưu hóa quá trình sinh phản chứng sao cho phản chứng vừa khác đầu ra ban đầu (thay đổi được dự đoán) vừa gần với mẫu gốc (giữ tính khả thi) đồng thời đảm bảo các phản chứng đa dạng (tránh trùng lặp), không những thế, nó còn có thể tạo ra các ví dụ có ý nghĩa và hợp lý để giải thích quyết định của mô hình:

$$C(x) = arg \min_{c_1,\dots,c_k} \left(\frac{1}{k}\sum_{i=1}^{k} y_{-loss(f(c_1),y)} + \lambda_1 \frac{1}{k}\sum_{i=1}^{k} dist(c_i, x) - \lambda_2 Diversity(c_i, \dots, c_k) \right) \quad (1)$$

Trong đó: $y_{-loss(f(c_1),y)}$ là hàm mất mát giữa đầu ra của mô hình f(ci) và nhãn mục tiêu y; λ_1, λ_2 là các siêu tham số để cân bằng giữa 3 phần của hàm mất mát. dist(ci,x) là khoảng cách giữa mỗi phản chứng ci và ví dụ đầu vào gốc x, dùng để đảm bảo tính khả thi của phản chứng (tức là độ gần gũi). Diversity(c1,..., ck) là chỉ số đo lường sự đa dạng giữa các phản chứng trong tập hợp {c1, c2,..., ck}.

- So sánh phản chứng với mẫu gốc: Từng phản chứng được so sánh với mẫu gốc bằng cách kiểm tra sự khác biệt ở từng đặc trưng. Nếu giá trị của một đặc trưng thay đổi trong phản chứng so với mẫu ban đầu thì có thể coi đó là một đặc trưng quan trọng nhất, ảnh hưởng nhiều nhất trong việc dự đoán kết quả.
- Tính toán tần suất thay đổi: Số lần mỗi đặc trưng thay đổi trong tất cả các phản chứng được đếm. Với mỗi đặc trưng, nếu giá trị của nó khác với mẫu ban đầu trong một phản chứng thì được tính là 1 lần thay đổi. Kết quả được thể hiện ở Hình 2 minh họa các đặc trưng có tần suất ảnh hưởng nhiều nhất.
- Tính toán chỉ số DCE (DCE Value): Tỷ lệ số lần đặc trưng đó thay đổi trong các phản chứng so với mẫu gốc là độ đo dùng để xếp hạng các đặc trưng. Khi tổng số lần thay đổi chia cho tổng số phản chứng, giá trị của đặc trưng cao nhất được xếp hạng là quan trọng nhất. Trong đó công thức của DCE Value dựa trên công thức tính Sparsity trong bài báo gốc và được tính như sau (Công thức 2):

$$DCE\_Value(f1) = \frac{1}{k}\sum_{i=0}^{k} 1_{[c_i^l \neq x^l]} \quad (2)$$

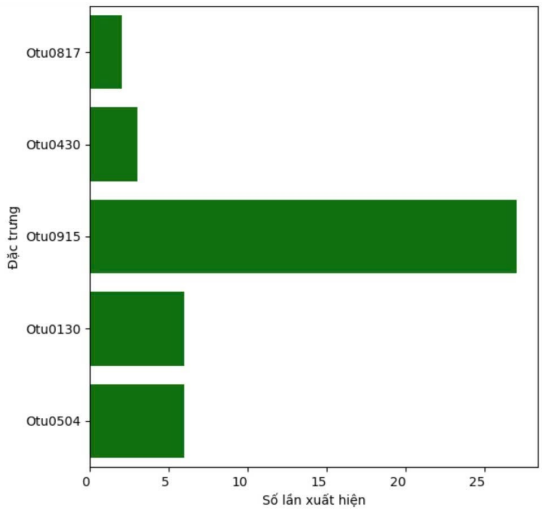
Trong đó:

k: số lượng phản chứng,

d: số đặc trưng,

$1_{[c_i^l \neq x^l]}$: đặc trưng thứ l trong phản chứng i khác với mẫu gốc.

- Chọn đặc trưng quan trọng nhất: Các đặc trưng được sắp xếp theo số lần thay đổi và đặc trưng có số lần thay đổi cao nhất được chọn.



Hình 2. Một minh họa về các đặc trưng có tần suất xuất hiện nhiều nhất trong 50 mẫu phân tích với DCE trên một bộ dữ liệu được khảo sát

Số lượng phản chứng càng lớn giúp hiểu rõ hơn về cách thức các đặc trưng có thể thay đổi để dẫn đến một kết quả khác. Sự ổn định của các đặc trưng được kiểm tra qua nhiều phản chứng cũng là một bước quan trọng. Nếu một đặc trưng thường xuyên thay đổi trong các phản chứng, đó là dấu hiệu cho thấy đặc trưng đó có tầm quan trọng cao. Cuối cùng, khi có nhiều phản chứng, khoảng cách trung bình giữa mẫu gốc và các phản chứng được tính toán sẽ giúp xác định các đặc trưng ảnh hưởng nhiều nhất, từ đó tăng độ chính xác cho các đặc trưng quan trọng. Trong dữ liệu thành phần vi khuẩn trong cơ thể bệnh nhân, mẫu được lựa chọn ngẫu nhiên 50 bệnh nhân, có thể tìm ra được trong 50 bệnh nhân đó các đặc trưng ảnh hưởng nhất đến kết quả. DCE tạo ra các phản chứng bằng cách tự thay đổi các đặc trưng để kết quả dự đoán có thể thay đổi. Khi mẫu ban đầu và phản chứng được so sánh, độ quan trọng của các vi khuẩn có thể được tính. Hình 2 thể hiện vi khuẩn có nhãn là Otu0915 có giá trị cao nhất, Otu0915 là loài Parvimonas micra, cũng được chỉ ra trong Dai et al. (2018) được xem như một trong chỉ dấu sinh học quan trọng dùng trong chẩn đoán bệnh ung thư đại trực tràng.

Sau khi có được bảng xếp hạng của các đặc trưng, giải thuật RF và GBC lần lượt được chạy để học và đánh giá trên các tập đặc trưng đã được xếp hạng lần lượt từ 1 đến 50 đặc trưng, với các siêu tham số của giải thuật RF và GBC được thiết lập như ở nghiên cứu (Nguyen and TN, 2020) để so sánh độ hiệu quả. Với mỗi tập đặc trưng quan trọng được chạy với 10-fold-cross-validation và tính giá trị trung bình kèm độ lệch chuẩn.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Dữ liệu thực nghiệm

Bốn bộ dữ liệu thu thập từ Dai et al. (2018) được sử dụng trong quá trình đánh giá phương pháp đề xuất được trình bày trong Bảng 1. Mỗi bộ có hơn 1900 đặc trưng được thu thập từ các nhóm người Áo, Mỹ, Trung Quốc, Đức và Pháp. Mỗi mẫu bao gồm tỷ lệ các loài vi sinh sống trong môi trường ruột người. Bộ dữ liệu được đặt tên là Ung thư (UT)1, UT2, UT3, UT4 lần lượt có 109, 100, 165, và 152 mẫu với số lượng đặc trưng thấp nhất là 1932 và cao nhất là 1981.

Bảng 1. Bốn bộ dữ liệu thực nghiệm

Tên bộ	#mẫu	#đặc trưng	#bệnh	#không bệnh
UT1	109	1981	46	63
UT2	100	1976	48	52
UT3	165	1932	92	73
UT4	152	1980	88	64

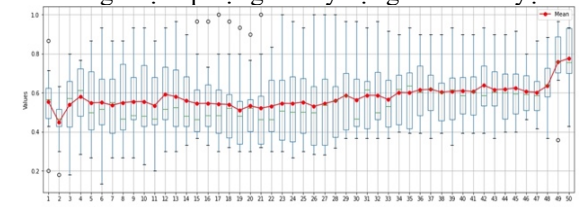
3.2. Môi trường thực nghiệm và độ đo đánh giá

Để triển khai thực nghiệm, các thư viện phổ biến như scikit-learn, numpy, pandas, math, keras, deepmg, matplotlib, và một số thư viện khác được sử dụng trong nghiên cứu để triển khai các bước xử lý và phân tích. Các mô hình học máy như RF và GBC cũng được áp dụng để xây dựng và huấn luyện các mô hình dự đoán. Ngoài ra, thư viện dice-ml (Mothilal et al., 2020) cũng được sử dụng để đánh giá lựa chọn đặc trưng từ các tập dữ liệu.

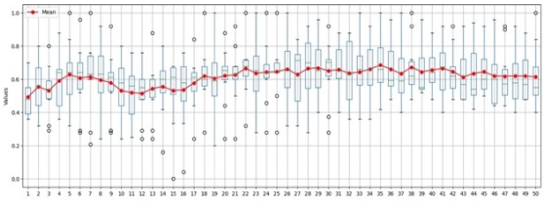
Độ hiệu quả được đánh giá với diện tích dưới đường cong (Area Under the Curve (AUC)) trung bình trên 10-fold-cross-validation. Kết quả được chạy với giải thuật RF và GBC với 500 cây (các siêu tham số còn lại để mặc định) để so sánh với Nguyen and TN (2020). Độ đo AUC là một độ đo đánh giá hiệu suất mô hình đáng tin cậy ngay cả trên các tập dữ liệu mất cân bằng và thường được sử dụng phổ biến trong các bài toán phân lớp trên dữ liệu y khoa.

3.3. Kết quả trên 50 đặc trưng quan trọng được lựa chọn bởi DCE

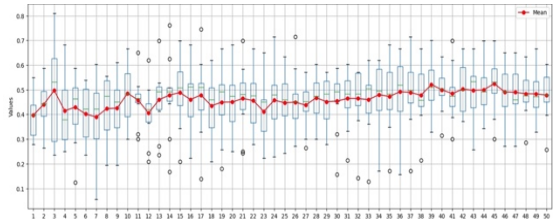
Kết quả với giải thuật RF được minh họa như Hình 3, thông qua quá trình thực nghiệm trên 10-fold-cross-validation với số lượng đặc trưng từ 1 đến 50, kết quả trên 4 tập dataset được hiển thị trên hình. Kết quả trên tập UT3 cho thấy, giá trị AUC tăng mạnh từ đặc trưng 1 đến khoảng đặc trưng thứ 3, sau đó có xu hướng ổn định, giá trị AUC cao nhất ở số lượng đặc trưng 45. Tiếp theo là tập UT1 có giá trị AUC giảm đột ngột ở đặc trưng thứ 2 nhưng sau đó tăng chậm ở các đặc trưng tiếp theo, đến khoảng 40 đặc trưng thì mới có xu hướng tăng nhanh, với kết quả AUC cao nhất được thực nghiệm là ở số lượng đặc trưng cao nhất được khảo sát.



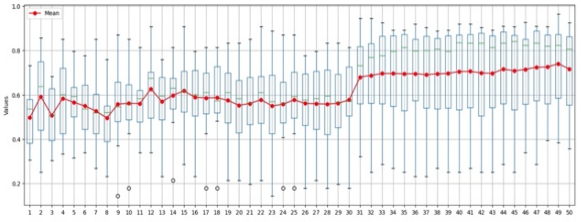
a) UT1



b) UT2

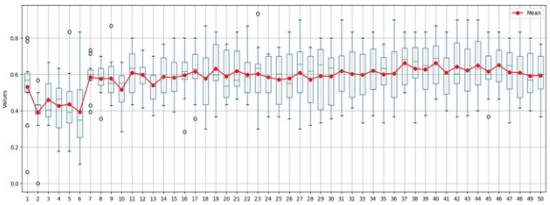


c) UT3

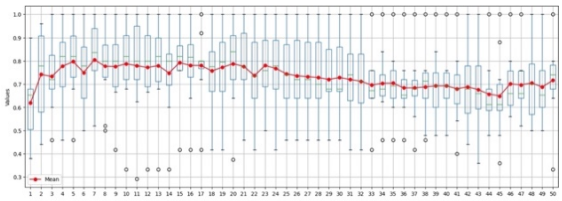


d) UT4

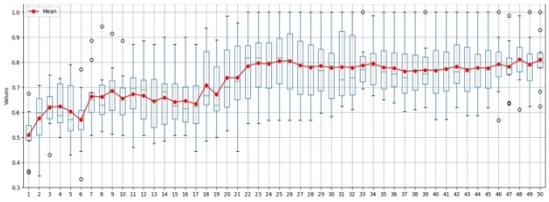
Hình 3. Kết quả phân lớp với độ đo AUC từ 1 đến 50 đặc trưng quan trọng với RF trên 4 bộ dữ liệu



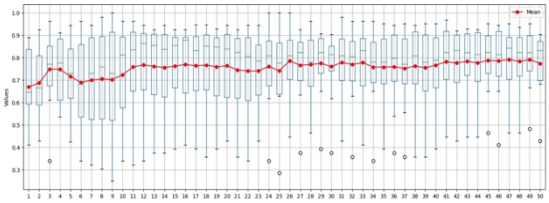
a) UT1



b) UT2



c) UT3



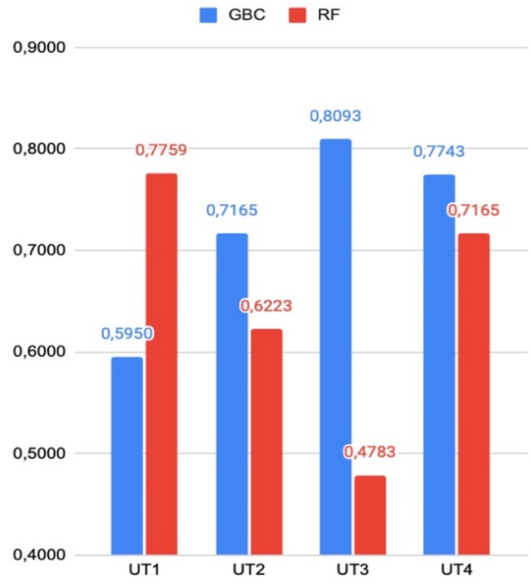
d) UT4

Hình 4. Kết quả phân lớp với độ đo AUC từ 1 đến 50 đặc trưng quan trọng với GBC trên 4 bộ dữ liệu

Kế tiếp là tập UT4, biểu đồ có xu hướng tăng giảm lúc đầu nhưng khoảng từ 30 đặc trưng đã thấy mức AUC cao và tăng liên tục, có kết quả thu được cao nhất với giá trị AUC là khoảng 49 đặc trưng. Tiếp tục là tập UT2 AUC có xu hướng tăng đến khoảng từ 10 đặc trưng sau đó có dấu hiệu ổn định, kết quả thực nghiệm có giá trị AUC cao nhất sau quá trình thử nghiệm là ở 35 đặc trưng. Trong hầu hết các bộ dữ liệu, giá trị AUC có xu hướng tăng nhẹ ở phần đầu nhưng sau đó đạt trạng thái ổn định và không cải thiện đáng kể khi thêm nhiều đặc trưng hơn; đặc trưng đầu tiên có ảnh hưởng lớn đến độ hiệu quả trong việc phân lớp, trong khi các đặc trưng bổ sung về sau có mức đóng góp giảm dần.

Hình 4 thể hiện kết quả với GBC. Kết quả trên tập UT1 cho thấy AUC không ổn định ở những số đặc trưng đầu cho đến khoảng 7 đặc trưng và dần ổn định với sự chênh lệch nhẹ khi số đặc trưng được chọn từ 11 đến 36, kết quả đạt đỉnh ở số đặc trưng 37, trong khi trên tập UT2 sau 7 đặc trưng theo độ đo AUC cho thấy có chiều hướng giảm. Tập UT3 cho thấy là AUC có xu hướng tăng dần theo số lượng đặc trưng được chọn và đạt gần đỉnh ở những số đặc trưng lớn. Kết quả trên UT3 cho thấy khoảng 25 đặc trưng thì độ đo AUC đi vào ổn định. Tương tự với UT3, kết quả trên UT4 cũng cho thấy sau 25 đặc trưng thì kết quả cũng dần đi vào ổn định.

Kết quả cho thấy, phần lớn kết quả trên GBC tốt hơn. So sánh chung thì GBC cho kết quả cao với ít đặc trưng hơn so với RF. Ngoài trừ trên tập UT1, các tập khác GBC cho kết quả tốt hơn.



Hình 5. So sánh kết quả (độ đo AUC) với 50 đặc trưng với được chọn trên 4 bộ dữ liệu với RF và GBC

Hình 5 thể hiện sự so sánh của RF và GBC trên 50 đặc trưng được chọn. Kết qua thấy GBC vượt trội với kết quả đạt tốt hơn trên 3 bộ UT2, UT3, UT4, trong khi RF chỉ tốt hơn trên UT1 với độ chênh lệch là gần 0,2. Độ chênh lệch lớn nhất là trên bộ UT3 với độ lệch là hơn 0,3, trong khi trên UT4 là hơn 0,05 và UT2 là hơn 0,08.

3.4. Thảo luận kết quả

Dựa trên kết quả đạt được ở Bảng 2 cho thấy hầu hết kết quả với GBC cho kết quả cao với số đặc trưng cần hầu hết là ít hơn. So với kết quả chạy trên

bộ ban đầu với hơn 1900 đặc trưng từ nghiên cứu của Nguyen and TN (2020) với số lượng đặc trưng chỉ dưới 3% so với các bộ ban đầu. Tuy nhiên kết quả thể hiện cho thấy, các đặc trưng được lựa chọn có độ quan trọng ảnh hưởng đến hiệu suất phân lớp. Đáng chú ý là chỉ với 7 đặc trưng nhưng có thể đạt được chính xác rất cao trên bộ UT3 khi so sánh với tập ban đầu, tuy nhiên điều này cũng phù hợp với nghiên cứu của Dai et al. (2018), nơi các tác giả cũng đề xuất sử dụng 7 đặc trưng. Ngoài trừ trên tập UT1, các kết quả tốt nhất đạt được với GBC chênh lệch có thể lên đến hơn 28% với bộ UT3 và chênh lệch thấp nhất trên UT4 với khoảng 5%.

Kết quả cũng cho thấy RF không ổn định bằng GBC được thể hiện qua độ lệch chuẩn cao khi so sánh với GBC trên tất cả các tập dữ liệu. Sự khác biệt về độ lệch chuẩn khoảng 1-5%.

Bảng 2. So sánh kết quả với giá trị trung bình và độ lệch chuẩn với độ đo AUC (%). Số mở ngoặc là số đặc trưng để đạt được kết quả

	UT1	UT2	UT3	UT4
RF	**77,59** ±16,40 (#50)	68,60 ±18,77 (#35)	52,50 ±12,24 (#45)	74,04 ±19,93 (#49)
GBC	66,25 ±14,52 (#40)	**80,55** ±16,61 (#7)	**80,93** ±10,76 (#50)	**79,23** ±15,41 (#47)
GBC (Nguyen and TN, 2020)	76,83 (#1981)	66,69 (#1976)	70,38 (#1932)	78,64 (#1980)

4. KẾT LUẬN

DCE được đề xuất sử dụng trong nghiên cứu cho quá trình lựa chọn đặc trưng trên dữ liệu vi sinh vật đường ruột nhằm hỗ trợ chẩn đoán bệnh ung thư đại trực tràng. Trong nghiên cứu, 4 bộ dữ liệu Metagenomic với số lượng đặc trưng ban đầu là hơn 1900 được đưa ra phân tích trong thực nghiệm. Với kết quả đã được phân tích và so sánh cho thấy rằng chỉ với dưới 3% số lượng đặc trưng ban đầu, các giải thuật phân lớp cổ điển RF và GBC đã đạt hiệu suất cao hơn trên các bộ ban đầu. Điều này thể hiện rằng việc lựa chọn đã trích chọn ra được những đặc trưng có ý nghĩa, có thể cải thiện hiệu quả mô hình phân lớp để hỗ trợ chẩn đoán bệnh ung thư đại trực tràng trên dữ liệu hệ vi sinh đường ruột. Với kết quả thể hiện cũng cho thấy rằng GBC hiệu quả về độ chính xác phân lớp với độ đo AUC hơn RF trong đa phần các trường hợp với số lượng đặc trưng đa phần trong các trường hợp so sánh cũng ít hơn.

Mặc dù vậy, hiệu suất cao nhất đạt được không đồng đều và nhất quán ở số lượng đặc trưng được chọn, điều này đòi hỏi có những phương pháp cải tiến hiệu quả hơn cần được thực hiện trong tương lai. Hiện, nghiên cứu mới chỉ đánh giá tiềm năng của DCE trong việc lựa chọn đặc trưng, các nghiên cứu tương lai có thể phân tích sâu và khai phá thêm về tiềm năng của giải thuật này và so sánh thêm với các nghiên cứu về các phương pháp lựa chọn đặc trưng khác. Thêm nữa, các mô hình học sâu có thể được thêm vào so sánh ở các nghiên cứu sắp tới để thấy rõ ưu điểm và nhược điểm của mỗi loại giải thuật.

LỜI CẢM TẠ

Đề tài này được tài trợ bởi Trường Đại học Cần Thơ, mã số: T2024-87.

TÀI LIỆU THAM KHẢO (REFERENCES)

Nguyen A. T., Le N.T.N., Nguyen H.T.T., Tran P.M., Pham T.T.T., Dang H.T., Tran A.T., Deng X., Ho N.T.D., Tran N.T., Nguyen H.V., Nguyen T.D., Pham P.H.T., Nguyen C.V.V., Baker, S., Delwart, E., Thwaites, G., & Le T.V., (2021). Viral Metagenomic Analysis of Cerebrospinal Fluid from Patients with Acute Central Nervous System Infections of Unknown Origin, Vietnam. *Emerging Infectious Diseases*, *27*(1), 205–213. https://doi.org/10.3201/eid2701.202723

Gungor, B., Ersoz, N. S., & Yousef, M. (2025). Integrating Biological Domain Knowledge with Machine Learning for Identifying Colorectal-Cancer-Associated Microbial Enzymes in Metagenomic Data. *Applied Sciences*, *15*(6), 2940. https://doi.org/10.3390/app15062940

Gungor, B., Temiz, M., Jabeer, A., Wu, D., & Yousef, M. (2023). microBiomeGSM: The identification of taxonomic biomarkers from metagenomic data using grouping, scoring and modeling (G-S-M) approach. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/fmicb.2023.1264941

Dai, Z., Coker, O. O., Nakatsu, G., Wu, W. K. K., Zhao, L., Chen, Z., Chan, F. K. L., Kristiansen, K., Sung, J. J. Y., Wong, S. H., & Yu, J. (2018). Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome*, *6*(1). https://doi.org/10.1186/s40168-018-0451-2

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular

biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–R249. https://doi.org/10.1016/s1074-5521(98)90108-9

Jabeer, A., Kocak, A., Akkas, H., Yenisert, F., Nalbantoglu, O. U., Yousef, M., & Bakir Gungor, B. (2022). Identifying Taxonomic Biomarkers of Colorectal Cancer in Human Intestinal Microbiota Using Multiple Feature Selection Methods. *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 1–6. https://doi.org/10.1109/asyu56188.2022.9925551

Karwowska, Z., Aasmets, O., Metspalu, M., Metspalu, A., Milani, L., Esko, T., Kosciolek, T., & Org, E. (2025). Effects of data transformation and model selection on feature importance in microbiome classification data. *Microbiome*, *13*(1). https://doi.org/10.1186/s40168-024-01996-6

Kim, D. J., Yang, J., Seo, H., Lee, W. H., Ho Lee, D., Kym, S., Park, Y. S., Kim, J. G., Jang, I.-J., Kim, Y.-K., & Cho, J.-Y. (2020). Colorectal cancer diagnostic model utilizing metagenomic and metabolomic data of stool microbial extracellular vesicles. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-59529-8

LaPierre, N., Ju, C. J.-T., Zhou, G., & Wang, W. (2019). MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods*, *166*, 74–82. https://doi.org/10.1016/j.ymeth.2019.03.003

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. arXiv. https://doi.org/10.48550/ARXIV.1705.07874

Martín, R., Miquel, S., Langella, P., & Bermúdez, L. G. (2014). The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, *5*(3), 413–423. https://doi.org/10.4161/viru.27864

Merrick, L., & Taly, A. (2020). The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In *Machine Learning and Knowledge Extraction* (17–38). Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_2

Moges, B., & Mengistu, D. Y. (2024). Metagenomics approaches for studying the human microbiome. *All Life*, *17*(1). https://doi.org/10.1080/26895293.2024.2350166

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617. https://doi.org/10.1145/3351095.3372850

Nguyen, T., & TN, N. (2020). Diagnosis Approaches for Colorectal Cancer Using Manifold Learning and Deep Learning. *SN Computer Science*, *1*(5), 281. https://doi.org/10.1007/s42979-020-00297-7

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Computational Biology*, *12*(7), e1004977. https://doi.org/10.1371/journal.pcbi.1004977

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. https://doi.org/10.18653/v1/N16-3020

Sekaran, K., Varghese, R. P., Gopikrishnan, M., Alsamman, A. M., El Allali, A., Zayed, H., & Doss C, G. P. (2023). Unraveling the Dysbiosis of Vaginal Microbiome to Understand Cervical Cancer Disease Etiology—An Explainable AI Approach. *Genes*, *14*(4), 936. https://doi.org/10.3390/genes14040936

Tang, H., Chen, Y., Tang, X., Wei, M., Hu, J., Zhang, X., Xiang, D., Yang, Q., & Han, D. (2025). Yield of clinical metagenomics: Insights from real-world practice for tissue infections. *eBioMedicine*, *111*, 105536. https://doi.org/10.1016/j.ebiom.2024.105536

Telenti, A., Lippert, C., Chang, P.-C., & DePristo, M. (2018). Deep learning of genomic variation and regulatory network data. *Human Molecular Genetics*, *27*(Supplement_R1), R63–R71. https://doi.org/10.1093/hmg/ddy115

Wachter, S., Mittelstadt, B., & Russell, C. (2017). *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. https://doi.org/10.48550/ARXIV.1711.00399

Yagin, F. H., Cicek, İ. B., Alkhateeb, A., Yagin, B., Colak, C., Azzeh, M., & Akbulut, S. (2023). Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. *Computers in Biology and Medicine*, *154*, 106619. https://doi.org/10.1016/j.compbiomed.2023.106619