ĐÁNH GIÁ GIÁ TRỊ CHẨN ĐOÁN CỦA HSA-MIR-1203 LƯU HÀNH TRONG UNG THƯ TUYẾN TIỀN LIỆT BẰNG CÁC PHƯƠNG PHÁP HỌC MÁY

Nguyễn Tú Anh, Dương Thị Kim Chi, Lê Thị Thanh Thảo
Đỗ Mai Xuân Diệu và Quang Trọng Minh✉

Trường Dược, Đại học Y Dược Thành phố Hồ Chí Minh

Ung thư tuyến tiền liệt là một trong những bệnh ung thư phổ biến ở nam giới, trong đó việc tìm kiếm dấu ấn sinh học tuần hoàn có giá trị chẩn đoán vẫn là nhu cầu cấp thiết. Nghiên cứu này nhằm đánh giá giá trị của hsa-miR-1203 trong việc hỗ trợ phân biệt bệnh nhân ung thư tuyến tiền liệt và nhóm không ung thư bằng mô hình học máy. Dữ liệu biểu hiện microRNA được khai thác từ bộ GSE211692 trên cơ sở dữ liệu Gene Expression Omnibus, gồm 1027 mẫu huyết thanh bệnh nhân ung thư và 5893 mẫu không ung thư. Sau tiền xử lý và chuyển đổi $\log_2$, dữ liệu được chia thành tập huấn luyện (70%) và tập kiểm định nội bộ (30%) theo phương pháp phân tầng. Hsa-miR-1203 giảm biểu hiện rõ rệt ở nhóm ung thư ($\log_2 FC = -3,77$; $p < 0,001$). Các mô hình Extra Trees, Support Vector Machine, AdaBoost và Gaussian Naive Bayes cho hiệu năng phân loại cao trên tập kiểm định với diện tích dưới đường cong ROC xấp xỉ 0,98. Kết quả gợi ý hsa-miR-1203 có tiềm năng trở thành dấu ấn sinh học tuần hoàn hỗ trợ chẩn đoán ung thư tuyến tiền liệt.

Từ khóa: Ung thư tuyến tiền liệt, microRNA, hsa-miR-1203, học máy, dấu ấn sinh học tuần hoàn, chẩn đoán.

I. ĐẶT VẤN ĐỀ

Ung thư tuyến tiền liệt là một trong những bệnh ung thư thường gặp nhất ở nam giới và là nguyên nhân tử vong do ung thư đứng hàng đầu trên toàn cầu. Theo GLOBOCAN 2020, ung thư tuyến tiền liệt đứng thứ hai về tỷ lệ mắc mới và thứ năm về tỷ lệ tử vong do ung thư ở nam giới, phản ánh gánh nặng y tế đáng kể đối với nhiều quốc gia.[1] Tại Việt Nam, số ca mắc có xu hướng gia tăng trong những năm gần đây, một phần do quá trình già hóa dân số và sự cải thiện trong hệ thống ghi nhận và chẩn đoán ung thư.[2,3] Tuy nhiên, việc phát hiện sớm và phân biệt chính xác giữa bệnh nhân ung thư và người không mắc bệnh vẫn còn nhiều thách thức trong thực hành lâm sàng.

Hiện nay, chẩn đoán ung thư tuyến tiền liệt chủ yếu dựa vào định lượng kháng nguyên đặc hiệu tuyến tiền liệt (prostate-specific antigen - PSA), thăm khám trực tràng và sinh thiết mô bệnh học. Mặc dù PSA được sử dụng rộng rãi trong sàng lọc, độ đặc hiệu còn hạn chế, có thể dẫn đến chẩn đoán quá mức và can thiệp không cần thiết.[4] Bên cạnh đó, PSA không phản ánh đầy đủ các thay đổi phân tử liên quan đến quá trình sinh ung. Do đó, việc tìm kiếm các dấu ấn sinh học mới có giá trị chẩn đoán cao hơn và có thể tiếp cận bằng phương pháp ít xâm lấn là nhu cầu cấp thiết.

Trong những năm gần đây, microRNA (miRNA/miR) được quan tâm như một nhóm phân tử điều hòa hậu phiên mã có vai trò quan trọng trong sinh bệnh học ung thư. MiRNA là các RNA không mã hóa ngắn, tham gia

Tác giả liên hệ: Quang Trọng Minh
Trường Dược, Đại học Y Dược Thành phố Hồ Chí Minh
Email: qtminh@ump.edu.vn
Ngày nhận: 26/02/2026
Ngày được chấp nhận: 13/03/2026

điều hòa biểu hiện gen thông qua cơ chế ức chế dịch mã hoặc thúc đẩy phân hủy mRNA đích.[5,6] Sự rối loạn biểu hiện miRNA đã được ghi nhận trong nhiều loại ung thư, bao gồm ung thư tuyến tiền liệt.[7] Đáng chú ý, miRNA có thể được phát hiện ổn định trong huyết thanh và huyết tương nhờ được bảo vệ trong túi ngoại bào hoặc liên kết với protein, cho phép ứng dụng như các dấu ấn sinh học tuần hoàn không xâm lấn.[8,9] Nhiều miRNA lưu hành đã được nghiên cứu trong ung thư tuyến tiền liệt, trong đó một số miRNA tăng biểu hiện và liên quan đến tiến triển bệnh, trong khi một số khác giảm biểu hiện và có thể đóng vai trò ức chế khối u.[7,10] Sự giảm biểu hiện của các miRNA có chức năng điều hòa âm tính có thể góp phần làm mất kiểm soát các con đường tín hiệu liên quan đến tăng sinh và xâm lấn tế bào.[11,12]

Trong số các miRNA tiềm năng, hsa-miR-1203 gần đây bắt đầu được chú ý trong một số nghiên cứu phân tích dữ liệu hệ gen và tin sinh học. Các nghiên cứu sàng lọc miRNA ở mô sinh thiết bệnh nhân ung thư tuyến tiền liệt đã ghi nhận sự thay đổi biểu hiện của nhiều miRNA và gợi ý rằng một số miRNA ít được nghiên cứu, bao gồm hsa-miR-1203, có thể tham gia vào mạng lưới điều hòa phân tử liên quan đến sinh ung.[5,13] Tuy nhiên, dữ liệu về vai trò của hsa-miR-1203 trong ung thư tuyến tiền liệt vẫn còn hạn chế, đặc biệt trong các nghiên cứu sử dụng bộ dữ liệu huyết thanh quy mô lớn. Do đó, việc đánh giá sự thay đổi biểu hiện của hsa-miR-1203 lưu hành và khả năng phân biệt giữa bệnh nhân ung thư tuyến tiền liệt và nhóm không ung thư có thể cung cấp thêm bằng chứng về tiềm năng ứng dụng của miRNA này như một dấu ấn sinh học chẩn đoán.

Sự phát triển của các cơ sở dữ liệu công khai như Gene Expression Omnibus (GEO) cho phép khai thác dữ liệu biểu hiện miRNA trên số lượng mẫu lớn, từ đó tăng tính khách quan và khả năng tái lập của nghiên cứu.[14] Đồng thời, các phương pháp học máy ngày càng được ứng dụng rộng rãi trong lĩnh vực ung thư nhằm hỗ trợ phân loại bệnh dựa trên dữ liệu phân tử.[15,16] Việc tích hợp biểu hiện của một dấu ấn sinh học đơn lẻ với các thuật toán phân loại có thể cung cấp bằng chứng định lượng về khả năng phân biệt giữa nhóm bệnh và không bệnh thông qua các chỉ số như diện tích dưới đường cong ROC. Mặc dù nhiều miRNA lưu hành đã được đề xuất như các dấu ấn sinh học tiềm năng cho ung thư tuyến tiền liệt, kết quả giữa các nghiên cứu vẫn còn chưa đồng nhất và phần lớn tập trung vào một số miRNA đã được nghiên cứu rộng rãi. Trong khi đó, vai trò của nhiều miRNA ít được khảo sát, bao gồm hsa-miR-1203, vẫn chưa được đánh giá đầy đủ trên các bộ dữ liệu huyết thanh quy mô lớn. Việc khai thác các cơ sở dữ liệu công khai với số lượng mẫu lớn kết hợp với các phương pháp phân loại dữ liệu có thể giúp đánh giá khách quan hơn giá trị chẩn đoán của các miRNA tiềm năng này. Từ những cơ sở trên, nghiên cứu này được thực hiện nhằm: (1) đánh giá sự khác biệt biểu hiện của hsa-miR-1203 trong huyết thanh giữa bệnh nhân ung thư tuyến tiền liệt và nhóm không ung thư dựa trên bộ dữ liệu quy mô lớn GSE211692, (2) xây dựng và đánh giá các mô hình phân loại dựa trên mức biểu hiện hsa-miR-1203 bằng các thuật toán học máy; (3) định lượng khả năng phân biệt của dấu ấn sinh học này thông qua các chỉ số chẩn đoán như độ nhạy, độ đặc hiệu và diện tích dưới đường cong ROC.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

1. Đối tượng

Đối tượng nghiên cứu bao gồm các mẫu huyết thanh của bệnh nhân ung thư tuyến tiền liệt và nhóm không ung thư được trích xuất từ cơ sở dữ liệu GEO thuộc Trung tâm Thông tin Công nghệ Sinh học Quốc gia Hoa Kỳ (NCBI) (https://www.ncbi.nlm.nih.gov/geo/). Bộ dữ liệu được lựa chọn cho phân tích là GSE211692.

Bộ dữ liệu ban đầu bao gồm 1027 mẫu huyết thanh từ bệnh nhân ung thư tuyến tiền liệt và 5893 mẫu huyết thanh từ nhóm không ung thư, với tổng cộng 6920 mẫu. Nhóm không ung thư bao gồm các đối tượng được phân loại là "control" trong bộ dữ liệu GSE211692 và không được chẩn đoán mắc ung thư tuyến tiền liệt tại thời điểm thu thập mẫu theo mô tả của nghiên cứu gốc. Tuy nhiên, bộ dữ liệu công khai không cung cấp thông tin chi tiết về tình trạng bệnh lý tuyến tiền liệt lành tính hoặc các đặc điểm lâm sàng khác của nhóm chứng. Do đó, đơn vị phân tích trong nghiên cứu này là từng mẫu huyết thanh có dữ liệu định lượng biểu hiện miRNA và có nhãn phân loại rõ ràng ở mức tối thiểu "prostate cancer" hoặc "control".

Các dữ liệu được đưa vào phân tích phải đáp ứng các tiêu chuẩn sau: được thu nhận từ các nghiên cứu khảo sát biểu hiện miRNA tuần hoàn trong ung thư tuyến tiền liệt và có mô tả rõ phương pháp đo lường mức biểu hiện; sử dụng mẫu huyết thanh, huyết tương hoặc máu người; có phân nhóm rõ ràng giữa bệnh và không bệnh. Các nghiên cứu thực hiện trên dòng tế bào ung thư, mô hình động vật hoặc mẫu xenograft khối u người không được đưa vào phân tích. Các mẫu từ bệnh nhân đã hoặc đang được phẫu thuật, xạ trị, hóa trị hoặc liệu pháp hormone, cũng như các trường hợp mắc ung thư ở cơ quan khác, bị loại trừ nhằm hạn chế các yếu tố gây nhiễu ảnh hưởng đến mức biểu hiện miRNA.

Nghiên cứu tập trung vào mục tiêu phân loại chẩn đoán nên không khai thác các biến lâm sàng đi kèm.

2. Phương pháp

Nghiên cứu được thiết kế theo dạng phân tích hồi cứu dựa trên dữ liệu thứ cấp công khai, kết hợp phân tích thống kê và xây dựng mô hình học máy nhằm đánh giá khả năng phân loại bệnh nhân ung thư tuyến tiền liệt và nhóm không ung thư dựa trên biểu hiện hsa-miR-1203 trong huyết thanh. Trong nghiên cứu này, các thuật ngữ liên quan đến học máy và tin sinh học được sử dụng theo định nghĩa chuẩn trong lĩnh vực phân tích dữ liệu sinh học. Cụ thể, $\log_2$ fold change ($\log_2$FC) biểu thị mức độ thay đổi biểu hiện của miRNA giữa hai nhóm trên thang logarit cơ số 2. Đường cong ROC và diện tích dưới đường cong (AUC) được sử dụng để đánh giá khả năng phân biệt của dấu ấn sinh học giữa hai lớp dữ liệu. Các mô hình học máy được huấn luyện trên tập dữ liệu huấn luyện và được đánh giá trên tập dữ liệu kiểm tra độc lập nhằm hạn chế hiện tượng quá khớp và đảm bảo khả năng khái quát của mô hình.

Tiền xử lý dữ liệu

Dữ liệu biểu hiện miRNA được tải từ GEO bằng công cụ GEOquery trong môi trường R (phiên bản 4.3.1).[17] Với dữ liệu microarray, bước chú giải probe được thực hiện để xác định probe tương ứng với hsa-miR-1203. Trong trường hợp một miRNA có nhiều probe đại diện, giá trị biểu hiện được xác định theo quy tắc định trước nhằm bảo đảm tính nhất quán. Bên cạnh đó, dữ liệu biểu hiện miRNA được tiền xử lý trước khi xây dựng mô hình học máy. Các giá trị biểu hiện được biến đổi $\log_2$ nhằm giảm ảnh hưởng của các giá trị ngoại lai và đưa phân phối dữ liệu về gần phân phối chuẩn. Sau đó, các bước chuẩn hóa dữ liệu được thực hiện trên toàn bộ bộ dữ liệu biểu hiện miRNA. Vì dữ liệu được thu nhận từ cùng một nền tảng microarray (3D-Gene Human miRNA V21_1.0.0), nguy cơ sai lệch do hiệu ứng lô (batch effect) được xem là hạn chế. Tuy nhiên, dữ liệu vẫn được kiểm tra phân bố tổng thể nhằm phát hiện các sai lệch bất thường trước khi tiến hành phân tích. Sau bước tiền xử lý này, dữ liệu mới được chia thành tập huấn luyện và tập kiểm tra nhằm tránh nguy cơ rò rỉ thông tin giữa các tập dữ liệu.[18]

Biến phụ thuộc là tình trạng bệnh, được mã hóa nhị phân (UTTTL so với không ung thư). Biến độc lập duy nhất trong mô hình là mức biểu hiện hsa-miR-1203 sau tiền xử lý và chuẩn hóa.

Chia dữ liệu và xây dựng mô hình

Dữ liệu được chia thành tập huấn luyện (70%) và tập kiểm định nội bộ (30%) bằng phương pháp lấy mẫu ngẫu nhiên phân tầng nhằm duy trì tỷ lệ phân bố giữa hai nhóm trong cả hai tập. Thiết lập seed cố định được sử dụng nhằm bảo đảm khả năng tái lập phân tích. Tập huấn luyện được sử dụng cho toàn bộ các bước tối ưu siêu tham số và huấn luyện mô hình, trong khi tập kiểm định chỉ được sử dụng để đánh giá hiệu năng cuối cùng nhằm tránh hiện tượng rò rỉ dữ liệu. Do mô hình chỉ sử dụng một biến liên tục duy nhất, các phương pháp chọn đặc trưng không được áp dụng. Đối với các thuật toán nhạy cảm với thang đo dữ liệu như Support Vector Machine (SVM) và Gaussian Naive Bayes (GaussianNB), mức biểu hiện hsa-miR-1203 được chuẩn hóa theo phương pháp z-score. Các tham số chuẩn hóa được ước tính từ tập huấn luyện và sau đó được áp dụng cho cả tập huấn luyện và tập kiểm định. Tối ưu siêu tham số được thực hiện trên tập huấn luyện bằng phương pháp Randomized Search kết hợp với Stratified K-Fold Cross-Validation, với chỉ số tối ưu là AUC-ROC.[19] Sau khi tối ưu, từng mô hình được huấn luyện trên toàn bộ tập huấn luyện và đánh giá trên tập kiểm định nội bộ.

Phân tích thống kê và đánh giá hiệu năng

Để xác định sự khác biệt biểu hiện của hsa-miR-1203 giữa nhóm ung thư tuyến tiền liệt và nhóm đối chứng, phân tích biểu hiện khác biệt được thực hiện bằng gói limma trong môi trường R. Phương pháp limma sử dụng mô hình tuyến tính để ước lượng sự khác biệt biểu hiện giữa các nhóm và đã được sử dụng rộng rãi trong phân tích dữ liệu microarray và giải trình tự RNA. Các giá trị biểu hiện được biến đổi $\log_2$ trước khi phân tích. Giá trị $\log_2$FC được tính dựa trên sự khác biệt trung bình biểu hiện $\log_2$ giữa hai nhóm, và ý nghĩa thống kê được đánh giá thông qua kiểm định thống kê hai phía với ngưỡng $p < 0,05$.

Hiệu năng mô hình được đánh giá thông qua diện AUC, độ nhạy, độ đặc hiệu và độ chính xác. Khi cần so sánh AUC giữa các mô hình, phép kiểm định DeLong được áp dụng.[20] Ngưỡng ý nghĩa thống kê được xác định tại $p < 0,05$. Toàn bộ phân tích học máy được thực hiện bằng Python (phiên bản 3.10).

3. Đạo đức nghiên cứu

Nghiên cứu được thực hiện dưới hình thức phân tích hồi cứu trên dữ liệu thứ cấp công khai từ cơ sở dữ liệu GEO thuộc NCBI. Bộ dữ liệu GSE211692 được sử dụng trong nghiên cứu đã được công bố công khai và toàn bộ thông tin nhận dạng cá nhân của người tham gia đã được loại bỏ trước khi chia sẻ trên hệ thống này. Nghiên cứu không tiến hành thu thập mẫu bệnh phẩm mới, không can thiệp lâm sàng, không tiếp xúc trực tiếp với người tham gia và không sử dụng bất kỳ dữ liệu cá nhân định danh nào. Dữ liệu được phân tích hoàn toàn ở dạng ẩn danh.

Theo quy định hiện hành về đạo đức trong nghiên cứu y sinh học, các nghiên cứu sử dụng dữ liệu thứ cấp đã được công khai và ẩn danh hoàn toàn, không thể truy xuất danh tính người tham gia, được xem là không thuộc phạm vi nghiên cứu can thiệp trên đối tượng người và không yêu cầu xin chấp thuận riêng của Hội đồng Đạo đức trong nghiên cứu y sinh học. Việc khai thác và sử dụng dữ liệu tuân thủ các điều khoản của GEO/NCBI và các nguyên tắc đạo đức trong nghiên cứu y sinh học.

III. KẾT QUẢ

1. Đặc điểm bộ dữ liệu nghiên cứu

Bộ dữ liệu GSE211692 bao gồm tổng cộng 6920 mẫu huyết thanh, trong đó có 1027 mẫu từ bệnh nhân ung thư tuyến tiền liệt và 5893 mẫu từ nhóm không ung thư. Sau khi kiểm tra tính đầy đủ của nhãn phân loại và chất lượng dữ liệu biểu hiện, toàn bộ các mẫu hợp lệ được đưa vào phân tích.

Dữ liệu được chia thành tập huấn luyện (70%) và tập kiểm định nội bộ (30%) bằng phương pháp lấy mẫu ngẫu nhiên phân tầng nhằm duy trì tỷ lệ phân bố giữa hai nhóm. Tổng số mẫu trong tập huấn luyện là 4844 và trong tập kiểm định là 2076.

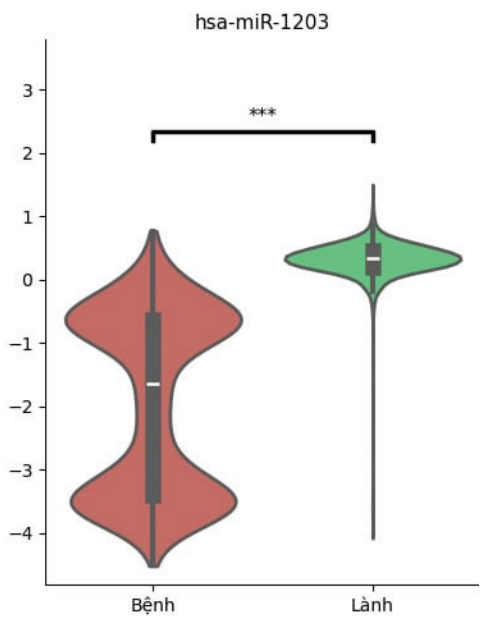
Bảng 1. Phân bố mẫu trong bộ dữ liệu GSE211692

Nhóm	Tổng số mẫu	Tập huấn luyện (70%)	Tập kiểm định (30%)
Ung thư tuyến tiền liệt	1027	719	308
Không ung thư	5893	4125	1768
Tổng cộng	**6920**	**4844**	**2076**

Chú thích: Dữ liệu được chia bằng phương pháp phân tầng để bảo toàn tỷ lệ bệnh/chứng trong cả hai tập

2. Sự khác biệt biểu hiện hsa-miR-1203 giữa hai nhóm

Sau khi thực hiện chuyển đổi $\log_2$ và chuẩn hóa dữ liệu theo quy trình, mức biểu hiện hsa-miR-1203 ở nhóm ung thư tuyến tiền liệt thấp hơn rõ rệt so với nhóm không ung thư (Biểu đồ 1). Phân tích biểu hiện cho thấy hsa-miR-1203 giảm mạnh ở nhóm bệnh ($\log_2$FC = −3,77), tương ứng với mức giảm khoảng 14 lần so với nhóm chứng. Sự khác biệt này có ý nghĩa thống kê ($p < 0,001$).



Biểu đồ 1. Phân bố mức biểu hiện hsa-miR-1203 giữa nhóm ung thư và nhóm không ung thư

Chú thích: Biểu đồ thể hiện phân bố mức biểu hiện của hsa-miR-1203 trong hai nhóm. Mức biểu hiện ở nhóm ung thư thấp hơn rõ rệt so với nhóm không ung thư

Biểu đồ 1 cho thấy sự dịch chuyển phân bố biểu hiện về phía giá trị thấp ở nhóm ung thư, phù hợp với kết quả phân tích biểu hiện (log_2FC = −3,77). Sự khác biệt phân bố giữa hai nhóm phản ánh khả năng phân biệt rõ ràng của hsa-miR-1203 trong mẫu huyết thanh bệnh nhân ung thư tuyến tiền liệt.

3. Hiệu năng phân loại của các mô hình học máy

Bốn mô hình phân loại đơn biến được xây dựng dựa trên mức biểu hiện của hsa-miR-1203 gồm Extra Trees, SVM (RBF), AdaBoost và GaussianNB. Hiệu năng được đánh giá trên tập kiểm định nội bộ (n = 2076).

Bảng 2. Hiệu năng phân loại trên tập kiểm định nội bộ

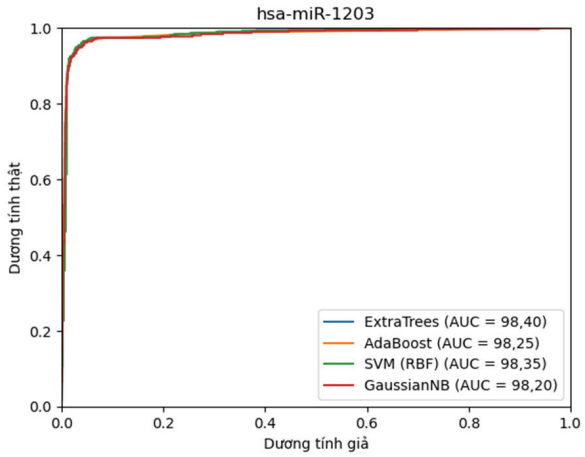
Mô hình	Độ nhạy (%)	Độ đặc hiệu (%)	Độ chính xác (%)	AUC
Extra Trees	95,5	96,4	96,2	0,9840
AdaBoost	93,5	97,5	96,9	0,9825
SVM (RBF)	94,8	97,0	96,7	0,9835
GaussianNB	89,6	98,6	97,3	0,9820

Chú thích: AUC được tính trên tập kiểm định nội bộ

Tất cả các mô hình đều cho hiệu năng phân loại rất cao với diện tích dưới đường cong ROC xấp xỉ 0,98. Extra Trees có AUC cao nhất (0,9840); tuy nhiên, chênh lệch AUC giữa các mô hình rất nhỏ (< 0,003). Kết quả này cho thấy khả năng phân biệt chủ yếu phản ánh giá trị tín hiệu sinh học của hsa-miR-1203, trong khi sự khác biệt về cấu trúc thuật toán chỉ đóng vai trò thứ yếu trong bối cảnh phân tích đơn biến.

4. Đường cong ROC

Các đường cong ROC của bốn mô hình đều tiến sát góc trên bên trái, phản ánh khả năng phân biệt rất cao giữa hai nhóm (Biểu đồ 2). Phép kiểm định DeLong cho thấy không có sự khác biệt có ý nghĩa thống kê giữa các mô hình (p > 0,05).



Biểu đồ 2. Đường cong ROC của các mô hình phân loại trên tập kiểm định

Chú thích: Các đường cong ROC gần như chồng lấp và tiến sát góc trên bên trái của biểu đồ. Extra Trees đạt AUC cao nhất, tiếp theo là SVM, AdaBoost và GaussianNB

Tổng thể, hsa-miR-1203 có sự giảm biểu hiện rõ rệt ở bệnh nhân ung thư tuyến tiền liệt so với nhóm không ung thư. Khi được tích hợp vào các mô hình máy học đơn biến, mức giảm biểu hiện của miRNA này cho hiệu năng phân loại rất cao trên tập kiểm định nội bộ, với AUC xấp xỉ 0,98 ở tất cả các thuật toán.

IV. BÀN LUẬN

Kết quả nghiên cứu cho thấy hsa-miR-1203 giảm biểu hiện rõ rệt trong huyết thanh bệnh nhân ung thư tuyến tiền liệt so với nhóm không ung thư, với mức thay đổi lớn ($log_2FC = -3,77$), tương ứng giảm khoảng 14 lần. Mức giảm mạnh này không chỉ có ý nghĩa thống kê mà còn phản ánh sự khác biệt sinh học đáng kể giữa hai nhóm. Trong sinh học ung thư, miRNA đóng vai trò quan trọng trong điều hòa hậu phiên mã các gen liên quan đến tăng sinh, biệt hóa và chết tế bào theo chương trình.[5] Sự suy giảm biểu hiện của các miRNA có chức năng ức chế khối u có thể góp phần làm mất kiểm soát các con đường tín hiệu liên quan đến sinh ung và tiến triển bệnh.[7] Do đó, mức giảm biểu hiện đáng kể của hsa-miR-1203 ghi nhận trong nghiên cứu này gợi ý vai trò tiềm năng của miRNA này trong cơ chế bệnh sinh của ung thư tuyến tiền liệt.

Một phát hiện đáng chú ý là hiệu năng phân loại rất cao của các mô hình học máy khi chỉ sử dụng một biến duy nhất. Giá trị AUC xấp xỉ 0,98 cho thấy mức biểu hiện hsa-miR-1203 có khả năng phân biệt mạnh giữa bệnh nhân ung thư và nhóm không ung thư. Sự chênh lệch rất nhỏ về AUC giữa các thuật toán khác nhau (< 0,003) cho thấy hiệu năng chủ yếu phụ thuộc vào tín hiệu sinh học của hsa-miR-1203 hơn là vào cấu trúc thuật toán. Điều này phù hợp với quan điểm rằng khi một dấu ấn sinh học có độ tách biệt rõ ràng giữa hai nhóm, nhiều phương pháp phân loại khác nhau có thể đạt hiệu năng tương đương.[15,16] Ngoài ra, mục tiêu của việc áp dụng nhiều thuật toán học máy không phải để xây dựng mô hình dự đoán đa biến phức tạp mà nhằm so sánh hiệu năng của các phương pháp phân loại khác nhau khi sử dụng cùng một đặc trưng sinh học. Cách tiếp cận này cho phép đánh giá tính ổn định của khả năng phân biệt của dấu ấn sinh học trên nhiều khung thuật toán khác nhau. Trong nhiều nghiên cứu dấu ấn sinh học ung thư, các mô hình dựa trên một biến thường đạt hiệu năng phân loại ở mức trung bình do tính dị hợp sinh học cao của bệnh ung thư cũng như sự chồng lấp biểu hiện phân tử giữa các nhóm bệnh và nhóm chứng. Tuy nhiên, trong nghiên cứu hiện tại, mức độ khác biệt biểu hiện rất lớn của hsa-miR-1203 giữa hai nhóm ($log_2FC = -3,77$) có thể tạo ra tín hiệu phân loại mạnh, qua đó góp phần nâng cao hiệu năng của mô hình. Bên cạnh đó, bộ dữ liệu GSE211692 có kích thước mẫu lớn và được thu thập trên cùng một nền tảng microarray, điều này có thể giúp giảm thiểu sai lệch kỹ thuật và cải thiện độ ổn định của tín hiệu sinh học. Tuy vậy, một số nghiên cứu trước đây đã chỉ ra rằng hiệu năng phân loại rất cao quan sát được trong các bộ dữ liệu công khai đôi khi có thể bị ảnh hưởng bởi đặc điểm của bộ dữ liệu hoặc các yếu tố kỹ thuật trong quá trình thu thập và xử lý dữ liệu. Vì vậy, các kết quả của nghiên cứu này cần được xác nhận thêm trên các quần thể bệnh nhân độc lập nhằm đánh giá đầy đủ tính khái quát của dấu ấn sinh học.[21]

Mặc dù kết quả của nghiên cứu cho thấy hsa-miR-1203 có sự thay đổi biểu hiện đáng kể và có khả năng phân biệt giữa bệnh nhân ung thư tuyến tiền liệt và nhóm không ung thư, vai trò sinh học cụ thể của miRNA này trong sinh bệnh học của ung thư tuyến tiền liệt hiện vẫn chưa được làm rõ. Trong các nghiên cứu về miRNA, việc dự đoán gen đích và phân tích làm giàu con đường tín hiệu thường được sử dụng để khám phá các cơ chế phân tử tiềm năng

mà miRNA có thể tham gia điều hòa. Các công cụ sinh tin học như TargetScan, miRDB hoặc miRTarBase cho phép dự đoán các gen đích tiềm năng của miRNA dựa trên các đặc điểm bảo tồn trình tự và tương tác miRNA-mRNA đã được xác nhận thực nghiệm.[6,22] Ngoài ra, phân tích làm giàu con đường tín hiệu thông qua các cơ sở dữ liệu như KEGG hoặc Gene Ontology có thể giúp xác định các mạng lưới sinh học liên quan đến quá trình sinh ung, bao gồm các con đường điều hòa chu kỳ tế bào, apoptosis hoặc tín hiệu tăng trưởng.[23,24] Do đó, các nghiên cứu trong tương lai nên kết hợp các phương pháp dự đoán gen đích và phân tích làm giàu con đường tín hiệu cùng với các thí nghiệm chức năng *in vitro* hoặc *in vivo* để làm rõ vai trò phân tử của hsa-miR-1203 trong ung thư tuyến tiền liệt.

Trong thực hành lâm sàng, xét nghiệm PSA hiện vẫn là phương pháp sàng lọc được sử dụng phổ biến nhất cho ung thư tuyến tiền liệt. Tuy nhiên, PSA có độ nhạy cao nhưng độ đặc hiệu còn hạn chế, đặc biệt trong vùng "xám" của nồng độ PSA, dẫn đến nguy cơ chẩn đoán quá mức và sinh thiết không cần thiết.[25] Do đó, nhiều nghiên cứu đã tập trung vào việc phát triển các dấu ấn phân tử mới, bao gồm các miRNA lưu hành, nhằm cải thiện độ chính xác chẩn đoán. Một số nghiên cứu đã đề xuất các bộ (panel) miRNA trong huyết thanh hoặc huyết tương có khả năng phân biệt ung thư tuyến tiền liệt với nhóm chứng với hiệu năng phân loại đáng kể.[13] So với các bộ dấu ấn đa biến này, nghiên cứu hiện tại chỉ đánh giá một miRNA đơn lẻ. Mặc dù, hsa-miR-1203 cho thấy khả năng phân biệt cao trong bộ dữ liệu được phân tích, cần có thêm các nghiên cứu so sánh trực tiếp với PSA hoặc các bộ miRNA đã được công bố để đánh giá đầy đủ hơn giá trị bổ sung của dấu ấn này trong bối cảnh lâm sàng.

Việc sử dụng bộ dữ liệu quy mô lớn với 6920 mẫu là một ưu điểm quan trọng của nghiên cứu. Cỡ mẫu lớn giúp tăng độ ổn định của ước lượng và giảm nguy cơ sai lệch do biến thiên ngẫu nhiên. Đồng thời, quy trình chia dữ liệu phân tầng thành tập huấn luyện và tập kiểm định nội bộ góp phần hạn chế nguy cơ quá khớp và nâng cao độ tin cậy của kết quả. So với các nghiên cứu trước đây thường thực hiện trên số lượng mẫu hạn chế, cách tiếp cận dựa trên dữ liệu lớn giúp tăng tính khách quan và khả năng tái lập của phát hiện.Tuy nhiên, nghiên cứu vẫn tồn tại một số hạn chế. Thứ nhất, đây là phân tích hồi cứu dựa trên dữ liệu thứ cấp công khai, do đó phụ thuộc vào quy trình thu thập và xử lý dữ liệu của nghiên cứu gốc. Thứ hai, nghiên cứu chỉ tập trung vào mục tiêu phân loại chẩn đoán và chưa đánh giá mối liên quan giữa mức biểu hiện hsa-miR-1203 với đặc điểm lâm sàng hoặc tiên lượng bệnh. Thứ ba, nghiên cứu chưa có tập dữ liệu kiểm định độc lập từ quần thể khác để xác nhận thêm khả năng khái quát hóa của mô hình. Các nghiên cứu tiền cứu và đa trung tâm trong tương lai là cần thiết để xác nhận giá trị lâm sàng của hsa-miR-1203.

Tổng hợp các kết quả cho thấy hsa-miR-1203 giảm biểu hiện đáng kể trong huyết thanh bệnh nhân ung thư tuyến tiền liệt và có khả năng phân biệt tốt giữa nhóm bệnh và nhóm không bệnh khi được tích hợp vào các mô hình máy học đơn biến. Phát hiện này góp phần bổ sung bằng chứng về tiềm năng ứng dụng của miRNA tuần hoàn như dấu ấn sinh học hỗ trợ chẩn đoán ung thư tuyến tiền liệt.

V. KẾT LUẬN

Nghiên cứu này cho thấy hsa-miR-1203 có sự giảm biểu hiện rõ rệt trong huyết thanh của bệnh nhân ung thư tuyến tiền liệt so với nhóm không ung thư và có khả năng phân biệt hai nhóm với hiệu năng phân loại cao trong bộ dữ liệu được phân tích. Các mô hình phân loại dựa trên mức biểu hiện của miRNA này cho thấy

tiềm năng của hsa-miR-1203 như một dấu ấn sinh học chẩn đoán. Tuy nhiên, do nghiên cứu được thực hiện trên dữ liệu công khai từ một bộ dữ liệu duy nhất và chưa có bước kiểm định trên các tập dữ liệu độc lập, các kết quả này cần được diễn giải một cách thận trọng. Trong tương lai, các nghiên cứu trên các quần thể bệnh nhân độc lập, cũng như các nghiên cứu tiền cứu và phân tích chức năng sinh học, là cần thiết để xác nhận thêm giá trị chẩn đoán và làm rõ vai trò sinh học của hsa-miR-1203 trong ung thư tuyến tiền liệt.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ kinh phí bởi Đại học Y Dược Thành phố Hồ Chí Minh theo Hợp đồng số 498/2025/HĐ-ĐHYD, ngày 30/09/2025.

Nhóm tác giả xin trân trọng cảm ơn các tác giả và đơn vị đã công bố dữ liệu trên cơ sở dữ liệu Gene Expression Omnibus (GEO), tạo điều kiện cho việc khai thác và phân tích dữ liệu trong nghiên cứu này.

Các tác giả cam kết không có xung đột lợi ích liên quan đến nội dung và kết quả nghiên cứu.

TÀI LIỆU THAM KHẢO

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021; 71(3): 209-249. doi:10.3322/caac.21660.

2. Van Dong H, Lee AH, Nga NH, et al. Epidemiology and prevention of prostate cancer in Vietnam. *Asian Pac J Cancer Prev APJCP*. 2014; 15(22): 9747-9751. doi:10.7314/apjcp.2014.15.22.9747.

3. Eala MA, Dee EC, Jacomina LE, et al. Prostate Cancer in Southeast Asia: An Analysis of 2022 Incidence and Mortality Data. *Int J Radiat Oncol*. 2024; 120(2, Supplement):e528. doi:10.1016/j.ijrobp.2024.07.1170.

4. Moyer VA, U.S. Preventive Services Task Force. Screening for prostate cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2012; 157(2): 120-134. doi:10.7326/0003-4819-157-2-201207170-00459.

5. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 2004; 116(2): 281-297. doi:10.1016/s0092-8674(04)00045-5.

6. Nguyen MT, Quang MT. Integrated Bioinformatics Analysis of hsa-miR-4783-3p Target Genes and Functions in Prostate Cancer. *Pharm Sci Asia*. 2024; 51(3): 233-240. doi:10.29090/psa.2024.03.24.ap0911.

7. Selth LA, Townley S, Gillis JL, et al. Discovery of circulating microRNAs associated with human prostate cancer using a mouse model of disease. *Int J Cancer*. 2012; 131(3): 652-661. doi:10.1002/ijc.26405.

8. Chen X, Ba Y, Ma L, et al. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res*. 2008; 18(10): 997-1006. doi:10.1038/cr.2008.282.

9. Mitchell PS, Parkin RK, Kroh EM, et al. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A*. 2008; 105(30): 10513-10518. doi:10.1073/pnas.0804549105.

10. Quang MT, Nguyen MN, Than VT. The role and regulation of cell death in cancer. *Prog Mol Biol Transl Sci*. 2025; 217: 135-161. doi:10.1016/bs.pmbts.2025.06.014.

11. Ghamlouche F, Yehya A, Zeid Y, et al. MicroRNAs as clinical tools for diagnosis, prognosis, and therapy in prostate cancer. *Transl Oncol*. 2023; 28: 101613. doi:10.1016/j.tranon.2022.101613.

12. Luo X, Wen W. MicroRNA in prostate

cancer: from biogenesis to applicative potential. *BMC Urol*. 2024; 24(1): 244. doi:10.1186/s12894-024-01634-1.

13. Urabe F, Matsuzaki J, Yamamoto Y, et al. Large-scale Circulating microRNA Profiling for the Liquid Biopsy of Prostate Cancer. *Clin Cancer Res Off J Am Assoc Cancer Res*. 2019; 25(10): 3016-3025. doi:10.1158/1078-0432.CCR-18-2849.

14. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002; 30(1): 207-210. doi:10.1093/nar/30.1.207.

15. Deo RC. Machine Learning in Medicine. *Circulation*. 2015; 132(20): 1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593.

16. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015; 13: 8-17. doi:10.1016/j.csbj.2014.11.005.

17. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007; 23(14): 1846-1847. doi:10.1093/bioinformatics/btm254.

18. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015; 43(7): e47. doi:10.1093/nar/gkv007.

19. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006; 27(8): 861-874. doi:10.1016/j.patrec.2005.10.010.

20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44(3): 837-845.

21. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008; 19(5): 640-648. doi:10.1097/EDE.0b013e31818131e7.

22. Agarwal V, Bell GW, Nam JW, et al. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. 2015; 4:e05005. doi:10.7554/eLife.05005.

23. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000; 28(1): 27-30. doi:10.1093/nar/28.1.27.

24. Lao DT, Quang MT, Le TAH. The Role of hsa-miR-21 and Its Target Genes Involved in Nasopharyngeal Carcinoma. *Asian Pac J Cancer Prev APJCP*. 2021; 22(12): 4075-4083. doi:10.31557/APJCP.2021.22.12.4075.

25. Merriel SWD, Pocock L, Gilbert E, et al. Systematic review and meta-analysis of the diagnostic accuracy of prostate-specific antigen (PSA) for the detection of prostate cancer in symptomatic patients. *BMC Med*. 2022; 20:54. doi:10.1186/s12916-021-02230-y.

Summary

EVALUATION OF CIRCULATING HSA-MIR-1203 AS A DIAGNOSTIC BIOMARKER FOR PROSTATE CANCER USING MACHINE LEARNING APPROACHES

Prostate cancer is one of the most common malignancies among men, and the identification of reliable circulating biomarkers remains essential for improving diagnostic accuracy. This study aimed to evaluate the diagnostic potential of circulating hsa-miR-1203 in distinguishing prostate cancer patients from non-cancer controls using machine learning models. MicroRNA expression data were obtained from the GSE211692 dataset in the Gene Expression Omnibus, comprising 1,027 prostate cancer serum samples and 5,893 non-cancer samples. After preprocessing and $\log_2$ transformation, the dataset was divided into training (70%) and internal testing (30%) sets using stratified sampling. hsa-miR-1203 expression was significantly decreased in prostate cancer samples ($\log_2$FC = -3.77; $p < 0.001$). Four classification algorithms, including Extra Trees, Support Vector Machine (RBF kernel), AdaBoost, and Gaussian Naive Bayes, demonstrated excellent discriminative performance on the testing set, with area under the ROC curve (AUC) values approaching 0.98. These findings suggest that circulating hsa-miR-1203 may serve as a promising non-invasive biomarker to support prostate cancer diagnosis.

Keywords: Prostate cancer, microRNA, hsa-miR-1203, machine learning, circulating biomarker, diagnosis.