Ứng dụng mô hình học máy vào sàng lọc ảo các chất ức chế HIV integrase

Phan Tiểu Long, Trần Đình Xuân Trúc,
Trịnh Chương, Lê Lại Hoàng Sơn, Nguyễn Quí Hiển,
Huỳnh Hoàng Thúc, Trương Ngọc Tuyền*
Khoa Dược, Đại học Y Dược Thành phố Hồ Chí Minh

Summary

Applications of machine learning in drug design is an emerging and fast-growing field of research. In silico models allow the speeding up of drug discovery and developments. HIV, once known as "the disease of the century", is currently with no specific treatment. New drugs discovered by machine learning may have the potential to complement ARV-therapy to help prolong the lifespan of people who live with HIV. The machine learning model uses the multi-layer perceptron (MLP) architecture along with the Extra Trees algorithm for feature selection, and handling imbalanced data with the Tomek Links algorithm gave the results: 95.3% accuracy, 82.6% sensitivity, 86.5% precision, F1 score of 0.845, area under the ROC curve (0.953) and Average Precision (0.917). The internal dataset was screened through the PAINS filter and QSAR machine learning models and substance DI081 was determined to be the best candidate with the probability of inhibiting HIV integrase up to 98.57%.

Keywords: *QSAR, machine learning, HIV integrase, area under curve of ROC, F1 score.*

Đặt vấn đề

Việc ứng dụng khoa học máy tính và trí tuệ nhân tạo vào lĩnh vực thiết kế thuốc *in silico* không những rút ngắn được thời gian, công sức và tài nguyên mà còn tăng cường hiệu quả sàng lọc. Từ đó có thể nghiên cứu và phát triển thuốc mới nhanh chóng, kết quả đã gặt hái được nhiều thành công như xây dựng mô hình dự đoán hoạt tính kháng ung thư, độc tính, khả năng đáp ứng của thuốc [1]…

Chịu trách nhiệm: Trương Ngọc Tuyền
Email: truongtuyen@ump.edu.vn
Ngày nhận: 19/4/2022
Ngày phản biện: 14/6/2022
Ngày duyệt bài: 25/7/2022

Theo thống kê từ "Chương trình phối hợp của Liên Hợp Quốc về HIV/AIDS" (UNAIDS) năm 2020, có hơn 37,6 triệu người trên toàn thế giới đang sống cùng với căn bệnh thế kỷ. Hiện nay vẫn chưa có thuốc đặc trị hay vaccin, người nhiễm HIV chỉ có thể điều trị bằng liệu pháp kháng virus bằng thuốc ARV (Antiretroviral), nhằm ức chế các triệu chứng và giữ cho sự nhiễm trùng không phát triển thành AIDS [2]. Enzym integrase (IN) là một loại enzym được sản xuất bởi virus phiên mã ngược, tích hợp tạo thành các liên kết cộng hóa trị giữa thông tin di truyền của virus với tế bào vật chủ mà nó lây nhiễm [3]. Chức năng chính của IN là tích hợp ADN của virus vào vật chất di truyền của vật chủ, một bước rất cần thiết cho quá trình nhân lên của virus HIV. Chính vì vậy, việc ức chế enzym integrase, cụ thể là trong quá trình

chuyển sợi, có thể ngăn chặn sự lây lan của virus, kéo dài thời gian sống cho vật chủ. Các chất ức chế integrase thường được phối hợp với các loại thuốc điều trị HIV khác để giảm thiểu sự kháng thuốc [4].

Đối tượng và phương pháp nghiên cứu

Thiết bị và phần mềm

Nghiên cứu *in silico* trong đề tài được tiến hành trên máy tính với Mainboard B550 AORUS PRO, CPU AMD Ryzen 9 3900X 12-Core Processor 3.79 GHz, Ram 32 GB, card đồ họa VGA Radeon RX6900XT 16 GB, hệ điều hành Linux 20.04 64 bit và các phần mềm: ChemDraw 18.1, Python 3.9.7 cùng với thư viện học máy Scikit-learn [5].

Dữ liệu xây dựng mô hình

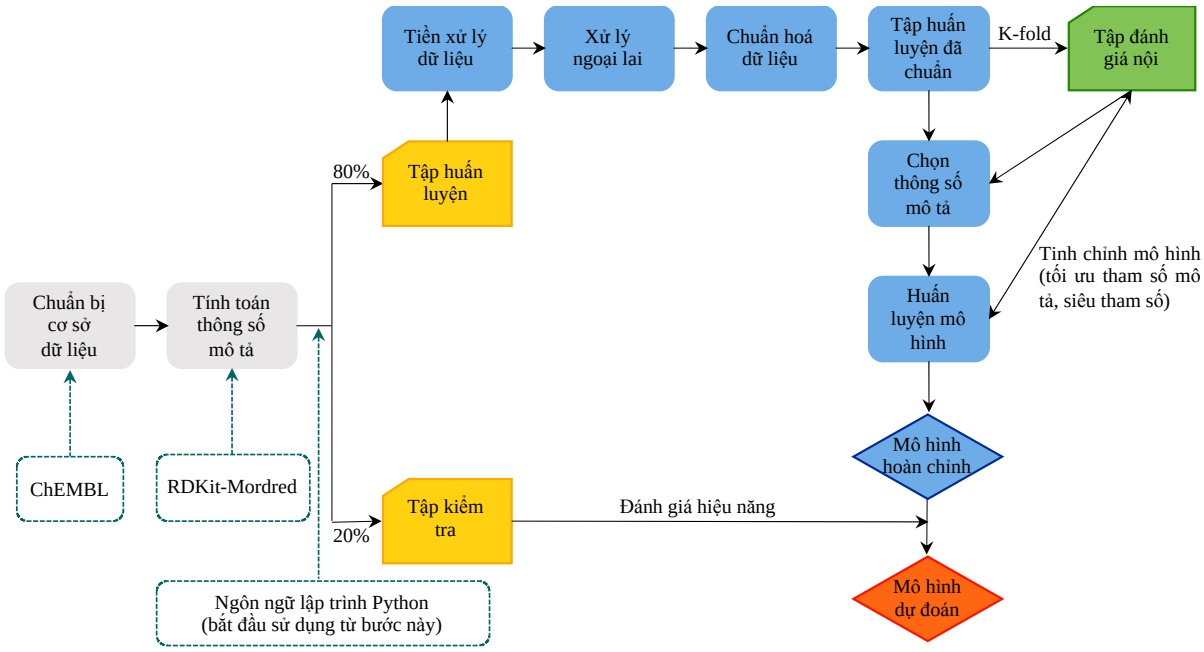
Các cấu trúc hai chiều (2D) được lấy từ bộ dữ liệu trên thư viện ChEMBL Database, sau đó được xử lý để chọn lựa các cấu trúc có hoạt tính ức chế HIV integrase với cùng phương pháp đo hoặc phương pháp đo gần giống nhau. Tập dữ liệu lớn và đa dạng về cấu trúc gồm 5685 chất có hoạt tính ức chế integrase từ thư viện ChEMBL. Trong đó, tất cả các chất đều được xác định là có IC50 dao động từ 0,46 nM đến 10000 nM. Tuy nhiên, tập dữ liệu này được thu thập từ nhiều nguồn khác nhau với các đích tác động, cũng như phương pháp thử riêng biệt nên tập dữ liệu cần được xử lý thô để chuẩn hoá về cùng hệ quy chiếu. Phép chuẩn hoá được thực hiện bởi thư viện ChEMBL Database bằng cách xấp xỉ các giá trị thực nghiệm giữa các phép đo khác nhau thành giá trị "pChEMBL value". Giá trị này cũng đã được nhiều nghiên cứu trên các tạp chí uy tín chấp thuận sử dụng [6].

Dữ liệu sàng lọc

Thư viện nội bộ bao gồm 1.000 chất chưa được công bố trên SciFinder (truy cập vào ngày 20/03/2021) có thể tổng hợp được trong điều kiện của Phòng thí nghiệm Hoá Hữu Cơ, Khoa Dược, Đại học Y Dược Thành phố Hồ Chí Minh bao gồm các khung dị vòng pyrazol, oxadiazol, thiadizol, benzotriazol, benzimidazol.

Phương pháp nghiên cứu

Quy trình xây dựng mô hình 2D–QSAR của đề tài được thể hiện tóm tắt ở hình 1. Các giá trị ngẫu nhiên (random_state) trong nghiên cứu đều được để về số 42 để đảm bảo tính lặp lại của các thử nghiệm.



Hình 1. *Quy trình xây dựng mô hình 2D-QSAR ứng dụng học máy*

Chuẩn bị cơ sở dữ liệu

Các cấu trúc hai chiều (2D) được lấy từ thư viện ChEMBL Database, sau đó được sàng lọc ra để chọn lựa các cấu trúc có hoạt tính ức chế HIV integrase với cùng phương pháp đo hoặc phương pháp đo tương đồng với nhau. Nghiên cứu sử dụng ngôn ngữ lập trình python và thư viện RDKit để tối thiểu hoá năng lượng, tính toán 1.613 thông số mô tả cấu trúc hai chiều từ thư viện Mordred [7].

Phân chia tập dữ liệu

Nghiên cứu xây dựng mô hình phân loại nên phải biến đổi giá trị "pChEMBL Value" sang hệ nhị phân, các dữ liệu được gán nhãn "1" cho "chất có hoạt tính" và "0" cho "chất không có hoạt tính" dựa trên ngưỡng phân loại được chọn. Trong phạm vi của nghiên cứu, chọn ngưỡng hoạt tính là 7 (tương đương IC50 = 100 nM).

Tập dữ liệu được phân chia thành tập huấn luyện (training set) và tập kiểm tra (test set) theo tỷ lệ 80 : 20 theo nguyên lý Pareto. Dữ liệu được chia theo nguyên tắc "phân tầng", bằng thư viện Scikit-learn.

Tiền xử lý dữ liệu

Tập huấn luyện được xử lý lần lượt qua các bước:

+ Loại bỏ dữ liệu trùng lặp và các thông số mô tả trùng lặp

+ Xử lý dữ liệu bị mất:

- Đối với cột: Nếu tỷ lệ dữ liệu bị thiếu trong 1 cột lớn hơn 50% số lượng dữ liệu trong cột thì sẽ tiến hành xoá cột.

- Đối với hàng: Xử lý bằng thuật toán học máy láng giềng gần nhất, KNNImputer từ thư viện Scikit-learn.

+ Loại bỏ các thông số có phương sai thấp: Tính toán phương sai của từng thông số, chọn ngưỡng phương sai là 0,05.

Xử lý ngoại lai

Tập huấn luyện tiếp tục được xử lý dữ liệu ngoại lai, gồm có xử lý ngoại lai đơn biến (univariate outliers) và ngoại lai đa biến (multivariate outliers), lần lượt qua các bước:

+ Xử lý ngoại lai đơn biến:

- Ánh xạ dữ liệu đến phân phối chuẩn hoặc phân phối đều bằng công cụ QuantileTransformer từ thư viện Scikit-learn.

- Các thông số mô tả có chứa ngoại lai được biến đổi thành biến định tính bằng công cụ KbinDiscretizer từ thư viện Scikit-learn. Đây là quá trình biến đổi một biến định lượng thành một tập hợp gồm hai hay nhiều nhóm định tính (còn gọi là danh mục).

+ Xử lý ngoại lai đa biến:

- Lựa chọn công cụ xử lý ngoại lai để hạn chế lượng dữ liệu phải xoá mà vẫn xử lý ngoại lai đa biến hiệu quả.

- Nghiên cứu sử dụng công cụ Local Outlier Factor (yếu tố ngoại lai cục bộ), xác định ngoại lai dựa trên mật độ cục bộ của từng điểm dữ liệu. Local Outlier Factor (LOF – yếu tố ngoại lai cục bộ) được xác định dựa trên mật độ cục bộ (local density).

Chuẩn hoá dữ liệu

Tập huấn luyện sau khi xử lý ngoại lai sẽ được chuẩn hoá về khoảng [0,1] bằng MinMaxScaler với công thức $y = \frac{x - \min}{\max - \min}$.

Lựa chọn thông số mô tả

Dữ liệu có số chiều ban đầu khá lớn với 1613 thông số mô tả, nên việc sử dụng hết các thông số mô tả này để xây dựng mô hình là điều bất khả thi. Vì vậy, việc lựa chọn các thông số mô tả đặc trưng nhất cho việc xây dựng mô hình là điều cần thiết. Phương pháp lựa chọn thông số mô tả của đề tài được chọn bằng cách so sánh hiệu năng của các thuật toán học máy, thông qua đánh giá chéo nội gập 10 lần, lặp lại 3 lần (RepeatedStratifiedKFold). So sánh kết quả F1 score của các mô hình bằng điểm trung bình và độ lệch chuẩn của điểm đánh giá chéo được tính toán bằng hàm cross_val_score trong thư viện Scikit-learn. F1-score, hay còn gọi là F-measure, f-score là trung bình điều hòa (Harmonic mean) của độ chính xác (precision) và độ nhạy (recall), là thông số phổ biến nhất

dùng trong đánh giá mô hình với bộ dữ liệu mất cân bằng [8]. Nếu chỉ dùng độ chính xác và độ nhạy để đánh giá thì không thể diễn tả hết toàn bộ khả năng của mô hình, chúng ta có thể có thể độ chính xác cực kỳ cao nhưng độ nhạy (recall) lại thấp và ngược lại. Do đó, F1 score giúp chúng ta giải quyết vấn đề này [9].

$$\textbf{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Lựa chọn mô hình học máy

Để có thể tận dụng được hết hiệu năng của các mô hình học máy, nghiên cứu sử dụng phương pháp đánh giá chéo nội gập 10 lần, lặp lại 3 lần cho 15 mô hình học máy khác nhau với các thông số được để mặc định theo thư viện Scikit-learn. So sánh kết quả của các mô hình bằng điểm trung bình và độ lệch chuẩn của điểm đánh giá chéo tương tự như lựa chọn thông số mô tả

Xử lý mất cân bằng dữ liệu và tối ưu mô hình

Dữ liệu mất cân bằng là tình trạng khi tỷ lệ dữ liệu của các lớp chênh lệch lớn, việc này đặt ra một thách thức đối với mô hình dự đoán vì hầu hết các thuật toán học máy sử dụng để phân loại được thiết kế dựa trên giả định về số lượng mẫu quan sát bằng nhau cho mỗi lớp. Điều này dẫn đến các mô hình có hiệu suất dự đoán kém, đặc biệt là với lớp thiểu số, mặc dù đây là lớp thường được quan tâm nhiều hơn. Nghiên cứu sử dụng phương pháp "lấy mẫu dữ liệu" (Data sampling) để xử lý tình trạng mất cân bằng này, bao gồm các phương pháp:

+ Oversampling: Là các phương pháp gia tăng kích thước mẫu thuộc nhóm thiểu số.

+ Undersampling: Là việc giảm số lượng các quan sát của nhóm đa số để nó trở nên cân bằng với số quan sát của nhóm thiểu số.

+ Kết hợp Oversampling và Undersampling.

Đánh giá khả năng tổng quát hoá của mô hình

Đánh giá khả năng tổng quát hoá của mô hình bằng tập đánh giá ngoại (20%), với các đại lượng tương tự với đánh giá nội.

Sàng lọc ảo

Dữ liệu sàng lọc lần lượt được đi qua lưới lọc PAINS và mô hình 2D-QSAR.

Lưới lọc PAINS loại bỏ những hợp chất thường xuất hiện như là một hit trong sàng lọc, chúng thực chất là chất dương tính giả. Các hợp chất này cho thấy có tác động ở nhiều mục tiêu hơn là một mục tiêu cụ thể, điều này được giải thích do sự gắn kết không đặc hiệu hoặc sự tương tác của những chất này với những thành phần trong quá trình định lượng. Baell và CS. đã nghiên cứu những cấu trúc thứ cấp can thiệp vào tín hiệu định lượng. Họ mô tả những cấu trúc thứ cấp này có thể giúp xác định PAINS và đưa ra danh sách có thể được sử dụng để lọc cấu trúc thứ cấp [10].

Kết quả

Xây dựng mô hình 2D-QSAR

Tiền xử lý dữ liệu

5685 dữ liệu từ thư viện ChEMBL được lọc theo giá trị "pChEMBL Value" còn lại 2296 chất và được chia thành 1700 chất của tập huấn luyện và 426 chất của tập kiểm tra. Tỷ lệ mất chênh lệch của nhóm có hoạt tính so với nhóm không hoạt tính là 1:2,9 ở cả tập huấn luyện và kiểm tra.

1613 thông số mô tả được tính toán bằng thư viện Mordred, sau khi qua bước kiểm tra cột có dữ liệu giống nhau còn 1245 cột (bao gồm cả giá trị "pChEMBL Value").

Sau khi loại bỏ các cột có dữ liệu thiếu hơn 50%, còn lại 982 cột. Kiểm tra lại dữ liệu bị thiếu, xác định được có 821 chất của tập huấn luyện và 203 chất của tập kiểm tra chứa dữ liệu bị thiếu, sử dụng công cụ KNNImputer để điền vào các dữ liệu bị thiếu đó.

Thực hiện phân tích ngưỡng phương sai các thông số mô tả, loại 355 thông số có ngưỡng phương sai dưới 0,05.

Xử lý ngoại lai

Tập dữ liệu được kiểm tra ngoại lai bằng phương pháp bách phân vị.

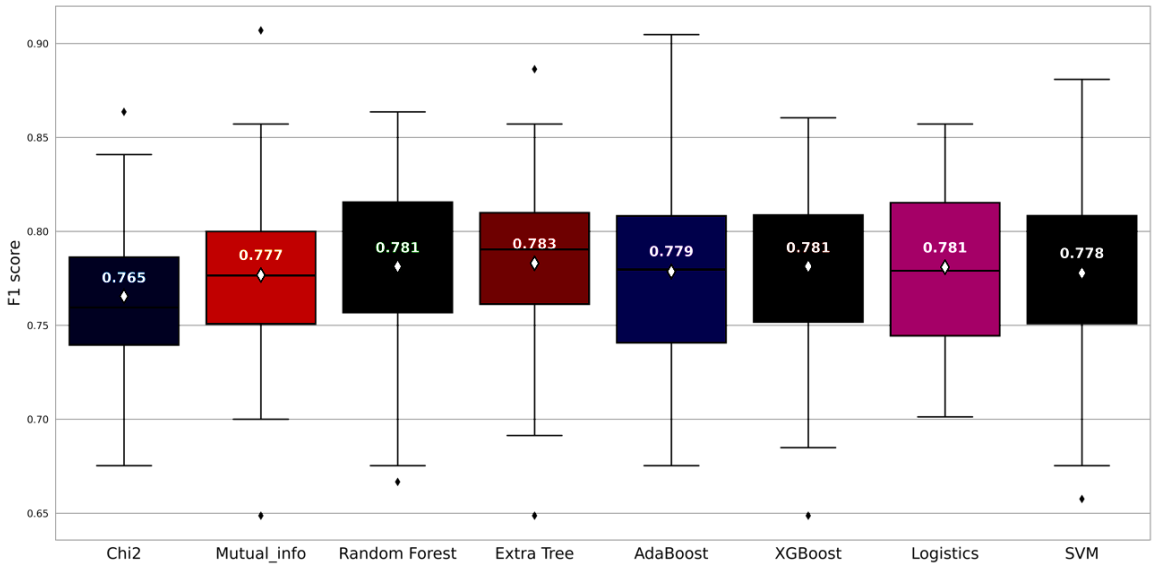
Tuy nhiên, việc xử lý ngoại lai bằng cách xoá các dữ liệu thường dẫn đến thiếu hụt dữ liệu nghiêm trọng, nên nghiên cứu sử dụng phương pháp ánh xạ dữ liệu đến phân phối đều QuantileTransformer. Sau khi ánh xạ sang phân phối đều, vẫn có 22 đặt thông số mô tả xuất hiện dữ liệu ngoại lai, nghiên cứu tiến hành xử lý bằng công cụ KbinDiscretizer. Kết quả dữ liệu được tinh sạch, không còn giá trị ngoại lai.

Tiếp tục xử lý ngoại lai bằng phương pháp phân tích đa biến, nhân tố ngoại lai cục bộ Local Outlier Factor. Kết quả loại 16 chất từ tập huấn luyện và 10 chất từ tập kiểm tra.

Lựa chọn thông số mô tả

Sau khi xử lý ngoại lai, dữ liệu sẽ được chuẩn hoá bằng phương pháp MinMaxScaler và tiến hành đánh giá chéo nội để chọn ra thuật toán chọn thông số mô tả phù hợp. Trong đó, phương pháp lọc thông số mô tả bằng thống kê được sử dụng là chi bình phương (Chi2) và thông tin tương hỗ (Mutual_info). Phương pháp lựa chọn nội tại sử dụng các thuật toán như rừng ngẫu nhiên (Random Forest), Extra Tree, AdaBoost, XGBoost, hồi quy Logic (Logistics) và máy vector hỗ trợ (SVM). Nghiên cứu sử dụng đại lượng F1 score làm tiêu chí chính khi so sánh các phương pháp với nhau, với kết quả được mô tả như Hình 2.

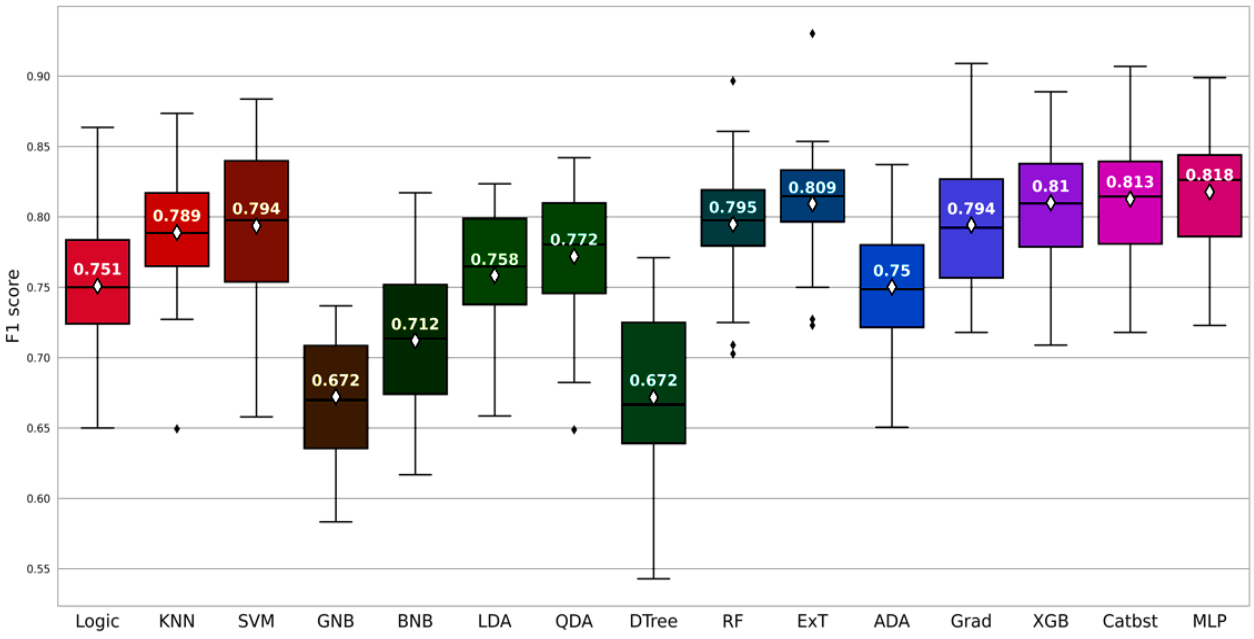
Kết quả là thuật toán Extra Tree cho kết quả tối ưu nhất, vì trung bình F1 score của 30 lần đánh giá nội cao nhất (0,783 ± 0,050). Nghiên cứu chọn thuật toán này và thu được 154 thông số mô tả.



Hình 2. *So sánh các phương pháp lựa chọn thông số mô tả*

Lựa chọn mô hình học máy

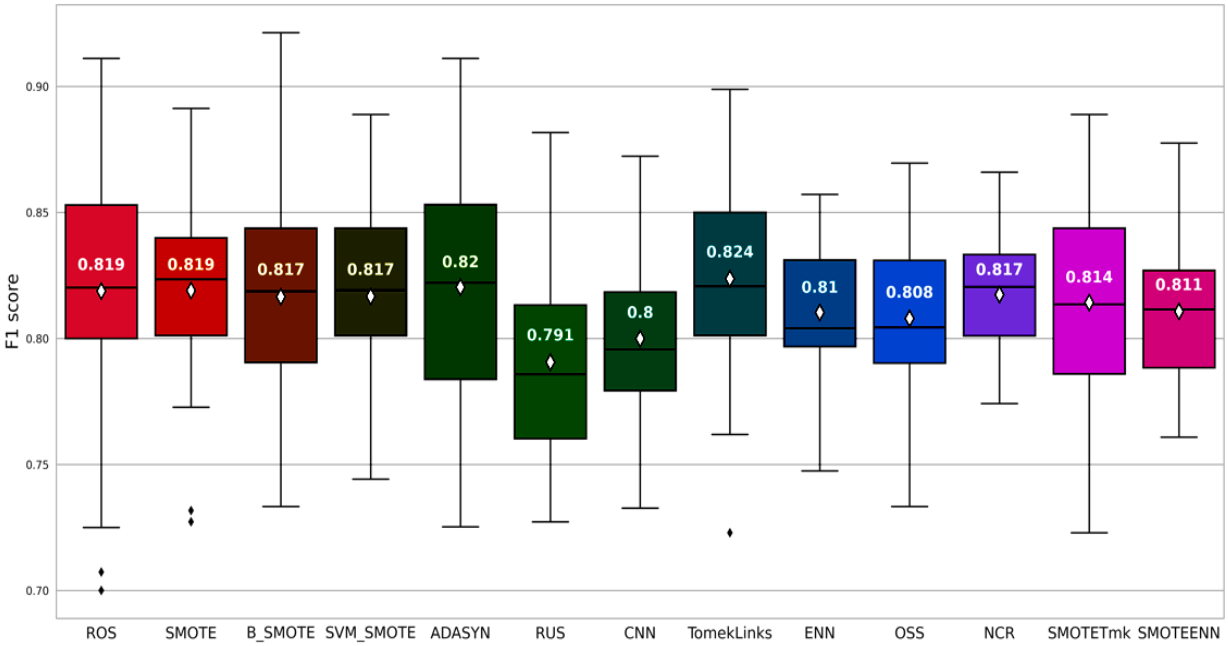
Nghiên cứu sử dụng 15 mô hình học máy khác nhau để thực hiện đánh giá chéo nội nhằm lựa chọn thuật toán tốt nhất, bao gồm các thuật toán như hồi quy logic (Logic), láng giềng gần nhất (KNN), máy vector hỗ trợ (SVM), Gaussion Naive bayes (GNB), Bernouli Naive bayes (BNB), phân tích phân biệt tuyến tính (LDA), phân tích phân biệt bậc 2 (QDA), cây quyết định (Dtree), rừng ngẫu nhiên (RF), Extra Tree (ExT), Adaboost (ADA), Gradient Boosting (Grad), XGboost (XGB), Catboost (Catbst) và mạng nơron truyền thẳng nhiều lớp (MLP), kết quả được mô tả tổng quát trong hình 3. Nghiên cứu lựa chọn mô hình MLP vì mô hình này cho kết quả trung bình F1 score đánh giá là cao nhất (0,818 ± 0,043).



Hình 3. *So sánh các thuật toán xây dựng mô hình*

Xử lý mất cân bằng dữ liệu và tối ưu mô hình

Nghiên cứu sử dụng 5 kỹ thuật Oversampling, 6 kỹ thuật Undersampling và 2 kỹ thuật kết hợp để chọn ra kỹ thuật phù hợp cho bộ dữ liệu này. 13 thuật toán sampling kết hợp cùng mô hình MLP đã được lựa chọn, tiến hành đánh giá chéo nội để tìm ra kỹ thuật phù hợp nhất. Kết quả được mô tả như hình bên dưới, kỹ thuật Tomek Links cho kết quả tốt nhất (0,824 ± 0,043).



Hình 4. *So sánh các phương pháp xử lý mất cân bằng dữ liệu*

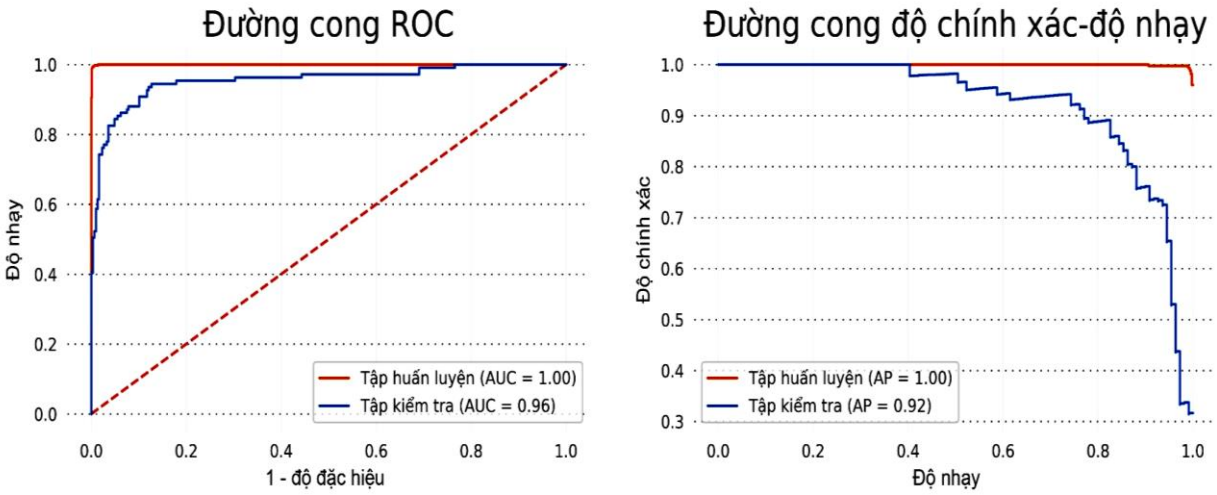
Đánh giá khả năng tổng quát hoá của mô hình

Thực hiện đánh giá ngoại với mô hình đã tối ưu và so sánh với 12 thuật toán khác. Kết quả tương tự như đánh giá nội, mô hình MLP cho kết quả tối ưu hơn so với các thuật toán về hầu hết các chỉ số như F1 score, Average precision, ROC AUC.

Bảng 1. *Kết quả đánh giá ngoại với 12 thuật toán khác nhau*

	AUC ROC	***Average precision***	***Accuracy***	***Recall***	***Precision***	***F1 score***
Hồi quy Logic	0,931	0,872	0,875	0,734	0,777	0,755
Láng giềng gần	0,943	0,854	0,906	0,798	0,837	0,817
Máy vector hỗ trợ	0,913	0,854	0,877	0,725	0,790	0,756
Cây quyết định	0,740	0,548	0,841	0,569	0,765	0,653
Rừng ngẫu nhiên	0,952	0,907	0,911	0,725	0,919	0,810
Extra Tree	0,962	0,910	0,904	0,761	0,856	0,806
Ada Boost	0,933	0,870	0,894	0,743	0,835	0,786
Gradient Boost	0,956	0,905	0,901	0,752	0,854	0,80
XGBoost	0,953	0,905	0,887	0,688	0,852	0,761
CAT Boost	0,959	0,916	0,911	0,752	0,891	0,816
MLP	0,955	0,917	0,921	0,826	0,865	0,845
Gaussian	0,831	0,627	0,796	0,743	0,587	0,656
Bernoulli	0,885	0,775	0,815	0,771	0,618	0,686

Ngoại trừ đại lượng F1 score, các đại lượng như Average precesion và ROC AUC cũng phù hợp trong việc đánh giá mô hình mất cân bằng. Kết quả thu được, điểm AP (trung bình độ chính xác) có giá trị 0,92, điểm ROC AUC có giá trị 0,955, đều cao hơn 0,9 và chênh lệch không quá 0,1 đơn vị so với tập huấn luyện, chứng tỏ mô hình không bị quá khớp khi huấn luyện.

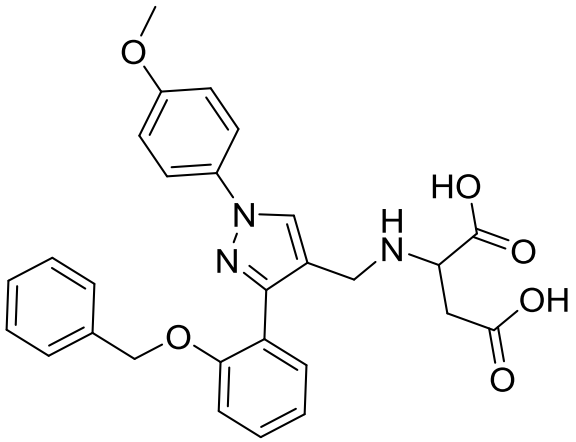


Hình 5. *Kết quả đường cong ROC và đường cong Precision Recall*

Sàng lọc ảo

Sàng lọc qua lưới lọc PAINS: 199 cấu trúc khung pyrazol thuộc bộ dữ liệu nội bộ được xác định là không thuộc danh sách PAINS.

Sàng lọc qua mô hình MLP-QSAR. Kết quả sàng lọc có 5 chất tiềm năng ức chế HIV integrase, trong đó chỉ có cấu trúc **DI081** có xác suất dự đoán cho hoạt tính kháng HIV integrase lớn hơn 90% (98,57%).



Hình 6. *Cấu trúc của chất DI081*

Bàn luận

Ứng dụng khai phá dữ liệu (data mining) vào việc xây dựng mô hình đã giúp cải thiện hiệu năng của mô hình đáng kể. Thông thường, các dữ liệu ngoại lai trong mô hình QSAR sẽ bị xoá bỏ do không có phương pháp xử lý phù hợp, nhưng việc sử dụng phương pháp ánh xạ dữ liệu đến phân phối đều sẽ giải quyết nhanh chóng các dữ liệu ngoại lai đó. Phương pháp này có tác dụng giảm tác động của giá trị ngoại lai, do đó đây là một chu trình tiền xử lý mạnh mẽ. Sự biến đổi này được áp dụng trên từng thông số mô tả một cách độc lập. Đầu tiên, ước tính về hàm phân phối tích luỹ (CDF) của một thông số mô tả cấu trúc được sử dụng để ánh xạ các giá trị ban đầu thành một phân phối đều. Các giá trị thu được sau đó được ánh xạ tới phân phối đầu ra mong muốn với hàm lượng tử liên quan. Ngoài ra, các thông số mô tả của dữ liệu mới hoặc nằm dưới hoặc trên phạm vi phù hợp (miền ứng dụng), sẽ được ánh xạ tới các giới hạn của phân phối đầu ra. Chính vì vậy, việc áp dụng ánh xạ tới phân phối đều sẽ bỏ qua việc xác định miền ứng dụng của mô hình, vì các giá trị nằm ngoài miền ứng dụng sẽ được ánh xạ thành đúng các giới hạn trên và dưới của phân phối đầu ra.

Nghiên cứu sử dụng phương pháp Intrinsic để lựa chọn thông số mô tả, thay vì các phương pháp chọn thông số mô tả dựa trên tương quan (CFS) thực hiện bằng phần mềm Weka. Phương pháp Intrinsic là thuật toán thực hiện việc lựa chọn thông số mô tả trong quá trình huấn luyện mô hình. Ưu điểm của phương pháp này là cho kết quả tốt hơn phương pháp CFS với thời gian tương tự. Kết quả đã chọn ra được thuật toán Extra Trees (F1 score = 0,783 ± 0,050) là phù hợp nhất để lựa chọn thông số mô tả cho bộ dữ liệu này. Phương pháp đánh giá nội để chọn ra thuật toán phù hợp để xây dựng mô hình là mạng nơron truyền thẳng nhiều lớp (MLP) (F1 score = 0,818 ± 0,043). Sau đó, nhận thấy dữ liệu mất cân bằng mức độ trung bình (tỷ lệ mất cân băng là 34,6%), nên việc xử lý dữ liệu mất cân bằng này bằng các thuật toán như SMOTE, Tomek Links,… có thể giúp gia tăng hiệu năng của mô hình. Và kết quả so sánh đã chọn ra phương pháp Tomek Links (F1 score = 0,824 ± 0,043).

So sánh hiệu năng của mô hình khi trải qua các phương pháp xử lý dữ liệu, nhận thấy F1 score của mô hình tăng lên đáng kể, từ 0,783 khi chọn lọc thông số mô tả cấu trúc bằng Extra Trees, lên 0,818 khi chọn được mô hình MLP, tăng lên 0,824 khi sử dụng Tomek Links để xử lý mất cân bằng. Cuối cùng, để khẳng định khả năng tổng quát hoá của mô hình, đánh giá ngoại đã được thực hiện và cho kết quả rất tốt là 0,845, không chênh lệch quá nhiều khi so với phương pháp đánh giá nội.

Bảng 2. *So sánh kết quả nhóm thực hiện với nghiên cứu của Zhou và CS.* [11]

	Zhou và CS.	***Nghiên cứu này***
Số chất	1785	2293
Mốc hoạt tính	4600 μM	100 nM
Tỷ lệ mất cân bằng dữ liệu	0,618	0,346
Số lượng thông số	8	154
Mô hình tốt nhất	Naive Bayesian	Multilayer Perceptron
Chỉ số đánh giá	Độ chính xác = 88,3% Độ đặc hiệu = 87,2%	Độ chính xác = 86,5% Độ đặc hiệu = 94%

Khi so sánh nghiên cứu của nhóm thực hiện với của Zhou, kết quả của chúng tôi có phần tốt hơn dù tỷ lệ mất cân bằng dữ liệu cao hơn. Kết quả nghiên cứu của Zhou tuy khá tốt, các trị số trên 85% đối với thuật toán Naive Bayesian, tuy nhiên vẫn có một số vấn đề:

+ Mốc hoạt tính 4600 μM được chọn của Zhou là không thuyết phục. Các chất ức chế hiện có trên thị trường có IC50 dao động từ 1 – 10 nM, chính vì vậy mốc hoạt tính tối thiểu nên lấy phải là 100 nM (tương đương với pIC50 = 7).

+ Thứ hai, nghiên cứu của Zhou chỉ sử dụng 8 thông số mô tả, khó bao quát hết được đặc điểm phân tử của không gian hoá học, nên khả năng tổng quát hoá của mô hình cũng sẽ khiêm tốn. Chính vì vậy, thuật toán tốt nhất của tác giả Zhou là Naive Bayesian chỉ trình bày kết quả đánh giá nội, không phải kết quả đánh giá ngoại, nên khả năng khái quát hoá của mô hình này chưa được chứng minh.

+ Khi so sánh với Zhou, đề tài nghiên cứu này sử dụng mốc hoạt tính hợp lý hơn và chấp nhận tỷ lệ mất cân bằng dữ liệu cao, bù lại kết quả độ chính xác lại tốt hơn.

Tiến hành sàng lọc ảo qua lọc PAINS và mô hình QSAR đã chọn ra được 5 chất tiềm năng có hoạt tính (pIC50 ≥ 7), có một chất mã số **DI081** có xác suất dự đoán đúng lên đến 98,57%.

Kết luận

Nhóm nghiên cứu đã xây dựng thành công được mô hình QSAR bằng thuật toán MLP với các kết quả đánh giá ngoại đều cao như độ đúng 95,3%, độ nhạy 82,6%, độ chính xác 86,5% và F1 score đạt 0,845. Bên cạnh đó, các chỉ số diện tích dưới đường cong ROC (0,953) và trung bình độ chính xác (Average Precision) (0,917) cũng đạt được kết quả cao hơn 0,9 và không chênh lệch quá 0,1 đơn vị khi so với bộ huấn luyện. Nghiên cứu cũng đã sàng lọc ra được chất **DI081** có hoạt tính ức chế HIV intergrase với xác xuất dự đoán đúng lên đến 98,57%. Kết quả này là tiền để để tổng hợp và thử hoạt tính kháng HIV intergrase của hợp chất này.

Tài liệu tham khảo

1. Baptista D. et al. (2021), "Deep learning for drug response prediction in cancer", *Brief. Bioinform.*, 22 (1), pp. 360-379.

2. WHO (2021), *Publishes new clinical and service delivery recommendations for HIV*

prevention, treatment and care, who.int, https://www.who.int/news/item/17-03-2021-who-publishes-new-clinical-recommendations-on-hiv-prevention-infant-diagnosis-antiretroviral-therapy-initiation-and-monitoring.

3. Stern T. A. et al. (2010), "Massachusetts General Hospital handbook of general hospital psychiatry - ebook", *Elsevier Health Sciences*, pp. 353-370.

4. Pau A. K. et al. (2014), "Antiretroviral therapy: Current drugs", *Infectious Disease Clinics*, 28 (3), pp. 371-402.

5. Pedregosa F. et al. (2011), "Scikit-learn: Machine learning in Python", *The Journal of Machine Learning Research,* 12, pp. 2825-2830.

6. Li G. et al. (2020), "Discovery and optimization of novel pyrazolopyrimidines as potent and orally bioavailable allosteric HIV-1 integrase inhibitors", *Journal of Medicinal Chemistry,* 63 (5), pp. 2620-2637.

7. Moriwaki H. et al. (2018), "Mordred: A molecular descriptor calculator", *Journal of Cheminformatics*, 10 (1), pp. 1-14.

8. Haibo He Y. M. (2013), *Imbalanced learning: Foundations, algorithms, and application.*

9. Brownlee J. (2020), *Imbalanced classification with python: Choose better metrics, balance skewed classes, and apply cost-sensitive learning*.

10. Baell J. B. et al. (2010), "New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays", *Journal of Medicinal Chemistry*, 53 (7), pp. 2719-2740.

11. Zhou J. et al. (2021), "Classification and design of HIV-1 integrase inhibitors based on machine learning", *Computational and Mathematical Methods in Medicine*, 2021, pp. 1-11.