

**BỘ GIÁO DỤC ĐÀO TẠO
TRƯỜNG ĐẠI HỌC LẠC HỒNG**



DƯƠNG THỊ KIM CHI

**NÂNG CAO HIỆU QUẢ MÔ HÌNH HỌC MÁY CHO DỮ
LIỆU Y SINH**

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Ngành: Khoa học máy tính

Mã số ngành: **9480101**

Đồng Nai, năm 2023

Công trình được hoàn thành tại: Trường Đại học Lạc Hồng

Người hướng dẫn khoa học: PGS.TS. Trần Văn Lăng

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án cấp Trường họp tại

.....

.....

Vào hồi giờ....., ngày.....tháng.....năm

Có thể tìm hiểu luận án tại thư viện:

- Thư viện trường Đại học Lạc Hồng
- Thư viện Quốc Gia

MỤC LỤC

CHƯƠNG 1. TỔNG QUAN	1
1.1 TỔNG QUAN VỀ ĐỀ TÀI NGHIÊN CỨU.....	1
1.1.1 Giới thiệu	1
1.1.2 Bài toán nghiên cứu	1
1.1.3 Thách thức của bài toán nghiên cứu	2
1.2 MỤC TIÊU, ĐỐI TƯỢNG, PHẠM VI VÀ PHƯƠNG PHÁP NGHIÊN CỨU	2
1.2.1 Mục tiêu	2
1.2.2 Phạm vi nghiên cứu tập trung vào:	3
1.3 NHIỆM VỤ CỦA LUẬN ÁN.....	3
1.3.1 Thiết kế mô hình học máy hiệu quả cho dữ liệu sinh học phân tử trong các nhiệm vụ ứng dụng trong phát triển thuốc bằng kỹ thuật tái tổ hợp	4
1.3.2 Mô hình học máy hiệu quả cho dữ liệu sinh học phân tử trong các nhiệm vụ định danh loài sinh vật.....	4
1.3.3 Mô hình học máy hiệu quả trong các ứng dụng y sinh về chuẩn đoán bệnh dựa trên dữ liệu lâm sàng.	5
1.4 CÁC ĐÓNG GÓP CỦA LUẬN ÁN.....	6
1.5 BỐ CỤC CỦA LUẬN ÁN.....	7
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VÀ CÁC CÔNG TRÌNH LIÊN QUAN	9
2.1 CÁC KHÁI NIỆM TRONG Y SINH	9
2.2 CÁC NGHIÊN CỨU LIÊN QUAN CÓ SỬ DỤNG THUẬT TOÁN HỌC MÁY DÙNG TRONG CÁC BÀI TOÁN ĐỀ XUẤT	9
2.2.1 Các vấn đề về rút gọn chiều:	9
2.2.2 Phương pháp học tập không giám sát	9
2.2.3 Phương pháp học tập giám sát	9
2.2.4 Phương pháp học máy học kết hợp	10
2.2.5 Các thuật toán dùng trong luận án để giải quyết các vấn đề đã đặt ra trong dữ liệu y sinh.....	10
2.3 CÁC NGHIÊN CỨU LIÊN QUAN.....	10
2.4 DỮ LIỆU Y SINH ĐƯỢC SỬ DỤNG TRONG CÁC NHIỆM VỤ CỦA LUẬN ÁN	11
2.5 ĐÁNH GIÁ MÔ HÌNH MÔ HÌNH HỌC MÁY.....	11

CHƯƠNG 3. MÔ HÌNH HỌC MÁY TÌM GENE CHO HỆ THỐNG BIỂU HIỆN TRONG KỸ THUẬT DNA TÁI TỔ HỢP	12
3.1 BÀI TOÁN TÌM GENE BIỂU HIỆN CAO (HEG- HIGHLY EXPRESSED GENE).....	12
3.1.1 Kết quả thực nghiệm.....	12
3.2 BÀI TOÁN TÌM HỆ THỐNG BIỂU HIỆN PHÙ HỢP VỚI GENE MỤC TIÊU [CT3]	13
CHƯƠNG 4. MÔ HÌNH ĐỊNH DANH LOÀI SINH VẬT.....	13
4.1 GIỚI THIỆU VỀ ĐỊNH DANH LOÀI	13
4.2 XÂY DỰNG TẬP ĐỦ LIỆU CHO QUÁ TRÌNH HUẤN LUYỆN.....	14
4.3 THUẬT TOÁN ĐỀ XUẤT XÂY DỰNG MÔ HÌNH ĐỊNH DANH LOÀI DỰA TRÊN HỌC KẾT HỢP (ENSEMBLE LEARNING)	14
4.4 MÔ HÌNH HỌC MÁY ĐỊNH DANH LOÀI NĂM MỐI.....	15
4.5 KẾT QUẢ THỰC NGHIỆM	15
4.6 KẾT LUẬN:	17
CHƯƠNG 5. MÔ HÌNH HỌC MÁY CHO CHUẨN ĐOÁN BỆNH DỰA TRÊN DỮ LIỆU LÂM SÀNG.....	18
5.1 MÔ HÌNH DỰ ĐOÁN BỆNH DỰA TRÊN DỮ LIỆU LÂM SÀNG	18
5.1.1 Giới thiệu bài toán dự đoán bệnh và mô hình đề xuất	18
5.1.2 Kết quả thực nghiệm	19
5.2 MÔ HÌNH PHÂN LOẠI BỆNH CoViD-19 VÀ BỆNH CÚM MÙA.....	19
5.2.1 Giới thiệu bài toán và mô hình giải quyết.....	19
5.2.2 Hiệu năng của mô hình đề xuất.....	21
5.3 KẾT LUẬN.....	22
CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	23
6.1 KẾT LUẬN.....	23
6.2 HƯỚNG PHÁT TRIỂN	24

CHƯƠNG 1. TỔNG QUAN

1.1 Tổng quan về đề tài nghiên cứu

1.1.1 Giới thiệu

Tính toán y sinh (hay còn gọi là tin y sinh) là một lĩnh vực nghiên cứu liên ngành giữa y học và khoa học máy tính. Đó là sự kết hợp các phương pháp phân tích dữ liệu, học máy, thống kê và lý thuyết thông tin để giải quyết các vấn đề trong lĩnh vực y sinh như: phát hiện và chẩn đoán bệnh, thiết kế thuốc và nghiên cứu sinh học phân tử. Các nghiên cứu quan trọng của lĩnh vực tin y sinh như có thể kể đến là: dự báo dịch bệnh, phát triển thuốc, sản xuất vaccine, hay chẩn đoán và điều trị bệnh. Phương pháp học máy có thể thúc đẩy các chương trình nghiên cứu cơ bản và ứng dụng về tin y sinh [2]. Các mô hình này có thể giúp xác định nguy cơ mắc bệnh, phát hiện sớm bệnh lý và dự đoán kết quả của điều trị.

Để xây dựng các mô hình học máy có hiệu quả và có tính ứng dụng cao, hầu hết các thuật toán học máy đều cần dữ liệu cấu trúc đồng nhất, dữ liệu các thuộc tính đã được số hóa và không trống. Tuy nhiên, điều này là rất khó thực hiện đối với dữ liệu y sinh vì: i) phần lớn dữ liệu sinh học phân tử thì có cấu trúc dạng chuỗi ký tự dài cơ chế sinh học phức tạp và khối lượng dữ liệu rất lớn; ii) dữ liệu của bệnh viện như hồ sơ y tế, kết quả xét nghiệm y khoa của các loại bệnh thì bị phân mảnh, trùng lặp, bị thiếu và mất cân bằng lớp trong các bộ dữ liệu y sinh thế giới thực. Các thách thức này là động lực thúc đẩy luận án thực hiện nghiên cứu quan trọng này.

1.1.2 Bài toán nghiên cứu

Với nguồn dữ liệu y sinh được tạo ra ngày càng nhiều và được công bố rộng rãi bởi các dự án lớn về khoa học sự sống (Life science) đã thúc đẩy mạnh các nhà nghiên cứu về ứng dụng học máy trong lĩnh vực y sinh. Một trong các thách thức lớn của hướng nghiên cứu này là dữ liệu y sinh có đặc tính chuyên ngành phức tạp, dữ liệu chiều cao và chứa nhiều lỗi. Việc xử lý hiệu quả dữ liệu thô từ các nguồn dữ liệu y sinh giúp các thuật toán học

máy hoạt động tốt và nâng cao hiệu năng mô tả dữ liệu hay dự đoán phân loại [1, 2].

Mục tiêu chính của luận án là đề xuất các phương pháp tiếp cận mới cho bài toán "Ứng dụng mô hình học máy hiệu quả cho dữ liệu y sinh thông qua lựa chọn các giải pháp tiền xử lý hiệu quả hai loại dữ liệu sinh học phân tử và dữ liệu khám bệnh cận lâm sàng trong các thuật toán phân cụm và phân lớp cho các ứng dụng y sinh về *dự đoán nguồn gốc loài sinh vật, hỗ trợ điều chế vaccine/thuốc, và hỗ trợ chuẩn đoán bệnh*"

1.1.3 Thách thức của bài toán nghiên cứu

Việc ứng dụng các mô hình học máy cho dữ liệu y sinh đang hết sức cần thiết mang lại nhiều cơ hội được chữa bệnh tốt hơn, giảm tải tối ưu hóa được chi phí điều trị cũng như giảm bớt sự quá tải của hệ thống y tế công khi bùng phát dịch bệnh. Để đưa ra dự đoán chính xác, hầu hết các thuật toán học máy đều cần dữ liệu cấu trúc đồng nhất, dữ liệu các thuộc tính đã được số hóa và không trống. Tuy nhiên, điều này là rất khó thực hiện đối với dữ liệu y sinh vì:

- ❖ *Đối với dữ liệu sinh học phân tử*: dữ liệu dạng trình tự gene có số chiều rất lớn (hàng ngàn chiều), cơ chế sinh học phức tạp, và dữ liệu không cân bằng đều là các vấn đề lớn trong loại dữ liệu này, đây cũng là thách thức lớn của ứng dụng học máy cho bài toán y sinh trong lĩnh vực sản xuất thuốc.
- ❖ *Đối với dữ liệu lâm sàng*: Dữ liệu lâm sàng của bệnh nhân từ của bệnh viện như hồ sơ y tế, kết quả xét nghiệm y khoa của các loại bệnh thì bị phân mảnh, trùng lặp, bị thiếu, mất cân bằng đối với một số loại bệnh mới hiếm.

1.2 Mục tiêu, đối tượng, phạm vi và phương pháp nghiên cứu

1.2.1 Mục tiêu

Luận án tập trung thực hiện hai mục tiêu chính là giải quyết hai thách thức lớn đang tồn tại trong việc triển khai các mô hình học máy tự động trong hai nhóm dữ liệu y sinh, cụ thể các công việc như sau:

- ❖ Đối với nhóm dữ liệu sinh học phân tử
 - Thiết kế mô hình học máy hiệu quả cho dữ liệu sinh học phân tử trong các nhiệm vụ xác định gene cho biểu hiện, được ứng dụng trong lĩnh vực phát triển thuốc
 - Xây dựng mô hình học máy hiệu quả cho nhiệm vụ định danh loài sinh vật.

- ❖ Nhóm dữ liệu lâm sàng, cận lâm sàng.

Xây dựng cơ chế tự động tiền xử lý dữ liệu chuẩn đoán lâm sàng trong y sinh, cụ thể: xử lý dữ liệu lỗi, mã hóa dữ liệu cũng như rút gọn tập thuộc tính. Tập dữ liệu đã được chuẩn hóa này được dùng làm đầu vào cho quá trình huấn luyện của các thuật toán học máy trong các ứng dụng y sinh lâm sàng về chuẩn đoán bệnh.

1.2.2 Phạm vi nghiên cứu tập trung vào:

- i) Thiết kế phương pháp biểu diễn, mã hóa, chọn lọc đặc trưng cho dữ liệu gene cho bài toán tìm gene mục tiêu, lựa chọn môi trường tế bào vật chủ phù hợp với gene mục tiêu cho mục đích thiết kế thuốc bằng kỹ thuật tái tổ hợp;
- ii) Đề xuất thuật toán định danh loài dựa trên kỹ thuật Boosting;
- iii) Thiết kế phương pháp tự động mã hóa, rút gọn đặc trưng cho dữ liệu lâm sàng về chẩn đoán bệnh CoViD-19.

1.3 Nhiệm vụ của luận án

Nhiệm vụ chính của mô hình học máy cho các ứng dụng y sinh như khả năng giải thích mô hình, tăng độ chính xác của mô hình dự đoán càng chính xác càng tốt. Đối với đặc thù của loại dữ liệu này, luận án tiếp cận theo ba hướng sau:

1.3.1 Thiết kế mô hình học máy hiệu quả cho dữ liệu sinh học phân tử trong các nhiệm vụ ứng dụng trong phát triển thuốc bằng kỹ thuật tái tổ hợp

Dữ liệu dạng trình tự gene có số chiều rất lớn (hàng ngàn chiều), cơ chế sinh học phức tạp, và dữ liệu không cân bằng đều là các vấn đề lớn trong loại dữ liệu này, đây cũng là thách thức lớn của ứng dụng học máy cho bài toán y sinh trong lĩnh vực sản xuất thuốc. Chẳng hạn như trong quá trình sản xuất thuốc bằng công nghệ tái tổ hợp, việc tìm được tập gene cho biểu hiện protein cao, hay việc chọn lựa môi trường vật chủ phù hợp với gene gene mục tiêu đều giúp cho chất lượng sản phẩm protein tái tổ hợp tốt hơn. Cụ thể việc tìm được môi trường vật chủ thích hợp cho gene mục tiêu đồng nghĩa với việc quyết định mức đáp ứng codon của môi trường vật chủ với sản phẩm protein tái tổ hợp cần sản xuất thuốc.

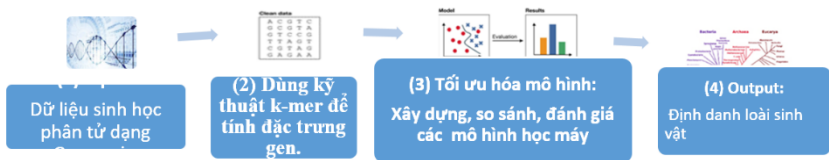
Thách thức của nhiệm vụ này là làm sao có thể *tìm được tập gene có khả năng biểu hiện protein tốt nhất trong một hệ gene*, số lượng gene này chỉ chiếm 5% tổng số trình tự của gene trong toàn hệ gene chứa hàng ngàn gene; và làm sao để có thể *tìm được môi trường vật chủ phù hợp với gene mục tiêu*. Cụ thể luận án đã đề xuất hai giải pháp hiệu quả trên tập dữ liệu gene này là: i) Giải pháp thứ nhất xây dựng mô hình "Dự đoán gene biểu hiện protein cao cho thiết kế gene dùng trong tái tổ hợp"; ii) Giải pháp thứ hai là xây dựng "Mô hình dự đoán gene tương quan với hệ thống vật chủ dùng trong tái tổ hợp".



Hình 1.1 Quy trình tổng quát của bài toán nhiệm vụ thứ 1

1.3.2 Mô hình học máy hiệu quả cho dữ liệu sinh học phân tử trong các nhiệm vụ định danh loài sinh vật.

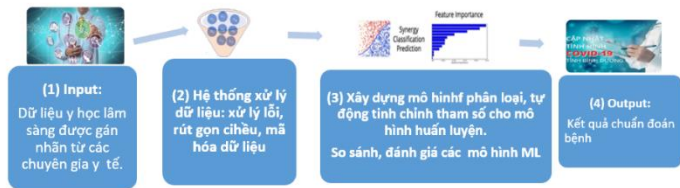
Trong các ứng dụng phát triển thuốc có sử dụng dữ liệu trình tự gene (genomic) thường có các nhiệm vụ như sau: định danh loài sinh vật, phân tích cơ chế bệnh, phát hiện bất thường trong trình tự gene. Xác định nguồn gốc đặc điểm di truyền loài để hiểu biết rõ về nguồn gốc xuất xứ của loài sinh vật đang xem xét, từ đó có bước nhận định về cơ chế sinh tồn và phát triển của sinh vật. Trước đây công việc này thường dựa và ghi chép về sinh học, và đặc điểm biểu hiện bên ngoài để xác định tên loài. Hoặc các phương pháp dò tìm trình tự bằng các thuật toán so sánh trình tự như: Neighbor Joining (NJ), phương pháp khoảng cách, phương pháp phân cụm [10, 11]. Các thuật toán này hiệu quả trên so sánh từng trình tự, khi tiến hành đối sánh trên nhiều hệ gene thì không hiệu quả về hiệu năng. Đặc biệt các hệ gene chứa các gene có chiều dài rất lớn và khác biệt. Sử dụng kỹ thuật học máy có thể xác định cùng lúc hàng ngàn gene thuộc loài nào cũng như cung cấp thông tin di truyền về các loại gene này hiệu quả hơn về thời gian và giải thích mức cấu trúc các gene. Phương pháp tổng quát này được mô tả bằng hình 1.2.



Hình 1.2 Quy trình tổng quát của bài toán định danh loài.

1.3.3 Mô hình học máy hiệu quả trong các ứng dụng y sinh về chuẩn đoán bệnh dựa trên dữ liệu lâm sàng.

Dữ liệu y sinh bao gồm dữ liệu cận lâm sàng và lâm sàng đây là dữ liệu y sinh được thu thập từ kết quả xét nghiệm sàng lọc khi khám bệnh của các cơ sở y tế. Dữ liệu này có đặc điểm chiều cao, dữ liệu thường chứa lỗi, dữ liệu bị thiếu, mất cân bằng nghiêm trọng đối với lớp bệnh hiếm. Để giải quyết các thách thức của nhiệm vụ này, luận án đề xuất phương pháp tổng quát như hình 1.2.



Hình 1.3 Quy trình tổng quát của bài toán chuẩn đoán bệnh từ dữ liệu lâm sàng.

1.4 Các đóng góp của luận án

Để thực hiện các mục tiêu của các bài toán đã nêu, các nghiên cứu về mặt lý thuyết, đồng thời các thực nghiệm thực tế đã được thực hiện trong luận án để đề xuất cách giải quyết các nhiệm vụ nghiên cứu. Các đóng góp chính của luận án bao gồm:

Thứ nhất, đã giải quyết vấn đề giảm chiều dữ liệu cho mô hình học máy trong ứng dụng sinh học phân tử cho các bài toán về chọn lựa gene thích hợp trong ứng dụng sản xuất protein tái tổ hợp. Dựa vào phương pháp tính toán các chỉ số HSCU của gene để biểu diễn thông tin đặc trưng từng gene, làm tin gọn tập dữ liệu đầu vào cho hai nhiệm vụ trong bài toán học máy: *chọn tập gene biểu hiện cao* trong hệ thống biểu hiện, và *lựa chọn được tế bào vật chủ phù hợp với gene* mục tiêu trong ứng dụng sinh học phân tử trong sản xuất protein tái tổ hợp [CT1, CT2, CT3]

Thứ hai, đã giải quyết vấn đề biểu diễn thông tin chiều cao cho mô hình học máy trong ứng dụng định danh loài sinh học. Dựa vào phương pháp k-mer để biểu diễn thông tin đặc trưng từng gene, học máy đã dùng các thuật

toán ensemble và kỹ thuật phân cụm phân cấp để tạo mô hình định danh tên loài sinh vật. Luận án đã xây dựng mô hình dự đoán tên của 17 loài nấm mốc. Mô hình có độ chính xác 91% AUCROC đạt 99% [CT4, CT7]

Thứ ba, Luận án đã sử dụng phương pháp kết hợp các thuật toán Boosting để xây dựng mô hình tự động rút gọn tính năng cho tập dữ liệu lâm sàng trong nhiệm vụ chuẩn đoán bệnh. Cụ thể luận án xây dựng mô hình dự đoán bệnh nhân nhiễm bệnh CoViD-19 dựa trên bộ dữ liệu lâm sàng này. Luận án đã sử dụng kết hợp hai kỹ thuật LightGBM và XGBoost để tự động rút gọn tính năng từ tập dữ liệu lâm sàng về bệnh CoViD-19 và Bệnh Cúm. Đồng thời xây dựng mô hình dự đoán bệnh với độ chính xác đến 99,85% [CT5, CT6]

1.5 Bố cục của luận án

Chương 1: Trình bày tính cần thiết của luận án, xác định mục tiêu, đối tượng và phạm vi nghiên cứu, luận án xác định nhiệm vụ và hướng tiếp cận của luận án. Luận án trình bày các đóng góp của luận án và tóm tắt cấu trúc của luận án.

Chương 2: Trình bày các vấn đề cơ bản liên quan đến dữ liệu y sinh. Luận án trình bày nghiên cứu tổng quan của từng yêu cầu của ba bài toán đã đề xuất, cách tiếp cận và cách giải quyết cho từng bài toán.

Chương 3: Xây dựng Chọn tập gene biểu hiện cao trong hệ thống biểu hiện bằng các thuật toán phân cụm PAM, CLARA. Chọn được tế bào vật chủ phù hợp với gene mục tiêu bằng các thuật toán Random Forest, và XGBoost.

Chương 4: Mô hình học máy định danh loài, đề xuất mô hình định danh loài sinh vật, các vấn đề của định danh loài cũng như cách quy trình xử lý dữ liệu và mô hình dự đoán tên loài sinh vật.

Chương 5: Mô hình dự đoán bệnh từ dữ liệu lâm sàng. Cách tiếp cận phương pháp chẩn đoán bệnh từ dữ liệu bệnh CoViD-19, sử dụng học máy để phân định bệnh CoViD-19 và bệnh cúm mùa.

Chương 6: Trình bày kết luận tóm tắt các đóng góp của luận án cũng như đề xuất hướng nghiên cứu tiếp theo

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VÀ CÁC CÔNG TRÌNH LIÊN QUAN

2.1 Các khái niệm trong y sinh

Nêu các khái niệm cơ bản của y sinh học có sử dụng trong các bài toán luận án như: genomics, dna tái tổ hợp, codon đồng nghĩa, hệ thống biểu hiện, định danh loài sinh vật.

2.2 Các nghiên cứu liên quan có sử dụng thuật toán học máy dùng trong các bài toán đề xuất

2.2.1 Các vấn đề về rút gọn chiều:

Đặc tính dữ liệu của các bài toán sinh học là dữ liệu chiều cao, nên việc cần giảm kích thước của tập dữ liệu ở bước trước tiên xử lý trước khi thực hiện các phân tích tiếp theo. Có hai cách tiếp cận chính để đạt được điều này, đó là kỹ thuật giảm kích thước chiều và chọn lọc tính năng. Trong các đề xuất của luận án tập trung vào việc lựa chọn tính năng: (i) kỹ thuật sử dụng trong luận án cho dữ liệu trình tự là mã hóa gene bằng kỹ thuật tính tỷ lệ codon đồng nghĩa của từng gene, và kỹ thuật k-mer để chọn lựa đặc trưng; (ii) Đối với dữ liệu lâm sàng dùng phối hợp nhiều phương pháp để tự động loại bỏ thuộc tính chứa dữ liệu trống, dữ liệu tương quan, dữ liệu độ lợi thông tin thấp.

2.2.2 Phương pháp học tập không giám sát

Phân cụm là kỹ thuật đặc trưng của phương pháp này tương tự như chính là dựa vào kết quả đo của độ đo tương tự và phương pháp định nghĩa độ đo này. Kỹ thuật phân cụm rất hiệu quả trong rút trích thông tin cho bộ dữ liệu. Luận án áp dụng kỹ thuật phân cụm để dự đoán tập gene cho biểu hiện cao và áp dụng phân cụm phân cấp để phân loài sinh vật.

2.2.3 Phương pháp học tập giám sát

Phân lớp là đại diện được gọi phương pháp học có giám sát. Học có giám sát được định nghĩa là vấn đề ước tính mối quan hệ chức năng giữa *tập thuộc tính* x_i và biến kết quả y_i , với $y_i \approx f(x_i)$. Tùy vào đặc tính của biến

y_i mà xác định là phân lớp hay hồi quy. Với y_i là nhóm giá trị cố định thì mô hình bài toán được gọi là phân lớp. Trong đề xuất một phần nhiệm vụ thứ nhất, nhiệm vụ thứ 2, nhiệm vụ thứ 3 đều sử dụng kỹ thuật này.

2.2.4 Phương pháp học máy học kết hợp

Học kết hợp (Ensemble learning) là một phương pháp học máy mà huấn luyện nhiều mô hình cùng lúc để giải quyết cùng một vấn đề, và sau đó kết hợp đầu ra của chúng để cải thiện độ chính xác. Các phương pháp học kết hợp, đặc biệt là kỹ thuật Boosting, đã được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm cả sinh học. Kỹ thuật Boosting có thể được sử dụng để phân loại các mẫu dữ liệu về bệnh nhân hoặc để dự đoán phản ứng của một loại thuốc trên một bệnh nhân cụ thể.

2.2.5 Các thuật toán dùng trong luận án để giải quyết các vấn đề đã đặt ra trong dữ liệu y sinh

Bốn phương pháp học máy như đã nêu trên đều được sử dụng trong luận án cụ thể là các thuật toán: PAM, CLARA, Random Forest, Xgboost, Catboost.

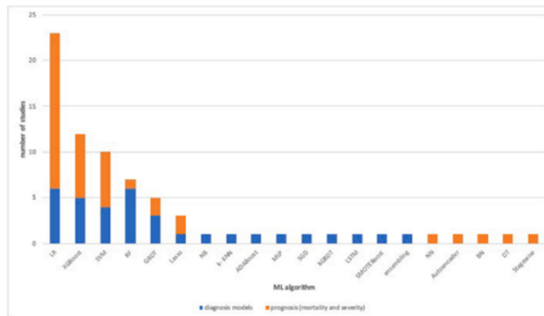
2.3 Các nghiên cứu liên quan

Đã có nhiều thành công trong việc khám phá thuốc được hỗ trợ bởi phương pháp học máy, chẳng hạn như Deep genomics đã sử dụng nền tảng bản làm việc AI để tạo ra mục tiêu di truyền mới và ứng cử viên thuốc oligonucleotide tương ứng DG12P1 để quản lý một dạng bệnh Wilsons di truyền bất thường [25] [26].

Học máy còn được sử dụng trong định danh tên chủng loại sinh vật và kiểm soát đột biến của chuỗi gene cũng hỗ trợ đắc lực cho việc phát triển vắc xin và thuốc, chẳng hạn như xác định tên loài vi rút CoViD-19 và các biến thể của nó [28].

Ngoài ra Học máy còn hỗ trợ để có được các tác nhân phòng ngừa và chữa bệnh tích cực hỗ trợ kiểm soát các dịch bệnh có thể xảy ra trong

tương lai. Trong nghiên cứu Norah Alballa đã khảo sát trên 93 luận án đã công bố từ tháng 1/2020 đến tháng 1/2021 trên các tạp chí uy tín như PubMed, Scopus, IEEE Xplore, and Google Scholar về áp dụng ML để chẩn đoán bệnh CoViD-19, và nguy cơ nhập viện. Tác giả đã thống kê các nhóm thuật toán được áp dụng như mô tả hình 2.6.



Hình 2.6: Các xu hướng sử dụng thuật toán ML trong chẩn đoán bệnh CoViD-19 dựa trên dữ liệu cận lâm sàng [29].

2.4 Dữ liệu y sinh được sử dụng trong các nhiệm vụ của luận án

Đây cũng là nhiệm vụ thứ nhất của luận án đối với dữ liệu nhóm dữ liệu genomic với hơn 10000 trình tự thuộc ba nhóm sinh vật tế bào vật chủ. Trong nhiệm vụ thứ hai, luận án vẫn dùng dữ liệu trình tự cho nhiệm vụ định danh tên loài sinh vật với trên 2000 trình tự ITS thuộc về 29 loài nấm mới. Nhiệm vụ cuối cùng là dữ liệu cận lâm sàng được thu thập với số mẫu tổng lên hơn 7000 ca của bệnh nhân CoViD-19 và Cúm H1N1.

2.5 Đánh giá mô hình mô hình học máy

Để đánh giá mô hình học máy luận án so sánh kết quả phân lớp của mô hình đề xuất với các mô hình khác. Cụ thể đã sử dụng luận án sử dụng phương pháp k-fold cross-validation (CV), Accuracy, Sensitivity (độ nhạy) - còn được gọi là Recall (PR), Specificity, FPR(False Positive Rate/Fall-out), F1 score, ROC (Receiver Operating Characteristics).

CHƯƠNG 3. MÔ HÌNH HỌC MÁY TÌM GENE CHO HỆ THỐNG BIỂU HIỆN TRONG KỸ THUẬT DNA TÁI TỔ HỢP

3.1 Bài toán tìm gene biểu hiện cao (HEG- Highly Expressed Gene)

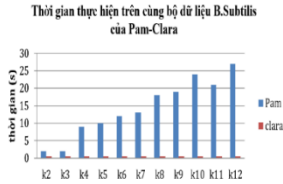
Gene biểu hiện cao là những gene có khả năng tạo ra một lượng lớn protein có tính chất đặc biệt trong một tế bào hoặc hệ thống tế bào cụ thể [5]. Những gene này được ưa chuộng trong sản xuất protein tái tổ hợp vì khả năng tăng năng suất sản xuất protein và giảm chi phí sản xuất. Tìm được HEG trong một hệ gene là việc làm khó thường có mức độ biểu hiện cao hơn so với các gene khác trong cùng một điều kiện môi trường và điều kiện thực nghiệm. Luận án đã sử dụng phương pháp phân hoạch được sử dụng để phân cụm gene trong một hệ gene nhằm lựa chọn các nhóm gene có xu hướng sử dụng codon tương tự nhau trong cùng một nhóm. Việc này để có thể chọn được các HEG có khuynh hướng gần nhất đến trung tâm của mỗi nhóm. Luận án đề xuất phương pháp phân tích cụm PAM [36], CLARA [38] cho việc phân cụm gene, có thể khái quát phương pháp này như sau:

- (1). Mã hóa gene dùng kỹ thuật RSCU cho toàn bộ gene trong hệ gene
- (2). Áp dụng các thuật toán phân cụm PAM, CLARA dữ liệu tìm ra ở bước (1), hình thành các cụm và tìm tâm cụm mới.
- (3). Đánh giá một nhóm được phân cụm. Chọn nhóm gene gần tâm cụm nhất này có khả năng cao là HEG.

3.1.1 Kết quả thực nghiệm

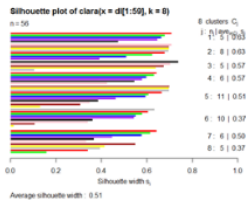
1. Thời gian thực thi PAM-CLARA trên bộ dữ liệu Bsubtilis

Tiêu chí thời gian thực nghiệm rất được quan tâm khi chọn lựa các thuật toán áp dụng tính toán với các bộ dữ liệu lớn. Thời gian thực thi của PAM và CLARA trên cùng tập gen B.Subtilis.



Hình 3.1 Thông kê thời gian thực thi PAM, CLARA

2. **Chỉ số Silhouette:** được dùng để đánh giá chất lượng phân cụm bằng chất lượng phân cụm như hình 3.3 và hình 3.4



Hình 3.2 ảnh phân cụm bằng CLARA

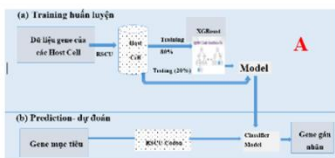
```

objective function: 83.84559
clustering vector: 1re [13543] 1 2 3 4 5 6 7 3 8 1 9 10 11 9 9 3 2 .
--
cluster sizes: 86 53 40 72 56 66 75 33 45 20 17 53 51 94 75 27 42 76
21 15 25 31 44 48 20 72 70 76 56 46 38 110 14 46 50 74 60 69 26 39 77 27 59 60
99 70 29 54 48 44 11 30 28 41 30 26 41 36 36 13 40 40 30 50 21 26 20 80 46 57 30 5 35 35
52 76 30 5
best sample:
[3] 17 45 51 79 82 84 89 138 152 193 219 242 259 1277
[15] 285 285 287 300 306 329 358 368 397 450 461 473 529 540
[28] 561 606 631 638 647 649 676 685 722 744 751 769 781 786
[43] 787 803 837 866 883 897 914 955 1004 1020 1028 1033 1065 1071
[57] 1088 1089 1139 1152 1159 1172 1174 1190 1196 1205 1206 1210 1215 1244
[73] 1246 1279 1282 1301 1306 1320 1321 1322 1328 1335 1379 1417 1451 1469
[85] 1482 1485 1490 1502 1528 1530 1561 1621 1640 1651 1669 1675 1685 1687
[99] 1709 1714 1786 1790 1802 1829 1842 1818 1841 1937 1944 1959 1975 1983
[113] 2063 2087 2098 2137 2146 2147 2148 2150 2170 2153 2260 2300 2326 2347
[127] 2352 2360 2366 2418 2438 2454 2455 2462 2510 2512 2533 2547 2562 2599
[143] 2621 2626 2645 2669 2692 2710 2717 2762 2765 2784 2785 2836 2838 2898
[155] 2906 2916 2944 2951 2996 2997 3006 3020 3025 3088 3102 3109 3161 3164
[169] 3211 3214 3234 3281 3347 3357 3376 3378 3385 3400 3425 3458 3461 3490
[183] 3492 3497 3521 3524 3528 3536 3540 3543
    
```

Hình 3.3 .5% gen HEG được chọn

3.2 Bài toán tìm hệ thống biểu hiện phù hợp với gene mục tiêu [CT3]

Để nâng cao năng suất biểu hiện protein tái tổ hợp trong sản xuất protein tái tổ hợp thì tìm Host Cell phù hợp với gene mục tiêu là rất quan trọng. Đây cũng là bài toán phân lớp được luận án đề xuất phương pháp học máy để nâng cao chất lượng dự đoán. Có thể mô tả vị trí của bài toán trong quá trình sản xuất protein tái tổ hợp. Quy trình tổng quát cho mô hình dự đoán gene tương quan với tế bào vật chủ **hình 3.4.A** Bộ dữ liệu dùng trong thực nghiệm là: E. coli, vi khuẩn chùng Bacillus subtilis và Latococcus Lactis. Hiệu năng mô hình đề xuất trên từng host cell của hai thuật toán tương ứng với các mô dự báo trong **hình3.4.B**.



Host cell	B	
	Random Forest	Xgboost
Host Ecoli (Class 1)	0,97	0,99
Host 2 B.Subtilis (Class 2)	0,94	0,99
Host 3 Latococcus (Class 3)	0,88	0,97

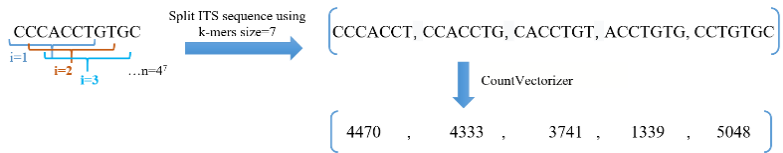
4.

A. Mô hình dự đoán gene tương quan với tế bào vật chủ và **B.** kết quả dự đoán phân loại của các thuật toán.

DNA thường được dùng để phân loại được gọi là DNA mã vạch (DNA barcode) đây là trình tự ITS (Internal transcribed spacer). Các nhóm gen này thường được sử dụng gen rRNA 18S, 5S và 16S là gene dùng để đánh giá mối quan hệ tiến hoá giữa các sinh vật.

4.2 Xây dựng tập dữ liệu cho quá trình huấn luyện

K-mer là một đoạn ngắn gồm k nucleotide liên tiếp nhau của một trình tự. Các đoạn k-mer có được từ việc dùng cửa sổ trượt có kích thước k dịch chuyển từ vị trí đầu chuỗi trình tự cho đến hết chiều dài của chuỗi trình tự [11]. Với 4 base cơ bản (A, G, T, C) có thể có $4k$ vị trí cho chuỗi trình tự



Hình 4.1 Minh họa cách tính k-mer cho chuỗi trình tự với $k=7$.

Dữ liệu về chuỗi ITS cho các loài nấm mốc trong các ngân hàng gene là không đầy đủ, tên gọi cũng không thống nhất gây khó cho việc tra cứu thông tin. Luận án đã tổng hợp dữ liệu chuỗi ITS từ hai ngân hàng gene BOLD và NCBI. Loại bỏ các loài nấm mốc có ít hơn 7 chuỗi, luận án thu được 1704 chuỗi thuộc về 17 loài nấm mốc. Áp dụng phương pháp mã hóa CountVectorizer để xây dựng bộ dữ liệu cho mô hình phân loại nấm mốc.

4.3 Thuật toán đề xuất xây dựng mô hình định danh loài dựa trên học kết hợp (Ensemble learning)

Mô hình được đề xuất sử dụng kỹ thuật Tối ưu Bayesian [63] để điều chỉnh siêu tham số trong quá trình xây dựng bộ phân loại XGBoost: 'max_depth', 'gamma', 'n_estimators', và 'learning_rate'. Điều này làm cho mô hình được xây dựng có hiệu năng tốt hơn chi tiết về thuật toán 4.1 được trình bày như sau:

Thuật toán 4.1: xây dựng một mô hình phân loại tối ưu

Input: $D_{\text{ter}} = \{(x_i, y_i) \in R^m \times \{0, 1\}, \forall i = 1, n\}$; hyperparameter is $\Theta = \{\text{'max_depth': int(max_depth), 'gamma': gamma, 'n_estimators': int(n_estimators), 'learning_rate': learning_rate}\}$

Output: Best_Model

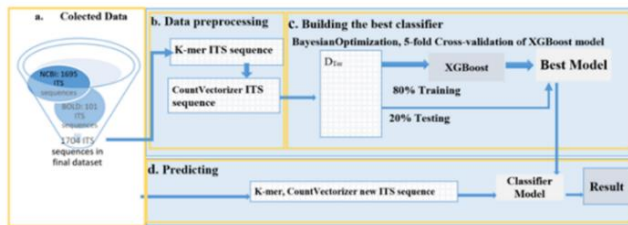
Begin

- 1: **Initialize:** FeatureImportances={}
- 2: Model $\leftarrow$ **XGBoostClassifier** (Θ)
- 3: KFold $\leftarrow$ StratifiedKFold (n_splits=5, shuffle=True, random_state=2020)
- 4: **For** i=1 **to** KFold
 - Divide the D_{ter} dataset into D_{train} and D_{test}
 - Train the model based on D_{train} and early-ending hyperparameters
- 5: **Calculate** the roc_auc_score, accuracy_score, precision_score, recall_score, and f1_score over iterations
- 6: **Select** the best model based on step 4
- 7: **Visualize** the mean value from step 4
- 8: **Return** the Best_Model from step 4

End

4.4 Mô hình Học máy định danh loài nấm mốc

Quy trình xây dựng mô hình dự đoán tên loài nấm mốc được tiến hành theo có 17 lớp được thiết lập trong mô hình tương đương với 17 loài nấm mốc. Quá trình được tiến hành như hình 4.2



Hình 4.2 Mô hình ứng dụng học máy cho dự đoán tên loài.

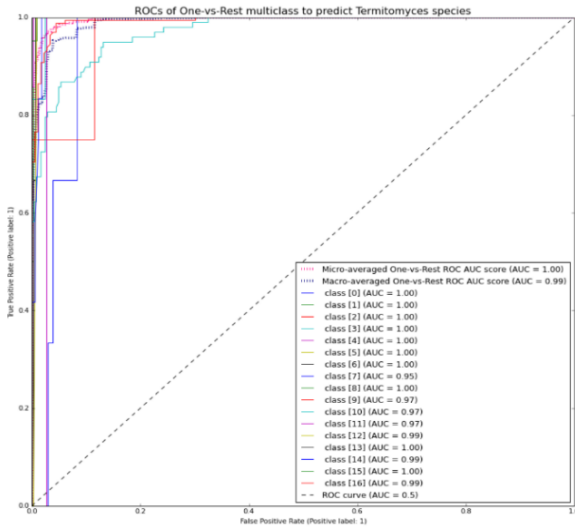
4.5 Kết quả thực nghiệm

Luận án đã tiến hành thực nghiệm trên cùng bộ dữ liệu với bốn thuật toán khác nhau để so sánh hiệu năng về các độ đo đặc trưng của mô hình học máy Kết quả chi tiết được trình bày trong Bảng 4.1

Bảng 4.1. So sánh hiệu năng các thuật toán học máy trên cùng bộ dữ liệu.

Method	AUROC	Accuracy	Precision	Recall	F1 score
Naive Bayes	0.93	0.60	0.84	0.60	0.62
RandomForest	0.98	0.88	0.88	0.88	0.88
XGBoost	0.99	0.91	0.90	0.91	0.90
Catboost	0.99	0.87	0.87	0.87	0.87

Ngoài ra AUCROC của từng lớp được tính theo phương pháp ROC OvR macro-average cho mô hình đa lớp được sử dụng [64]. Luận án sử dụng lớp cuối (là lớp 16) làm lớp tích cực các lớp còn lại là tiêu cực. Và kết quả trực quan của mỗi lớp được trình bày trong hình 8.



Hình 4.3 The ROC curve using the OvR macro-average for each class in the XGBoost method by size k-mer=7

❖ So sánh hiệu năng với các bộ phân lớp về nấm:

Kết quả đề xuất của luận án có hiệu năng cao hơn các mô hình phân lớp khác khi áp dụng k-mer size với kích cỡ là 7 với thuật toán phân loại được chọn là XGBoost. Kết quả so sánh được trình bày trong bảng 4.2

Bảng 4.2: So sánh hiệu năng với các bộ phân lớp khác nhau về năm cũng sử dụng trình tự ITS.

Nghiên cứu	Phương pháp	Độ chính xác
(Schloss PD 2009)	K-mer (k=5), kNN, PGMA	0.86
(Delgado 2016),	K-mer (k=5), Naïve Bayes	0.87
(Deshpande 2016)	K-mer (k=8) Bayesin	0.87
(Prabina Kumar 2019)	K-mer (k=4), Random Forest.	0.89
Đề xuất của luận án	K-mer (k=7), XGBoost	0.91

❖ So sánh kết quả dự đoán của mô hình đề xuất với kết quả dự đoán của BLAST

Các trình tự ITS của nấm mốc được thu thập tại tỉnh Bình Dương, Việt Nam đã được công bố trên NCBI. Các trình tự này được dự đoán bằng mô hình phân loại được đề xuất của luận án cũng cho kết quả giống như của kết quả định danh loài được lưu trên NCBI. Bên cạnh đó các trình tự thuộc chi nấm mốc MF163445-BD3, MF163446-BD6, MF163149-BD4 nhưng chưa rõ thuộc loài nào. Bộ phân loại của luận án đã định danh được hai trình tự MF163445-BD3, MF163446-BD6 thuộc loài *Termitomyces striatus* giống kiểu hình của mẫu nấm được thu thập khi giải trình tự.

4.6 Kết luận:

Đây là phương pháp học máy kết hợp mã hóa thông tin di truyền theo phương pháp ngôn ngữ tự nhiên kết hợp với kỹ thuật tối ưu hóa tham số mô hình học đã xây dựng mô hình học máy. Với cách kết hợp như đã mang đến kết quả tốt hơn so với các đề xuất cho bài toán tương tự.

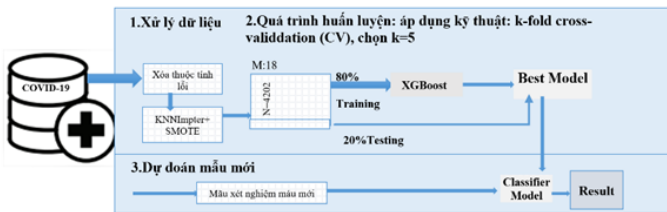
CHƯƠNG 5. MÔ HÌNH HỌC MÁY CHO CHẨN ĐOÁN BỆNH DỰA TRÊN DỮ LIỆU LÂM SÀNG.

5.1 Mô hình dự đoán bệnh dựa trên dữ liệu lâm sàng

5.1.1 Giới thiệu bài toán dự đoán bệnh và mô hình đề xuất

Dữ liệu từ kết quả xét nghiệm máu, là loại dữ liệu cận lâm sàng thường được dùng nhiều trong chẩn đoán bệnh ban đầu. Đây cũng là loại dữ liệu được lưu trữ nhiều trong hồ sơ y tế. Chẩn đoán bệnh với sự hỗ trợ tự động bằng học máy dựa vào kết quả xét nghiệm máu có thể phát hiện các về bệnh về máu, bệnh tim mạch, bệnh về gan. Luận án sử dụng kết quả xét nghiệm máu của bệnh nhân mắc bệnh CoViD-19 được cung cấp bởi bệnh viện Israelita Albert Einstein ở Brazil làm dữ liệu huấn luyện [32] .

Hình 5.2.Mô hình tổng quan đề xuất chẩn đoán bệnh CoViD-19



Để có được dữ liệu tốt nhất cho mô hình phân loại, luận án đã áp dụng phương pháp loại bỏ thuộc tính có tỉ lệ lỗi (null) hơn 99,9% và loại bỏ các thuộc tính có mức độ liên quan cao trong tập dữ liệu. Để bỏ tất cả các thuộc tính của dữ liệu đầu vào có tỷ lệ lỗi lớn hơn 99,9 %. Luận án đã đề xuất loại bỏ thuộc tính có tỷ lệ lỗi cao bằng Thuật toán 5.1

Thuật toán 5.1: Thuật toán loại bỏ thuộc tính có độ lỗi cao

Đầu vào: set $D_{ogr} = X_{ogr} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}, x \in R^m, y \in \{0, +1\}$ là không gian m -chiều, $Y \in [0, 1], thresh$: ngưỡng dùng để loại bỏ dữ liệu

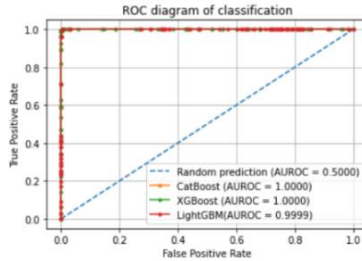
Đầu ra: D_{lean}

Begin

- 1: missing_total $\leftarrow$ Tính giá trị null trong các cột
- 2: missing_percent $\leftarrow$ missing_total / len(D_{ogr})
- 3: data_missing $\leftarrow D_{ogr}.index[missing_percent \geq thresh]$
- 4: Clean $\leftarrow D_{ogr}.drop[data_missing]$

5.1.2 Kết quả thực nghiệm

Luận án sử dụng nhiều thuật toán khác nhau để xây dựng bộ phân lớp dựa trên Datasmall như: XGBoost [22], LGBMClassifier [47], CatBoost [48]. Kết quả đánh giá AUROC của từng thuật toán được mô tả như hình 4. Nhận thấy XGBoost, cho giá trị AUROC tốt nhất nên dùng chọn dùng để xây dựng bộ phân lớp.



Hình 5.2. So sánh hiệu năng ROC của các thuật toán tăng cường độ độc

Các kết quả của mô hình đề xuất còn được so sánh với kết quả dự đoán cùng chức năng như nghiên cứu của tác giả Maryam AlJame [25].

Bảng 5.1 Bảng so sánh hiệu năng của mô hình đề xuất

Bộ phân lớp	Thuật toán	AUC	Precision	Recall	F1 score
Mô hình đề xuất	XGBoost	0.998	1.0	1,0	1,0
Maryam AlJame [25]	XGBoost	0.994	0.98	0.99	-

Nhận thấy mô hình đề xuất với các cách xử lý dữ liệu đầu vào như loại bỏ thuộc tính có độ lỗi hơn 99%, xóa dữ liệu trống bằng kỹ thuật KNNimputer, giảm mất cân bằng giữa hai lớp bằng kỹ thuật SMOTE. Việc kết hợp này giúp cho mô hình phân loại đề xuất có kết quả dự đoán tốt hơn nghiên cứu cùng mục tiêu và bộ dữ liệu.

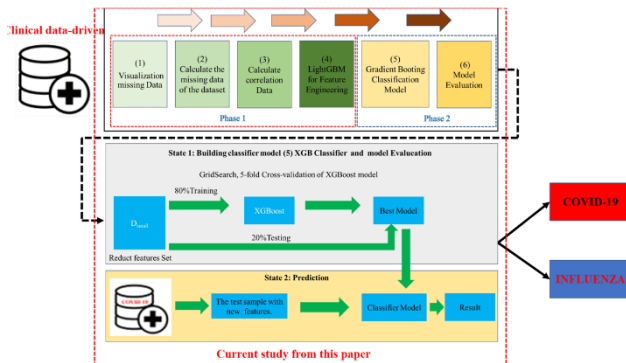
5.2 Mô hình phân loại bệnh CoViD-19 và bệnh cúm mùa

5.2.1 Giới thiệu bài toán và mô hình giải quyết

Dữ liệu lâm sàng và cận lâm sàng khi được thu thập tại các bệnh viện hay hồ sơ bệnh án luôn có rất nhiều thuộc tính bị thiếu, lỗi. Nếu sử dụng các phương pháp xử lý bằng cách xóa bỏ lỗi thì có thể ảnh hưởng đến kết quả

dự đoán bằng cách bỏ qua dữ liệu chứa các thuộc tính quý được dùng trong chẩn đoán và điều trị. Luận án đề xuất một mô hình xử lý dữ liệu tự động dữ liệu lỗi và phân biệt hai loại bệnh có triệu chứng gần giống nhau là bệnh CoViD-19 và cúm H1N1.

Mô hình đề xuất có thể học trực tiếp từ dữ liệu thô mà không cần sự can thiệp để xóa dữ liệu trống. Phương pháp đề xuất bao gồm hai giai đoạn: Ở giai đoạn đầu tiên, nó đánh giá và xử lý dữ liệu để giảm các thuộc tính không có giá trị dự đoán trong tập dữ liệu bằng thuật toán Light Gradient Boosting (LightGBM); ở giai đoạn thứ hai, nó xây dựng một mô hình phân loại cho mỗi nhóm bệnh dựa trên phương pháp XGBoost. Mô hình đề xuất trong hình 5.3 mô tả đề xuất này:



Hình 5.3.Mô hình tổng quan phân biệt Covid -19 và Cúm

Luận án đã sử dụng LightGBM để lựa chọn các thuộc tính quan trọng trong tập dữ liệu lâm sàng từ nền tảng CoViD-19 và cúm. Sau đó, mô hình cây sử dụng phương pháp tính toán lợi ích thông tin (IG) để tính toán giá trị cho các thuộc tính quan trọng. Luận án đã sử dụng thuật toán 5.2 để trích xuất các thuộc tính quan trọng và để xây dựng bộ phân loại bệnh thì thuật toán 5.3 chi tiết về hai thuật toán được mô tả như sau:

Thuật toán 5.2. Trích chọn các thuộc tính

Đầu vào: D_{learn} ; iter = number of iterations, hyperparameters = {objective = 'binary', boosting_type = 'goss', n_estimators = 1000, class_weight = 'imbalanced'}

Đầu ra: D_{small} : The dataset

Begin

1: Initialize: FeatureImportances={}

2: Model $\leftarrow$ LGBM classifier (hyperparameters)

3: **Repeat** iter times

- Divide D_{learn} dataset into D_{Train} and D_{Test}
- Train model based on D_{Train} and early-ending hyperparameters
- FeatureImportances = model. FeatureImportances

End repeat

4: Tính trung bình FeatureImportances

5: Chọn ZeroFeatureImportances //Chọn các thuộc tính có giá trị là không

6: Return $D_{small} \leftarrow D_{learn} - \text{ZeroFeatureImportances}$

End

Thuật toán 5.3: Xây dựng bộ phân loại bệnh dùng kỹ thuật XGboost

Đầu vào : $D_{small} = \{(x_i, y_i) \in R^m \times \{0,1\}, \forall i = \overline{1, n}\}$; hyperparameter is $\Theta = \{\text{'n_estimators'}$, $\text{'colsample_bytree'}$, 'learning_rate' , 'eval_metric' : 'auc'}\}

Đầu ra: Best_Model

Begin

1: Initialize: FeatureImportances={}

2: Model $\leftarrow$ **XGBoostClassifier** (Θ)

3: KFold $\leftarrow$ StratifiedKFold (n_splits=5, shuffle=True, random_state=2020)

4: For i=1 to KFold

- Divide the D_{small} dataset into D_{Train} and D_{Test}
- Train the model based on D_{Train} and early-ending hyperparameters

5: Tính các điểm roc_auc_score, accuracy_score, precision_score, recall_score, and f1_score

6: Chọn the best_model ở bước 4

7: Visualize the mean value ở bước 4

8: Return the Best_Model ở bước 4

End

5.2.2 Hiệu năng của mô hình đề xuất

❖ So sánh mức độ chính xác về dữ liệu dự đoán

Từ tập dữ liệu đầu vào gồm 50 thuộc tính, luận án trích xuất được 24 thuộc tính quan trọng để dùng cho chuẩn đoán bệnh. Trong đó thuộc tính *GroundGlassOpacity* này có vai trò đặc biệt quan trọng cho chẩn đoán CoViD-19, vì thuộc tính này luôn xuất hiện trong hầu hết các trường hợp nhập viện điều trị với tỷ lệ 89%. Việc mô hình đã chọn thêm tính năng

này trong mô hình đề xuất để dự đoán CoViD-19 và cúm rất có ý nghĩa về việc phát hiện ý nghĩa tiềm ẩn của dữ liệu.

❖ Hiệu năng của mô hình đề xuất

Để kiểm tra tính chính xác của mô hình đề xuất của luận án, luận án đã so sánh hiệu suất của nó với mô hình của Li kết quả cụ thể được mô tả như bảng 5.6.

Bảng 5.6: . So sánh hiệu suất của mô hình đề xuất với mô hình của Li

Bộ phân loại		Accuracy (%)	AUC score (%)	Recall score (%)	Precision score (%)	F1 score (%)
Mô hình đề xuất của luận án	XGBoost	99.9	99	99	100	99
	Gradient-Boosting	99.6	99	98	100	99
	LGBM	99.7	99	99	100	99
	Random Forest	99.8	99	99	100	99
Li [77]	XGBoost	99	97.7	92	92	92
	RIDGE regression	-	96.6	-	-	-
	Random Forest	-	95.3	-	-	-
	LASSO regression	-	96.3	-	-	-

5.3 Kết luận

Dữ liệu lâm sàng và cận lâm sàng đều là dữ liệu quan trọng của nghiên cứu về y sinh. Dựa trên các bộ dữ liệu này, luận án đã đề xuất hai phương pháp xây hai mô hình học máy hỗ trợ chuẩn đoán bệnh và phân biệt bệnh. Khi so sánh kết quả của của các mô hình đề xuất với các nghiên cứu cùng mục tiêu và bộ dữ liệu, hiệu năng của mô hình đề xuất cho kết quả tốt hơn.

CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết luận

Các mô hình học máy có thể đảm nhận tốt vai trò dự đoán, phân loại, và trực quan hóa các loại dữ liệu phức tạp của sinh học, y học. Luận án đã đề xuất các giải pháp hiệu quả hỗ trợ tính toán trong các vấn đề sinh học cấp bách như sản xuất thuốc trong tái tổ hợp, định danh loài và chuẩn đoán y khoa. Tìm giải pháp mới để nâng cao hiệu quả cho mô hình học máy trong dữ liệu y sinh là nhiệm vụ chính của luận án. Trong các đề xuất, luận án đã tập trung thực hiện ba nhiệm vụ chính là giải quyết hai thách thức lớn đang tồn tại trong việc triển khai các mô hình học máy cho hai nhóm dữ liệu y sinh: dữ liệu sinh học phân tử và dữ liệu lâm sàng, cận lâm sàng. Trong từng nhiệm vụ luận án đã thực được như:

- Thiết kế mô hình học máy hiệu quả cho dữ liệu sinh học phân tử trong các nhiệm vụ xác định gene cho biểu hiện, được ứng dụng trong lĩnh vực phát triển thuốc;

- Xây dựng mô hình học máy hiệu quả cho nhiệm vụ định danh loài sinh vật.

- Xây dựng mô hình học máy hiệu quả chẩn đoán bệnh trong y sinh, Cụ thể luận án đã thực hiện các nghiên cứu như sau:

Trong nhiệm vụ thứ nhất, luận án đã đề xuất hai phương án để chuyển đổi dữ liệu từ dạng trình tự gene thô thành dạng có thể xử lý bằng học máy như dùng chỉ số RSCU để biểu diễn mức độ codon tương đồng cho trình tự gene ở bài toán 1, phục vụ cho mục tiêu tối ưu hóa việc chọn lựa mô trường tổng hợp protein tái tổ hợp cho tế bào vật chủ.

Ở nhiệm vụ thứ hai luận án đã đề xuất phương pháp trích xuất thông tin của gene cho mô hình học bằng phương pháp K-mer. Phương pháp này chuyển đổi từ định dạng trình tự của tập gene thô thành thông tin dạng số và đưa vào mô hình học máy để định danh loài sinh học. Đây là một

phương pháp định danh mới với độ chính xác cao và thuận lợi cho việc trực quan dữ liệu phân loài.

Trong nhiệm vụ thứ ba, việc xử lý dữ liệu lỗi và chọn các thông tin quan trọng từ dữ liệu lâm sàng về chẩn đoán bệnh trong y sinh là một vấn đề được luận án giải quyết bằng cách kết hợp hai mô hình học Gradient Boosting là LightGBM và XGBoost. Đề xuất này giúp việc chọn lựa thuộc tính quan trọng được thực hiện một cách tự động trước khi đưa vào mô hình học chẩn đoán bệnh với quy tắc tránh mất thông tin quan trọng mà vẫn giữ độ chính xác tốt nhất cho mô hình dự đoán.

6.2 Hướng phát triển

Ứng dụng y sinh là một lĩnh vực rộng và rất tiềm năng cho việc nghiên cứu các giải pháp tính hỗ trợ các ứng dụng sinh học trong chăm sóc sức khỏe con người. Việc tiếp tục nâng cấp độ chính xác của các mô hình dự đoán ở bài toán 1 và bài toán 2 cũng là vấn đề đặt ra. Giải pháp có thể sử dụng học sâu trích xuất thông tin tự động ở bài toán 2 thay vì phải sử dụng kỹ thuật K-mer để trích xuất tính năng vì rằng phương pháp K-mer có thể đã phù hợp với định danh loài nấm mốc nhưng không phù hợp với các loài có bộ gene với độ dài lớn hơn nên việc trích xuất tính năng tự động là cần thiết. Đối với bài toán thứ 3, cũng có thể bổ sung tính năng cho mô hình dự đoán là tự động phân tầng mức độ bệnh để việc triển khai vào thực tế cho thiết thực hơn.

CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ

[CT1]. **Dương Thị Kim Chi**, Trần Văn Lăng, Lê Mậu Long, *Xác định các tham số quan trọng cho việc thiết kế gene dùng trong tái tổ hợp*. Kỷ yếu Hội nghị Quốc gia lần IX về Nghiên cứu cơ bản và Ứng dụng Công nghệ thông tin, Cần Thơ, 04-05/8/2016, ISBN: 978-604-913-472-2, NXB. KHTN&CN, DOI: 10.15625/vap.2016.000103, tr. 846-853.

[CT2]. **Dương Thị Kim Chi**, Trần Văn Lăng, Huỳnh Xuân Hiệp, *Dự đoán gene biểu hiện cao cho thiết kế gene dùng trong tái tổ hợp*. Kỷ yếu Hội nghị Quốc gia lần IX về Nghiên cứu cơ bản và Ứng dụng Công nghệ thông tin, Cần Thơ, 04-05/8/2016, ISBN: 978-604-913-472-2, NXB. KHTN&CN, DOI: 10.15625/vap.2016.00017, tr. 134-142.

[CT3] **Dương Thị Kim Chi**, Trần Văn Lăng, *Mô hình dự đoán gene tương quan với hệ thống vật chủ dùng trong Tái tổ hợp*. Kỷ yếu Hội nghị Quốc gia lần X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin, Đà Nẵng, 17-18/8/2017, ISBN:978-604-913-614-6, Nxb. KHTN&CN, DOI:10.15625/vap.2017.00049, tr. 408 – 416.

[CT4]. **Dương Thị Kim Chi**, Nguyễn Thị Ngọc Nhi, Nguyễn Thế Bảo, Lê Mậu Long, Phạm Công Xuyên, *Ứng dụng học máy cho định danh loài nấm mốc*”, Kỷ yếu Hội nghị Quốc gia lần XI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin, Hà Nội, 09-10/8/2018, ISBN:978-604-913-749-5, Nxb. KHTN&CN, DOI:10.15625/vap.2018.00069 tr. 529 – 536.

[CT5]. **Dương Thị Kim Chi**, Trần Văn Lăng, Trần Bá Minh Sơn, *Mô hình học tăng cường độ dốc dự đoán bệnh CoViD-19 từ dữ liệu lâm sàng*, Kỷ yếu Hội thảo quốc gia STAIS-2022 “Ứng dụng Công nghệ thông minh trong công nghiệp 4.0 thành phố thông minh và phát triển bền vững”, Bình Dương, 14/7/2022, ISBN: 978-604-357-047-2, tr. 245-253

[CT6]. **Dương Thị Kim Chi**, Tran Van Lang, Thanh Q. Nguyen; *Clinical data-driven approach to identifying CoViD-19 and influenza from a*

gradient-boosting model, Cogenet Engineering; Volume 10, Issue 1, 2023, DOI: 10.1080/23311916.2023.2188683 (ESCI, Scopus, Scimago Q2)

[CT7]. **Thi Kim Chi Duong**, Van Lang Tran, The Bao Nguyen, Thi Thuy Nguyen, Ngoc Trung Kien Ho Thanh Q. Nguyen, *Ensemble learning-based approach for automatic classification of termite mushrooms*, Frontiers Genetics, Sec. Computational Genomics, 2023, DOI: 10.3389/fgene.2023.1208695 (SCI-E, Scopus, Scimago Q2)