

**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

Đồng Nai, năm 2023

**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

Mã số: 9480101

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS.TS. Trần Văn Lăng

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của bản thân. Luận án này được thực hiện dưới sự hướng dẫn của PGS.TS.Trần Văn Lãng. Các kết quả nghiên cứu trong luận án là trung thực và chưa từng được ai công bố trong các công trình nào khác. Việc tham khảo các nguồn tài liệu đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định. Các bài báo được công bố chung với nhiều tác giả đều được sự đồng ý của các đồng tác giả trước khi đưa vào luận án.

Người hướng dẫn chính

Đồng Nai, ngàytháng ...năm 2023
Nghiên cứu sinh

PGS.TS. Trần Văn Lãng

Dương Thị Kim Chi

LỜI CẢM ƠN

Để hoàn thành luận án này tôi đã nhận được sự hướng dẫn, quan tâm, giúp đỡ nhiệt tình từ Quý thầy cô, bạn bè và người thân. Tôi xin gửi lời cảm ơn chân thành đến:

Thầy đã tận tình chỉ bảo, hướng dẫn, động viên và tạo mọi điều kiện tốt nhất cho tôi trong quá trình học tập và nghiên cứu. Thầy cô và các anh, chị của Khoa Công nghệ thông tin, Phòng Sau Đại học, Ban Giám hiệu Trường Đại học Lạc Hồng đã cung cấp thêm kiến thức, tạo mọi điều kiện cho tôi và quan tâm, hỗ trợ tôi trong quá trình học tập.

Ban Giám hiệu Trường Đại học Thủ Dầu Một, Ban Chủ nhiệm Viện Kỹ thuật Công nghệ đã tạo điều kiện để tôi được tham gia học tập nâng cao trình độ chuyên môn, các bạn đồng nghiệp đã không ngừng động viên và giúp đỡ tôi trong suốt thời gian học tập.

Sau cùng tôi xin chân thành cảm ơn sâu sắc đến gia đình và người thân đã giúp đỡ, động viên tôi trong suốt quá trình học tập và tạo điều kiện tốt nhất để tôi hoàn thành luận án.

NCS. Dương Thị Kim Chi

TÓM TẮT

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. Tính toán y sinh giúp đẩy nhanh quá trình phát triển thuốc, tăng hiệu quả trong việc chẩn đoán bệnh và điều trị bệnh. Một trong công cụ hỗ trợ cho tính toán y sinh thuận lợi hiệu quả hơn là các phương pháp học máy. Các phương pháp học máy tạo ra các mô hình giúp quá trình nhận dạng, phân loại được thực hiện một cách tự động và đạt độ chính xác cao. Trong lĩnh vực tin y sinh các mô hình học máy được huấn luyện trên dữ liệu đầu vào sau đó sử dụng các thuật toán để phân loại hoặc dự đoán kết quả. Mô hình học máy cho dữ liệu y sinh có vai trò hết sức cần thiết và cấp bách nhằm phân loại các đối tượng để đưa ra các quyết định chính xác trong chẩn đoán và điều trị.

Việc xây dựng mô hình phân loại cho dữ liệu y sinh đòi hỏi kỹ năng chuyên môn, kinh nghiệm và sự hiểu biết sâu sắc về dữ liệu y sinh và các phương pháp tính toán phù hợp. Đặc biệt, việc lựa chọn các đặc trưng quan trọng, xử lý dữ liệu thiếu, cân bằng dữ liệu và đánh giá hiệu suất của mô hình là rất quan trọng để đạt được kết quả phân loại chính xác và đáng tin cậy. Cụ thể luận án đã giải quyết các vấn đề nâng cao hiệu quả các mô hình phân lớp, phân cụm trên dữ liệu y sinh với những đóng góp như sau:

Thứ nhất, 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*¹ 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

¹ Gene mục tiêu: gene của một loài sinh vật có khả năng biểu hiện sản phẩm protein tốt trong cần sản xuất thuốc

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". Đối với giải pháp 1, luận án đã sử dụng kỹ thuật codon đồng nghĩa để tính chỉ số codon đồng nghĩa RSCU (Relative Synonymous Codon Usage) qua đó biểu diễn đặc trưng cho từng gene; tiếp theo luận án đã áp dụng hai giải thuật PAM (Partitioning Around Medoids), CLARA (Clustering for Large Applications) cho việc phân cụm dự đoán gene cho biểu hiện protein cao. Đối với giải pháp 2, luận án đã xây dựng mô hình dự đoán gene tương quan phù hợp với tế bào vật chủ với thuật toán XGBoost. Mô hình dự đoán của đề xuất này đạt độ chính xác cao nhất 0,99. Những kết quả này đã được công bố trong các công trình [CT1][CT2][CT3].

Thứ hai, 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. Việc định danh loài giúp xác định tên loài, phân tích thay đổi tiến hóa, hay hình thành loài mới. Với việc phân loại loài dựa trên kiểu hình của sinh vật ẩn chứa nhiều khả năng định dạng sai loài vì vật mẫu có thể bị đột biến nên biểu hiện bên ngoài thay đổi nên rất dễ nhầm lẫn thành loài mới. Định danh loài bằng kỹ thuật sinh học phân tử giúp xác định loài tốt hơn, có thể phát hiện loài mới và xác định đột biến trong loài. Số lượng trình tự các loài sinh vật từ các ngân hàng gene quốc tế rất lớn nhưng phân phối không đồng đều giữa các loài trong cùng một chi. Bên cạnh đó độ dài trình tự của các loài cũng rất khác biệt trong cùng loại. Đây là thách thức của nhiệm vụ định danh loài bằng kỹ thuật sinh học phân tử khi triển khai bằng các kỹ thuật định danh loài truyền thống như NJ, phương pháp khoảng cách, phương pháp phân cụm. Luận án đã đề xuất giải pháp mới sử dụng học máy để định dạng tên loài: i) Tự động trích xuất đặc trưng trình tự sinh học, ii) Vector hóa từ để số hóa dữ liệu chuỗi, iii) Tối ưu hóa tham số, iv) Xây dựng bộ phân loại. Thực nghiệm trên bộ dữ liệu trình tự nấm mốc đã cho ra kết mô hình định danh loài nấm mốc với hiệu năng và độ chính xác vượt trội. Cụ thể luận án đã tiến hành thực nghiệm trích xuất thông tin trên gene đặt trung ITS

của 17 loài nắm mỗi loài bằng kỹ thuật K-mer. Sau đó tiến hành phân loại bằng các thuật toán phân loại kết hợp, và phân cụm phân cấp để xác định tên loài. Kết quả mô hình phân lớp đạt kết quả về độ chính xác: 0,91; Multi-class area under the curve: 0.99; Thời gian thực thi 1.66 s. Với đề xuất này cho kết quả chính xác cao thời gian thực thi thấp và trùng khớp kết quả dự đoán với phần mềm BLAST của ngân hàng gene quốc tế NCBI. Mô hình này đạt hiệu quả cao về độ chính xác trong thời gian ngắn nên có thể triển khai khi trong thực tiễn. Kết quả đã công bố trong các công trình [CT4][CT7].

Thứ ba, 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 hai vấn đề nghiêm trọng dữ liệu trống và mất cân bằng dữ liệu luận án đã sử dụng hai giải pháp: i) Giải pháp thứ nhất: Sử dụng phương pháp KNNImputer để bổ sung thêm dữ liệu trống, và sử dụng kỹ thuật SMOTE (Synthetic Minority Oversampling Technique) để xử lý dữ liệu trước khi thử nghiệm các thuật toán tăng cường độ dốc để xây dựng bộ phân loại. Việc thử nghiệm mô hình dự đoán này trên bộ dữ liệu lâm sàng từ xét nghiệm mẫu máu của bệnh CoViD-19 của các bệnh nhân nhập bệnh viện Israelita Albert Einstein ở Brazil để dự đoán khả năng mắc bệnh CoViD-19. Hiệu suất của mô hình đạt độ chính xác tổng thể đạt trên 0,998. ii) Giải pháp thứ hai: sử dụng kết hợp hai bộ phân loại LightGBM và XGBoost để xây dựng mô hình phân loại bệnh CoViD-19 và Bệnh Cúm mùa, mô hình đề xuất đạt độ chính xác là 0,99. Khi tiến hành so sánh phương pháp đề xuất với các công bố khác trên cùng bộ dữ liệu COVIDandFLU cho chẩn đoán bệnh CoViD-19 và Bệnh Cúm mùa, mô hình đề xuất có kết quả vượt trội hơn về độ chính xác cũng như độ nhạy Recall, độ đặc hiệu (Specificity), F1 score, ROC. Kết quả tổng thể của mô hình đều đạt ở mức là 0.99 và đã được công bố trên [CT5][CT6].

Từ khóa: Genenomic, dữ liệu lâm sàng, học kết hợp, học máy tăng cường độ dốc, phân loại, Rừng Ngẫu Nhiên.

ABSTRACT

Biomedical computing (biomedical informatics) is an interdisciplinary research field that combines medicine and computer science. It involves the combination of data analysis methods, machine learning, statistics, and information theory to address issues in the biomedical field such as disease detection and diagnosis, drug design, and molecular biology research. Biomedical computing helps accelerate the drug development process, improve efficiency in disease diagnosis and treatment. Machine learning techniques are one of the useful tools in biomedical computing. Machine learning techniques create models that facilitate automatic identification and classification with high accuracy. In the field of biomedical informatics, machine learning models are trained on input data and then use algorithms to classify or predict outcomes. Machine learning models for biomedical data play a crucial and urgent role in classifying objects to make accurate decisions in diagnosis and treatment

Building classification models for biomedical data requires specialized skills, experience, and a deep understanding of biomedical data and appropriate computational methods. Specifically, selecting important features, handling missing data, balancing data, and evaluating model performance are crucial to achieve accurate and reliable classification results. In particular, the thesis addresses the challenges of improving the effectiveness of classification and clustering models on biomedical data, with the following contributions:

Firstly, gene sequence data has a very high dimensionality (thousands of dimensions), complex biological mechanisms, and imbalanced data distribution, which are significant challenges in this type of data and a major obstacle in applying machine learning to biomedical problems in the field of the drug production. For example, in the process of producing drugs using recombinant technology, finding a set of genes for high protein expression or selecting a suitable host environment for target genes can improve the quality of recombinant protein products. Specifically, finding the appropriate host environment for the target gene is synonymous with determining the codon responsiveness of the host environment to the desired recombinant protein. The challenge of this task is how to identify a set of genes with the highest potential for protein expression within a gene system, where this set of genes only accounts for 5% of the total gene sequences in the gene system containing thousands of genes. Furthermore, finding the appropriate host environment for the target gene is another challenge. In this regard,

the thesis proposes two effective solutions for this gene dataset: **i)** The first solution is to build a model for "Predicting high protein-expressing genes for gene design in recombinant technology"; **ii)** The second solution is to build a "Model for predicting gene correlation with the host system used in recombinant technology." For the first solution, the thesis utilizes synonymous codon techniques to calculate the Relative Synonymous Codon Usage (RSCU) index, representing features for each gene. Then, the thesis applies two algorithms, PAM (Partitioning Around Medoids) and CLARA (Clustering for Large Applications), for clustering and predicting genes for high protein expression. For the second solution, the thesis develops a gene correlation prediction model with the host cell using the XGBoost algorithm. The proposed prediction model achieves the highest accuracy of 0.99. These results have been published in the following studies [CT1], [CT2], [CT3].

Secondly, in drug development applications that utilize gene sequence (genomic) data, the following tasks are commonly performed: species identification, analysis of disease mechanisms, and detection of abnormalities in gene sequences. Species identification helps determine the name of the species, analyze evolutionary changes, or identify new species. Classifying species based on morphological characteristics of hidden organisms can lead to misidentifying them as new species, as the external appearance may change due to mutations. Species identification using molecular biology techniques enables more accurate species determination and the detection of new species and mutations within species. The number of sequences of different species in international gene banks is vast, but their distribution is uneven among species within the same genus. Additionally, the sequence lengths of species within the same group can vary significantly. These are the main challenges of species identification using molecular biology techniques when implementing traditional species identification methods such as NJ (Neighbor-Joining), distance-based methods, and clustering methods. The thesis proposes a novel solution using machine learning for species name assignment, which includes: **i)** Automatic extraction of biological sequence features; **ii)** Vectorization of words for sequence data encoding; **iii)** Parameter optimization; **iv)** Construction of a classifier. Experiments on termite mushroom sequence data yielded a model for termite mushroom species identification with outstanding performance and accuracy. Specifically, the thesis conducted experiments to extract information from the ITS gene

features of 17 termite mushroom species using the K-mer technique. Subsequently, classification was performed using combined classification algorithms and hierarchical clustering to determine the species' names. The classification model achieved the following results: Accuracy: 0.91, Multi-class area under the curve: 0.99, Execution time: 1.66 s. This proposal demonstrated high accuracy, low execution time, and matching prediction results with the NCBI's BLAST software, which is an international gene bank. This model achieved high effectiveness in terms of accuracy in a short period, making it suitable for practical implementation. The results have been published in the following studies [CT4], [CT7].

Thirdly, biomedical data includes clinical and laboratory data, which are collected from diagnostic screening results during medical examinations at healthcare facilities. This data has the characteristic of high dimensionality and often contains errors, missing values, and severe class imbalance for rare diseases. To address the two significant issues of missing data and data imbalance, the thesis utilized two solutions: **i)** The first solution: Using the KNNImputer method to impute missing data and applying the SMOTE (Synthetic Minority Oversampling Technique) technique to preprocess the data before experimenting with gradient boosting algorithms to construct a classifier. The predictive model was tested on clinical data from blood sample tests for COVID-19 patients admitted to the Israelita Albert Einstein Hospital in Brazil to predict the likelihood of COVID-19 infection. The model achieved an overall accuracy rate of over 0.998; **ii)** The second solution: Using a combination of two classifiers, LightGBM and XGBoost, to build a classification model for COVID-19 and seasonal influenza. The proposed model achieved an accuracy rate of 0.99. When comparing the proposed method with other publications on the same COVIDandFLU dataset for diagnosing COVID-19 and seasonal influenza, the model also demonstrated superior results in terms of accuracy, sensitivity (Recall), specificity, F1 score, and ROC. The overall performance of the model reached a level of 0.99 and has been published in [CT5] and [CT6]

Key words: Genenomic, clinical data, ensemble learning, gradient-boosting machine learning, classification, Random Forest, Ensemble learning.

MỤC LỤC

CHƯƠNG 1. TỔNG QUAN.....	1
1.1. TÍNH CẤP THIẾT CỦA LUẬN ÁN	1
1.2. MỤC TIÊU, ĐỐI TƯỢNG, PHẠM VI VÀ PHƯƠNG PHÁP NGHIÊN CỨU	1
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.....	3
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.	5
1.3.3. Mô hình học máy hiệu quả trong các ứng dụng y sinh về chẩn đoán bệnh dựa trên dữ liệu lâm sàng.	6
1.4. CÁC ĐÓNG GÓP CỦA LUẬN ÁN	8
1.5. BỐ CỤC CỦA LUẬN ÁN	8
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VÀ CÁC CÔNG TRÌNH LIÊN QUAN	10
2.1. DỮ LIỆU Y SINH.....	10
2.1.1. DNA, hệ gene, gene, protein.....	11
2.1.2. DNA tái tổ hợp.....	12
2.1.3. Codon đồng nghĩa (Synonymous Condon).....	13
2.1.4. Hệ thống biểu hiện.....	14
2.1.5. Định danh loài sinh vật	15
2.1.6. Dữ liệu lâm sàng, cận lâm sàng	16
2.2. CÁC NGHIÊN CỨU LIÊN QUAN CÓ SỬ DỤNG THUẬT TOÁN HỌC MÁY CHO DỮ LIỆU Y SINH.....	17
2.2.1. Rút gọn chiều	18
2.2.2. Phương pháp học tập không giám sát	18
2.2.3. Phương pháp học tập giám sát	19
2.2.4. Phương pháp học máy học kết hợp.....	19
2.3. CÁC NGHIÊN CỨU LIÊN QUAN	20
2.3.1. Nghiên cứu về ứng dụng các mô hình học máy trong các việc giải quyết các vấn đề trong sinh học phân tử.....	20
2.3.2. Nghiên cứu về việc áp dụng các mô hình học máy trong các chẩn đoán bệnh dựa trên dữ liệu lâm sàng.....	20
2.3.3. Các thuật toán học máy hiệu quả của ứng dụng y sinh trong các bài toán đề xuất.	22
2.4. ĐÁNH GIÁ MÔ HÌNH MÔ HÌNH HỌC MÁY	29
2.5. DỮ LIỆU THỰC NGHIỆM.....	32
2.6. KẾT CHƯƠNG	34
CHƯƠNG 3. MÔ HÌNH HỌC MÁY TÌM GENE CHO HỆ THỐNG BIỂU HIỆN TRONG KỸ THUẬT DNA TÁI TỔ HỢP	35
3.1. GIỚI THIỆU.....	35

3.2. BÀI TOÁN TÌM GENE BIỂU HIỆN CAO (HIGHLY EXPRESSED GENE - HEG)	37
3.2.1. Bài toán tìm HEG	37
3.2.2. Phương pháp giải quyết	37
3.2.3. Kết quả thực nghiệm.....	41
3.3. BÀI TOÁN TÌM HỆ THỐNG BIỂU HIỆN PHÙ HỢP VỚI GENE MỤC TIÊU	42
3.3.1. Phát biểu bài toán.....	42
3.3.2. Phương pháp giải quyết	43
3.3.3. Xử lý dữ liệu thực nghiệm	45
3.3.4. Thực nghiệm mô hình dự đoán gene tương quan	46
3.4. KẾT LUẬN	49
CHƯƠNG 4. MÔ HÌNH ĐỊNH DANH LOÀI SINH VẬT.....	51
4.1. GIỚI THIỆU	51
4.1.1. Định danh loài sinh vật	51
4.1.2. Giới thiệu định danh loài nấm	52
4.1.3. Định danh loài nấm môi bằng phương pháp học máy	52
4.2. MÔ HÌNH ĐỊNH DANH LOÀI NẤM BẰNG KỸ THUẬT HỌC KẾT HỢP.....	56
4.3. CÁC THUẬT TOÁN HIỆU QUẢ CHO MÔ HÌNH ĐỊNH DANH LOÀI NẤM	57
4.4. KẾT QUẢ THỰC NGHIỆM.....	59
4.4.1. Xây dựng tập dữ liệu cho mô hình định danh loài nấm.....	59
4.4.2. Đánh giá hiệu năng của mô hình đề xuất.....	61
4.5. GIAO DIỆN MÔ HÌNH DỰ ĐOÁN TÊN LOÀI	63
4.6. KẾT LUẬN	64
CHƯƠNG 5. MÔ HÌNH HỌC MÁY CHO CHẨN ĐOÁN BỆNH DỰA TRÊN DỮ LIỆU CẬN LÂM SÀNG.	65
5.1. GIỚI THIỆU	65
5.2. MÔ HÌNH DỰ ĐOÁN BỆNH DỰA TRÊN DỮ LIỆU CẬN LÂM SÀNG	65
5.2.1. Giới thiệu.	65
5.2.2. Mô hình dự đoán bệnh dựa trên dữ liệu lâm sàng	67
5.2.3. Kết quả thực nghiệm.....	71
5.3. MÔ HÌNH PHÂN LOẠI BỆNH COVID-19 VÀ BỆNH CÚM MÙA	75
5.3.1. Giới thiệu	75
5.3.2. Mô hình phân biệt bệnh CoViD-19 và Cúm H1N1	76
5.3.3. Các thuật toán dùng trong mô hình đề xuất	77
5.3.4. Kết quả thực nghiệm.....	79
5.3.5. So sánh hiệu năng	87
5.4. KẾT LUẬN	89
CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	91
6.1. KẾT LUẬN	91
6.2. HƯỚNG PHÁT TRIỂN.....	92

DANH MỤC TỪ VIẾT TẮT

STT	Viết tắt	Tiếng Anh	Ý nghĩa
1	BLAST	Basic Local Alignment Search	Tên phần mềm sinh học BLAS
2	BN	Bayesian Network	Mạng Bayes
3	BOLD	Barcode Of Life Data	Tên ngân hàng gene quốc tế chuyên cung cấp mã định danh loài.
4	CAI	Codon Adaptation Index	Chỉ số thích nghi codon
5	CNN	Convolutional Neural-Networks	Mạng nơ-ron tích chập
6	COI	Cytochrome Oxidase I	Vùng phiên mã nội của động vật.
7	DL	Deep Learning	Máy học sâu
8	DNN	Deep Neural Networks	Mạng nơ-ron
9	DT	Decision Tree	Cây quyết định
10	EMBL	European Molecular Biology Laboratory	Tên ngân hàng gene của châu âu
11	GBDT	Gradient Boosted Decision Tree	Cây quyết định tăng cường độ dốc
12	GOI	Gene of Interest	Gene quan tâm, hay gene mục tiêu
13	HEG	Highly Expressed Gene	Gene cho biểu hiện biểu hiện cao
14	KNN	K-Nearest Neighbors	K-hàng xóm gần nhất.
15	LR	Logistic Regression	Hồi quy logistic
16	LASSO	Least Absolute Shrinkage and Selection Operator	Phân tích hồi quy thực hiện cả lựa chọn biến và chính quy hóa.
17	ML	Machine Learning	Học máy.
18	MLP	Multilayer Perceptron	Phương pháp học máy có giám sát nhiều lớp thuộc Mạng nơ-ron nhân tạo
19	NCBI	National Center For Biotechnology Information	Tên ngân hàng gene quốc tế
20	NN	Neural Network	Mạng thần kinh

21	RNN	Recurrent Neural Networks	Mạng nơ-ron hồi quy
22	RSCU	Relative Synonymous Codon Usage	Chỉ số codon đồng nghĩa
23	SNP	Single Nucleotide Polymorphism	Đa hình đơn nucleotide là loại biến dị di truyền phổ biến nhất ở người
24	SVM	Support Vector Machine	Máy véctor hỗ trợ
25	XGBDT	Extreme Gradient Boosted Decision Tree	Cây quyết định tăng cường độ dốc cực cao

DANH MỤC HÌNH

Hình 1.1: Quy trình tổng quát của bài toán nghiên cứu 1,2.....	4
Hình 1.2: Quy trình tổng quát của bài toán nhiên cứu.....	6
Hình 1.3: Quy trình tổng quát của bài toán nhiên cứu.....	7
Hình 2.1: Minh họa quá trình từ DNA đến protein của sinh vật.	12
Hình 2.2: Minh họa hình thành DNA tái tổ hợp từ Plasmid và DNA mục tiêu [3].....	13
Hình 2.3: <i>Minh họa các codon đồng nghĩa.</i>	14
Hình 2.4: a. Hệ thống phân loại sinh học các bậc phân loại chính; b. Cây phân loại dựa vào gene	16
Hình 2.5: Quy trình tổng quát hướng tiếp cận áp dụng học máy trong chẩn đoán bệnh CoViD-19 dựa trên dữ liệu lâm sàng [30]	21
Hình 2.6: Các xu hướng sử dụng thuật toán ML trong chẩn đoán bệnh CoViD-19 dựa trên dữ liệu cận lâm sàng [30].....	21
Hình 2.7: Minh họa kỹ thuật xây dựng cây quyết định thuật toán RF.	24
Hình 2.8: Minh họa kỹ thuật xây dựng cây quyết định thuật toán GBDT.	25
Hình 2.9: Mô hình xây dựng cây phân loại với thuật toán XGBoost.....	26
Hình 2.10: Minh họa kỹ thuật EFB trong thuật toán LightGBM	28
Hình 3.1: <i>Các giai đoạn tiến hành thực kỹ thuật tái tổ hợp</i>	36
Hình 3.2: Quy trình tổng quát của bài toán tìm gene biểu hiện cao, và tìm host cell phù hợp với gene mục tiêu.....	37
Hình 3.3: . Minh họa cho biểu diễn gene với RSCU	38
Hình 3.4: . Thống kê thời gian thực thi PAM, CLARA	42
Hình 3.5: Hình ảnh phân cụm bằng CLARA	42
Hình 3.6: 5% HEG được chọn.....	42
Hình 3.7: Các giai đoạn trong quá trình sản xuất protein tái tổ hợp;	43
(i)Lựa chọn một gene mong muốn, (ii) phân lập gene và cắt gene bằng các enzyme hạn chế,	
(iii)Gene tách được gắn vào một vector tạo dòng (plasmid); (iv) Chọn tế bào vật chủ để nhân	

dòng các plasmid; (v) Sử dụng gene đã nhân dòng đưa vào một ‘hệ thống biểu hiện thích hợp để thu protein mong muốn.....	43
Hình 3.8: Minh họa cho biểu diễn tập gene của các HostCell với RSCU.....	44
Hình 3.9: Sơ đồ xây dựng mô hình dự đoán gene tương quan với tế bào vật chủ.	45
Hình 3.10: Quy trình xử lý dữ liệu tổng quát [47].....	46
Hình 3.11: Confusion Matrix của mô hình đề xuất [47]	47
Hình 3.12: Hiệu năng về độ chính xác của mô hình đề xuất [47]	47
Hình 3.13: Kết quả đặc trưng quan trọng dùng trong mô hình dự đoán [47].....	48
Hình 4.1: Cấu trúc của gene ITS [54].....	54
Hình 4.2: Quy trình dự đoán tên loài nấm mốc.....	57
Hình 4.3: Minh họa cách tính k-mer cho chuỗi trình tự với k=7.....	59
Hình 4.4: Đường cong ROC sử dụng phương pháp OvR macro-average cho mỗi lớp trong phương pháp XGBoost với kích thước K-mer=7.....	61
Hình 4.5: Minh họa giao diện định hỗ trợ người dùng sử dụng thuật toán [61]	63
Hình 5.1: Mô hình tổng quan đề xuất cho bộ phân phân lớp bệnh CoViD-19.....	67
Hình 5.2: Số mẫu dữ liệu của hai lớp dữ liệu [66]	71
.....	72
Hình 5.3: Mối tương quan giữa các thuộc tính khảo sát.....	72
Hình 5.4: Mối tương quan giữa các thuộc tính khảo sát.....	73
Hình 5.5: So sánh hiệu năng ROC của các thuật toán tăng cường độ dốc	74
Hình 5.6: Hiệu năng ROC của các thuật toán đề xuất.....	75
Hình 5.7: Mức độ quan trọng chi tiết của các thuộc tính từ của mô hình đề xuất.....	75
Hình 5.8: Mô hình tổng quan đề xuất phân biệt COVID-19 và Cúm H1N1.....	76
Hình 5.9: ROC curve cho mô hình dự đoán	86
Hình 5.10: Precision recall curve cho mô hình dự đoán.....	86
Hình 5.11: Confusion matrix cho mô hình dự đoán.	86
Hình 5.12: Thuộc tính quan trọng của mô hình dự đoán.....	87

DANH MỤC BẢNG

Bảng 2.1: Bảng so sánh năng lực biểu hiện của các hệ thống biểu hiện protein mục tiêu trong quá trình sản xuất sản phẩm protein tái tổ hợp	15
Bảng 2.2: Bảng tổng hợp các nghiên cứu về chẩn đoán bệnh CoViD-19	22
Bảng 2.3: Dữ liệu y sinh được sử dụng trong các nhiệm vụ của luận án	33
Bảng 3.1: Thống kê hiệu năng của mô hình đề xuất [47]	48
Bảng 3.2: Độ chính xác của hai thuật toán tương ứng với các mô hình dự báo [47]	49
Bảng 4.1: Các nghiên cứu về định danh nấm bằng phương pháp Học máy	55
Bảng 4.2: Các nhãn của mỗi loài nấm mốc	60
Bảng 4.3: Tập dữ liệu được sử dụng làm tập kiểm tra là các mẫu nấm được thu thập tại tỉnh Bình Dương	60
Bảng 4.4: So sánh hiệu năng các thuật toán học máy trên cùng bộ dữ liệu.....	61
Bảng 4.5: So sánh hiệu năng với các bộ phân lớp về nấm cũng sử dụng trình tự ITS.	62
Bảng 4.6: Kết quả so sánh việc định danh loài của trình tự ITS của nấm mốc thu thập tại tỉnh Bình Dương, Việt Nam với định danh trên NCBI.....	62
Bảng 5.1: Tổng hợp các nghiên cứu cùng mục tiêu dự đoán bệnh nhân CoViD-19 bằng mẫu máu [66]	67
Bảng 5.2: Thông tin chi tiết các thuộc tính lâm sàng của các mẫu máu từ dữ liệu bệnh viện Israelita Albert Einstein [66]	68
Bảng 5.3: Bảng so sánh hiệu năng của mô hình đề xuất với tác giả Maryam AlJame	74
Bảng 5.4: Thuộc tính của bộ dữ liệu D_{org}	82
Bảng 5.5: Kết quả so sánh giữa việc lựa chọn đặc trưng của mô hình của luận án và mô hình của Li [12].....	84
Bảng 5.6: . So sánh hiệu suất của hai mô hình.	88

CHƯƠNG 1. TỔNG QUAN

Trong chương này, luận án trình bày sự cần thiết của các mô hình học máy trong các ứng dụng bài toán y sinh. Tiếp theo là mục tiêu, đối tượng, phạm vi, phương pháp nghiên cứu và các đóng góp của luận án. Cuối cùng là bố cục của luận án.

1.1. Tính cấp thiết của luận án

Việc áp dụng trí tuệ nhân tạo và học máy (Machine Learning-ML) trong các hoạt động sản xuất đã mang lại những hiệu quả ưu việt cho hầu hết mọi ngành nghề trong cuộc sống, đặc biệt lĩnh vực chăm sóc sức khỏe [1]. Kết hợp với sự phát triển vượt bậc của công nghệ phân tích trình tự, đã thúc đẩy nghiên cứu lĩnh vực sinh học cấu trúc, sinh học phân tử, y sinh học cùng phát triển. 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]. Học máy có khả năng phân tích dữ liệu y tế để xây dựng các mô hình dự đoán bệnh. 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. Điều này giúp cải thiện chẩn đoán, tăng khả năng tiên lượng và giảm chi phí chăm sóc y tế. Học máy có thể phân tích dữ liệu về cấu trúc di truyền, hóa học, tác dụng của thuốc và dữ liệu lâm sàng để tìm kiếm các phân tử mới và phát hiện các liên kết giữa các thuốc và bệnh lý. Điều này có thể giúp tăng tốc quá trình phát triển thuốc và tìm ra những phương pháp điều trị mới. 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 xây dựng mô hình có hiệu quả 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 và bị thiếu 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.

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

Hiện tại, các nghiên cứu các mô hình học máy trong lĩnh vực y sinh tập trung vào hai nhóm dữ liệu lớn là sinh học phân tử và dữ liệu lâm sàng. Việc xử lý hiệu quả hai loại dữ liệu 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 mới cho các tiếp cận nâng cao hiệu quả cho mô hình học máy trong dữ liệu y sinh. 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 chẩn 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 chẩn đoán bệnh.

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.

Để thực hiện nghiên cứu luận án phân tích, tổng hợp các nghiên cứu có liên quan đến nội dung nghiên cứu từ tài liệu tham khảo: sách, luận án công bố trên tạp chí và kỹ yếu hội thảo có liên quan về các xử lý gene, rút trích đặc trưng gene, các phương pháp dùng học máy chẩn đoán bệnh CoViD-19 từ dữ liệu lâm sàng.

Phương pháp thực nghiệm được sử dụng để đề xuất các tiếp cận mới nhằm nâng cao độ chính xác của mô hình học máy cho các bài toán về phân lớp gene, chẩn đoán bệnh. Tính hiệu quả của các mô hình phân lớp được chứng minh bằng kết quả thực nghiệm trên dữ liệu thực được lấy từ các kho dữ liệu Ngân hàng Gene quốc tế NCBI, từ các luận án công bố trên các tạp chí về tin sinh học. Để đánh giá tính tổng quát của các mô hình đề xuất luận án đã thực nghiệm

trên cùng bộ dữ liệu với các nghiên cứu khác đã công bố và luận án tiến hành thu thập, xử lý dữ liệu, xây dựng, huấn luyện và đánh giá mô hình. Các mô hình đề xuất được đánh giá bằng cách so sánh kết quả độ chính xác phân lớp với các mô hình cơ bản khác. Ngoài ra, luận án cũng đo thêm thời gian thực hiện của các mô hình để phân tích và đánh giá hiệu năng giữa các mô hình

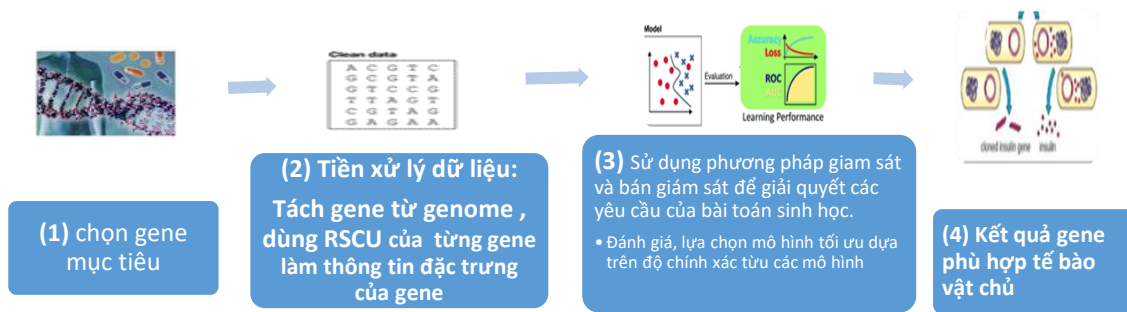
1.3. Nhiệm vụ của luận án

Mặc dù đã có nhiều nghiên cứu thực hiện áp dụng học máy cho các nhiệm vụ về dữ liệu genomic hay chẩn đoán bệnh. Việc ứng dụng học máy cho dữ liệu genomic cũng đặt ra những thách thức mới như: việc xử lý dữ liệu lớn, đảm bảo tính nhất quán và đáng tin cậy của kết quả phân tích. Tuy nhiên, đặc điểm số chiều lớn và sự hỗn độn của đặc trưng về dữ liệu y sinh làm hạn chế độ chính xác của mô hình y sinh. Vì vậy, hiện nay nhiều nghiên cứu vẫn tiếp tục thực hiện [1] để tìm ra các mô hình học máy hiệu quả hơn. 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

Cơ chế xử lý dữ liệu chiều cao cho dữ liệu sinh học phân tử trong các bài toán tìm gene, phân loại gene trong thiết kế thuốc. Trong các ứng dụng sinh học sử dụng kỹ thuật DNA tái tổ hợp cho lĩnh vực thiết kế thuốc, hormone, vaccine, enzyme hay cải tạo giống cây trồng đặc biệt là các loại thuốc điều trị các loại bệnh thiếu hụt nội tiết tố, hormone, tiểu đường. Trong kỹ thuật tạo ra DNA tái tổ hợp, tiến hành chọn lọc hai nguồn DNA của sinh vật khác nhau để tạo ra *gene mục tiêu* dùng cho sản xuất thuốc, hormone, vaccine, enzyme hay cải tạo giống cây trồng [3, 4]. Tuy nhiên trong quá trình sản xuất loại protein này có nhiều vấn đề xảy ra không mong muốn như: sản lượng thấp, protein không kích hoạt, protein ở thể không tan, protein bị bất hoạt. Do đó nhu cầu cần tìm gene mục tiêu tốt, cũng như môi trường tế bào vật chủ phù hợp nhất cho việc tạo ra sản phẩm tái tổ hợp tốt [5]. Dữ liệu của từng gene mục tiêu có độ dài khác biệt dao động từ 20 nucleotides đến 6 ngàn nucleotides. Hệ gene (genenome) sinh vật dùng trong kỹ thuật tái tổ hợp nhỏ nhất cũng chứa 4 ngàn gene. *Luận án sử dụng các phương pháp học máy để giải quyết vấn đề tìm nhóm gene cho biểu hiện protein cao và lựa chọn môi trường*

tế bào vật chủ phù hợp với gene mục tiêu. Đề xuất này nhằm mục tiêu nâng cao hiệu quả sản xuất được phẩm bằng công nghệ tái tổ hợp [6, 7]. Phương pháp học máy tổng quát để giải quyết nhiệm vụ này được trình bày như hình 1.1:



Hình 1.1: Quy trình tổng quát của bài toán nghiên cứu 1,2

Gene biểu hiện cao (Highly Expressed Gene, HEG) [5] [3] 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. 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. Các HEG 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. Việc lựa chọn gene mong muốn để sản xuất protein tái tổ hợp thường dựa trên các thông tin về tính chất của gene và sản phẩm của nó, ví dụ như cấu trúc của protein, đặc tính sinh học và vị trí trên gene. Ngoài ra, những gene được cho là HEG có thể được ưu tiên lựa chọn để tăng hiệu suất sản xuất protein trong hệ thống tái tổ hợp.

Phương pháp của Pere Puigbò [8] và cộng sự năm 2007 để tìm HEG dựa trên sự khác biệt về sử dụng codon giữa những gene mã hóa protein và những gene không mã hóa protein. Họ sử dụng chỉ số thích nghi codon CAI (Codon Adaptation Index- CAI) để đo lường mức độ ưu tiên sử dụng một loại codon nhất định trong một gene so với một tập hợp các gene tham chiếu. CAI được tính bằng cách so sánh tần suất sử dụng thực tế của mỗi codon trong gene đang xét với tần suất dự kiến của mỗi codon trong một tập hợp các gene tham chiếu có cùng điều kiện môi trường và đặc tính sinh học. Theo Pere Puigbò, một gene có CAI cao hơn rất nhiều so với phân phối của CAI trong tập hợp các gene thì nó có thể được xem là một HEG. Có thể chọn CAI tối thiểu để xác định một gene là HEG, và ngưỡng này được điều chỉnh dựa trên sự khác biệt về phân phối CAI giữa các loài. Số lượng gene biểu hiện cao theo phương pháp này chiếm khoảng 5% số gene trong bộ gene. Tuy nhiên, phương pháp này chỉ có thể áp dụng cho các loài có đầy

đủ thông tin về hệ gene và đặc tính sinh học của chúng. Ngoài ra, nó cũng có thể bỏ sót một số HEG do chúng không có sự khác biệt về sử dụng codon so với các gene khác. Để có thể tự động tìm tập HEG từ tập hệ gene bất kỳ, thì luận án đề xuất mô hình học máy để chọn lựa HEG từ hệ gene. Phương pháp tiến hành như sau:

(1). Biểu diễn thông tin đặc trưng của từng gene trong hệ gene dựa trên chỉ số codon đồng nghĩa RSCU (Relative Synonymous Codon Usage) [9]. Đây cũng là phương pháp rút gọn chiều dữ liệu của gene từ 20 nucleotides đến 6 ngàn nucleotides thành còn 64 đặc trưng.

(2). Sử dụng thuật toán phân cụm PAM, CLARA trên tập dữ liệu ở bước (1)

(3). Thay đổi giá trị k cụm, đánh giá chất lượng phân cụm theo biến đổi của tham số k

(4) Chọn HEG từ cụm k có giá trị tốt nhất

Bằng cách số hóa hệ gene của sinh vật có thể áp dụng các phương pháp gom cụm bằng mô hình học máy để tự động chọn lựa được tất cả HEG của hệ gene cho kỹ thuật tái tổ hợp trong thiết kế thuốc. Đây là nhiệm vụ đầu tiên của luận án.

Sau khi đã xác định được nhóm gene cho biểu hiện protein cao gọi là HEG, bước tiếp theo là cần xác định gene mục tiêu phù hợp (tương quan) với các loại tế bào vật chủ nào. Điều này là rất quan trọng để đảm bảo protein tái tổ hợp được biểu hiện tốt [5]. Mô hình xác định gene mục tiêu tương quan với tế bào vật chủ (cụ thể là hệ gene) là vấn đề thứ hai cần giải quyết của luận án. Phương pháp thực hiện nhiệm vụ này như sau:

(1). Biểu diễn thông tin đặc trưng của từng gene trong hệ gene dựa trên chỉ số codon đồng nghĩa RSCU

(2). Áp dụng các thuật toán phân lớp ensemble-learning cho dữ liệu bước (1)

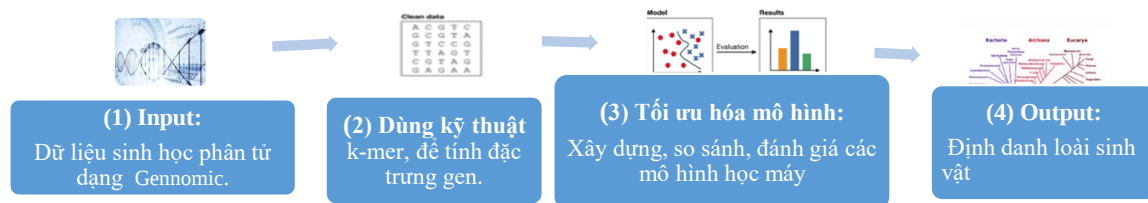
(3). Đánh giá mô hình phân lớp.

(4) Chọn mô hình phân lớp tối ưu.

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.

Xác định tên loài sinh vật là một là một nhiệm vụ quan trọng của ứng dụng y sinh. Điều này giúp 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. Tuy nhiên phương pháp dựa vào sin học phân tử được dùng phổ biến nhất hiện nay. Phương pháp này đánh giá sự tương quan giữa các trình tự genomic hoặc protein để xây dựng các cây phát sinh loài (phylogenetic tree) phản ánh quan hệ tiến hóa giữa các loài. Phương pháp này dựa vào việc dò tìm trình tự để đánh giá sự tương đồng giữa, việc này có thể được thực hiện bằng nhiều thuật toán khác nhau, trong đó các thuật toán so sánh trình tự như Neighbor Joining (NJ) và phương pháp khoảng cách, phương pháp phân cụm [10, 11]. Các thuật toán này hiệu quả trên so sánh từng trình tự, nhưng 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 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 này được mô tả bằng hình 1.2.

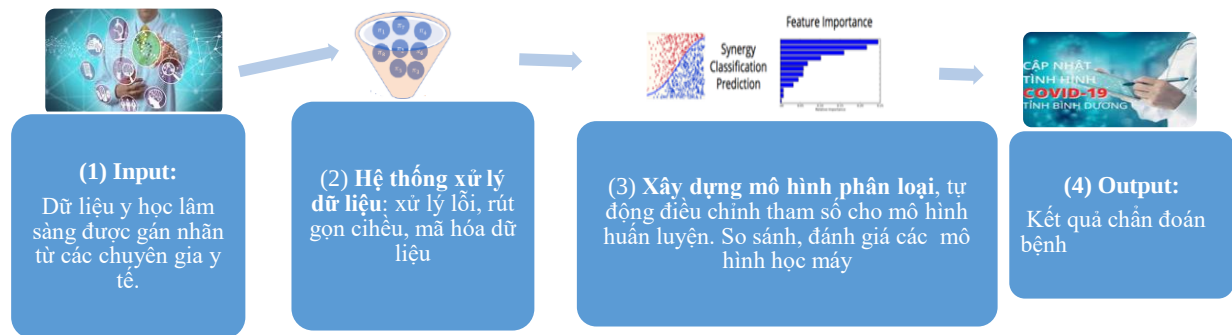


Hình 1.2: Quy trình tổng quát của bài toán nhiên cứu

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

Dữ liệu lâm sàng từ quá trình xét nghiệm chẩn đoán bệnh từ bệnh viện hay hồ sơ y tế cũng được xem là nguồn dữ liệu quý và rất hữu ích cho quá trình huấn luyện trong học máy cho nhiệm vụ chẩn đoán bệnh trong y tế. Đặc biệt khi có xảy ra dịch bệnh mô hình học máy này càng phát huy sức mạnh trong việc hỗ trợ bác sĩ trong chẩn đoán và phân tầng điều trị bệnh nhân. Điều này mang lại nhiều cơ hội cứu chữa cho bệnh nhân mắc bệnh nặng trong bối cảnh nguồn lực y tế khám chữa bệnh trong trạng thái quá tải. Thách thức chính của nhiệm vụ này là các loại dữ liệu lâm sàng có nhiều định dạng, thuộc tính bị thiếu dữ liệu, trùng lặp, chiều cao, số mẫu lớn và xuất hiện hiện tượng mất cân bằng dữ liệu trong các lớp phân loại. Dữ liệu thu được từ các xét nghiệm lâm sàng này có thể được xem là tập dữ liệu phức tạp và thuộc loại BigData [1, 2]. Xử lý dữ liệu y sinh là công việc đòi hỏi phải cần có chuyên môn về y tế. Một

trong các nhiệm vụ quan trọng của công đoạn này là xác định nhãn cũng như các đặc tính lâm sàng mô tả về bệnh. Bên cạnh đó dữ liệu này cần phải được xem xét nhiều yếu tố phụ thuộc lẫn giữa các đặc tính trước khi tiến hành rút gọn chiều. Mã hóa dữ liệu phân loại cũng cần được xem xét vì có thể dẫn đến việc làm sai lệch kết quả dự đoán. Chọn lựa thuật toán phân loại phù hợp, cũng như điều chỉnh các tham số cũng là vấn đề thách thức trong nhiệm vụ phân loại bệnh. Để giải quyết các thách thức của nhiệm vụ này, luận án đề xuất phương pháp tổng quát như hình 1.3.



Hình 1.3: Quy trình tổng quát của bài toán nhiên cứu

Ở giai đoạn (1): lựa chọn các tập dữ liệu y sinh được công bố từ các bệnh viện, các trung tâm xét nghiệm y khoa và được gán nhãn chẩn đoán từ các chuyên gia y tế về điều trị bệnh, cụ thể bệnh CoViD-19. Đây là một loại bệnh do virus SARS-COV-2 gây ảnh hưởng nghiêm trọng đến sức khỏe của người dân và nền kinh tế toàn cầu. Nguồn dữ liệu về CoViD-19 đã được nghiên cứu cẩn thận, quy trình lấy mẫu, gán nhãn dữ liệu đều được tiến hành trên dữ liệu thực tế tại các trung tâm y tế. Trong nghiên cứu luận án đã sử dụng các nguồn dữ liệu được chia sẻ bởi các tác giả [12] [13] [14].

Giai đoạn (2): đề xuất mô hình tự động xử lý dữ liệu có thể học trực tiếp từ dữ liệu thô mà không cần phải xóa bỏ cột có dữ liệu trống. Mô hình được đề giai đoạn (2) đánh giá, xử lý dữ liệu và rút gọn thuộc tính của tập dữ liệu bằng phương pháp LightGBM [15] (Light Gradient Boosting Model);

Giai đoạn (3): xây dựng mô hình phân loại từng nhóm bệnh dựa trên phương pháp XGBoost [16] (Extreme Gradient Boosting).

Giai đoạn (4): giai đoạn này sử dụng mô hình vừa xây dựng để dự đoán các mẫu xét nghiệm mới với tập thuộc tính đã rút gọn.

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, cũng như thực nghiệm đã được thực hiện trong luận án đã đề xuất cách giải quyết các nhiệm vụ nghiên cứu. Các đóng góp chính của luận án bao gồm:

Thứ nhất, để 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 RSCU của gene để biểu diễn thông tin đặc trưng từng gene dùng làm tập dữ liệu đầu vào cho hai nhiệm vụ 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 [CT2][CT3].

Thứ hai, để 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 của gene để biểu diễn thông tin đặc trưng từng gene, làm tin gọn tập dữ liệu đầu vào cho nhiệm vụ 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 dựa tên các đoạn gene chỉ định theo loài. Và luận án đã xây dựng mô hình dự đoán tên của 17 chi thuộc loài nấm mốc. Mô hình có độ chính xác 98%, và AUCROC đạt 91%. [CT4][CT7].

Thứ ba, sử dụng phương pháp kết hợp các thuật toán Gradient Boosting để xây dựng mô hình tự động rút gọn tính năng cho tập dữ liệu lâm sàng trong nhiệm vụ chẩn đoán bệnh. Cụ thể luận án xây dựng mô hình dự đoán bệnh nhân nhiễm bệnh CoViD-19, và bệnh cúm dựa trên bộ dữ liệu lâm sàng về hai nhóm bệnh 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

Chương này trình bày tổng quan về việc vận dụng mô hình học máy trong các bài toán có dữ liệu genomic và dữ liệu y sinh. Đầu tiên luận án trình bày các khái niệm sinh học, y học cho hướng tiếp cận dữ liệu y sinh. Tiếp theo luận án mô tả kiến trúc của các thuật toán của mô hình phân lớp dữ liệu biểu gene tổng quát và đề cập ý nghĩa và thách thức quan trọng của các bài toán này trong lĩnh vực học máy. Nội dung trình bày trong chương này cũng đã được công bố trong các công trình [CT1], [CT2], [CT3], [CT4].

2.1. Dữ liệu y sinh

Tin sinh học (Bioinformatics) là lĩnh vực nghiên cứu về việc sử dụng các công nghệ máy tính và thuật toán để phân tích và hiểu các dữ liệu sinh học như: hệ gene học, chuỗi nucleotide, protein, dữ liệu genomic, dữ liệu người bệnh, v.v... [10] Tin sinh học giúp phân tích và đánh giá tác động của các yếu tố di truyền đến sức khỏe con người và các loài sinh vật khác, tìm kiếm các phân tử được phẩm mới và phát triển các phương pháp chẩn đoán và điều trị bệnh.

Tin y sinh (Medical Informatics) [10] là lĩnh vực sử dụng các công nghệ thông tin và trí tuệ nhân tạo trong y học để cải thiện chất lượng chăm sóc sức khỏe và quản lý thông tin y tế. Tin y sinh bao gồm các ứng dụng như hồ sơ bệnh án điện tử, hệ thống hỗ trợ quyết định y tế, dữ liệu y tế và dữ liệu sinh lý học, v.v... Tin y sinh cải thiện khả năng chẩn đoán và điều trị bệnh, tăng cường quản lý thông tin y tế và giảm thiểu sai sót trong chăm sóc sức khỏe. Hiện tại trong lĩnh vực tin y sinh, dữ liệu y sinh đang rất được quan tâm là dữ liệu genomic, dữ liệu di truyền, dữ liệu tế bào, dữ liệu hình ảnh y học, dữ liệu lâm sàng. Cụ thể các nhóm dữ liệu về tin y sinh [17] [18]:

- Dữ liệu Genomic: Các dữ liệu genomic bao gồm thông tin về DNA và RNA của các sinh vật [3]. Các nguồn dữ liệu genomic từ các loài khác nhau được cung cấp bởi ba ngân hàng dữ liệu gene lớn GeneBank - NCBI (Ngân hàng CSDL sinh học của Mỹ), EMBL-Bank (CSDL của châu Âu về trình tự nucleotide) và DDBJ – DNA Data Bank of Japan (Ngân hàng CSDL DNA của Nhật).
- Dữ liệu di truyền: Đây là dữ liệu liên quan đến di truyền và biểu hiện gene.

- Dữ liệu hình ảnh y học: Đây là các tập hợp hình ảnh y học như hình ảnh siêu âm, hình ảnh CT Scan (Computed Tomography- chụp cắt lớp vi tính), MRI (Magnetic Resonance Imaging- chụp cộng hưởng từ).
- Dữ liệu lâm sàng: Đây là thông tin liên quan đến bệnh nhân, bao gồm thông tin về triệu chứng, kết quả xét nghiệm, lịch sử bệnh và phản ứng với điều trị.

Các nghiên cứu ứng dụng khoa học máy tính vào các lĩnh vực cho sự chăm sóc sức khỏe cộng đồng thường tập trung vào tìm các giải pháp tính toán tốt nhất cho các nhu cầu dự đoán đích thuốc, điều tra nguồn gốc sinh vật, cơ chế hình thành bệnh hay dự đoán bệnh tật. Việc vận dụng giải pháp học máy cho các bài toán này giúp nâng cao chất lượng ra quyết định cho quá trình sản xuất thuốc cũng như mang lại nhiều cơ hội được chữa bệnh tốt hơn cho cộng đồng. Luận án đã tiến hành nghiên cứu các giải pháp học máy hiệu quả trên hai nhóm dữ liệu genomic và dữ liệu lâm sàng.

2.1.1. DNA, hệ gene, gene, protein

Trong dữ liệu genomic bao gồm phần lớn các đối tượng như DNA, hệ gene và gene. Đây là các khái niệm quan trọng trong lĩnh vực di truyền học và sinh học phân tử. Dưới đây là giải thích về mỗi khái niệm:

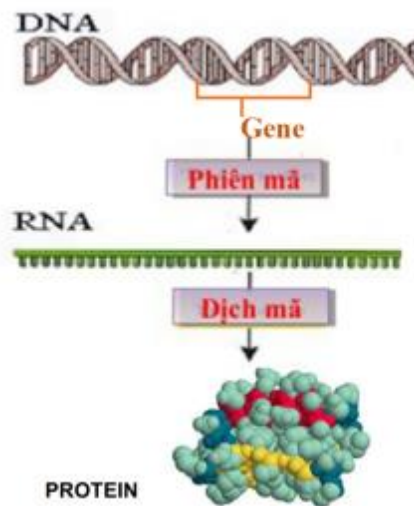
DNA (Deoxyribonucleic acid) [10]: DNA chứa thông tin di truyền cần thiết để điều chỉnh và điều hợp các hoạt động của tế bào. DNA được xem như là bộ nhớ di truyền của tất cả các sinh vật sống, từ vi khuẩn đến động vật và cây cối. DNA là chuỗi xoắn kép gồm 2 mạch đơn, mỗi mạch đơn là một chuỗi nucleotide. Cấu trúc của DNA là một chuỗi liên tiếp các nucleotide, trong đó mỗi nucleotide bao gồm một đường gốc đường xoắn kép (sugar deoxyribose) và một trong bốn loại nucleobase, gồm adenine (A), thymine (T), cytosine (C) và guanine (G). Hai chuỗi nucleotide trong DNA quấn quanh nhau tạo thành một cấu trúc xoắn ốc gọi là double helix (xoắn kép đôi).

Hệ gene (genome) [10]: là toàn bộ tập hợp các gene có mặt trong một tế bào hay một sinh vật. Nó đại diện cho toàn bộ thông tin di truyền được lưu trữ trong DNA của một sinh vật. Hệ gene bao gồm tất cả các gene đơn lẻ và các vùng không mã hóa khác nhau trên DNA.

Gene: Gene là một đơn vị cơ bản của di truyền học, được định nghĩa là một đoạn DNA cụ thể trên một chuỗi DNA. Mỗi gene chứa thông tin cần thiết để tổ chức và điều chỉnh một

tính chất hoặc một số tính chất cụ thể trong một sinh vật. Gene chứa mã hóa cho các sản phẩm protein, nhưng cũng có thể chịu trách nhiệm cho các hoạt động khác như điều chỉnh hoạt động gene khác. Chức năng của gene được thể hiện chính qua cơ chế tạo ra protein dựa trên thông tin được mã hóa trong gene.

Protein [10]: Protein là kết quả của việc tổng hợp protein từ gene. Quá trình này gồm hai giai đoạn chính là phiên mã (transcription) và dịch mã (translation). Có thể hình dung mối quan hệ giữa DNA, gene, Protein được mô tả như hình Hình 2.1. Protein chiếm phần lớn cấu trúc của tế bào và đóng vai trò quan trọng trong nhiều hoạt động tế bào như: cung cấp sự hỗ trợ cơ học và định hình cho tế bào; làm chất xúc tác trong các phản ứng sinh hóa của tế bào được gọi là Enzyme; điều chỉnh quá trình sinh lý trong cơ thể và tương tác với các tế bào như là Hormone v.v... Có tổng cộng 20 loại amino acid được sử dụng để xây dựng các protein. Trong số này, có 9 loại được gọi là "amino acid thiết yếu" vì chúng không thể được tổng hợp bởi cơ thể và phải được cung cấp qua chế độ ăn uống. Những amino acid thiết yếu bao gồm histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan và valine. Chúng cần thiết cho sự tăng trưởng, sửa chữa và duy trì sự hoạt động bình thường của cơ thể con người.

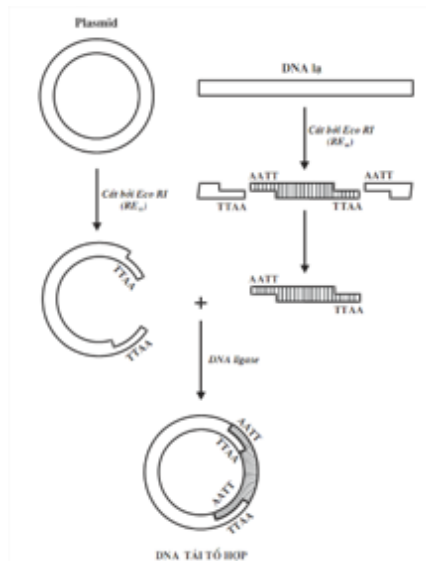


Hình 2.1: Minh họa quá trình từ DNA đến protein của sinh vật.

2.1.2. DNA tái tổ hợp

DNA tái tổ hợp (DNA recombination) được nghiên cứu thành công từ những năm đầu của thập niên 1970 [3]. DNA tái tổ hợp là quá trình mà các đoạn DNA được cắt và ghép lại với

nhau để tạo ra các đoạn DNA mới. Trong quá trình tái tổ hợp, các đoạn DNA được cắt bởi các enzyme cắt DNA, sau đó các đoạn này được ghép lại với nhau để tạo ra các đoạn DNA mới. Quá trình này có thể xảy ra trong cùng một phân tử DNA hoặc giữa các phân tử DNA khác nhau. Hình 2.2 Minh họa hình thành DNA tái tổ hợp từ Plasmid và DNA mục tiêu, và Hình 2.3 mô tả quá trình tạo ra các sản phẩm của DNA tái tổ hợp. Kỹ thuật này thường được dùng trong các lĩnh vực y sinh, dược phẩm. Các sản phẩm tái tổ hợp phổ biến thường là protein tái tổ hợp được sử dụng như thuốc, hormone, vaccine, enzyme hay cải tạo giống cây trồng-vật nuôi.



Hình 2.2: Minh họa hình thành DNA tái tổ hợp từ Plasmid và DNA mục tiêu [3]

2.1.3. Codon đồng nghĩa (Synonymous Condon)

Trong sinh học, các codon đồng nghĩa (Synonymous Condon) [3] có vai trò quan trọng trong quá trình dịch mã gene, tạo ra chuỗi protein cần thiết cho sự sống của tế bào và sinh vật. Các codon đồng nghĩa có thể ảnh hưởng đến tốc độ và chất lượng của quá trình dịch mã gene. Nghiên cứu đã chỉ ra rằng sự sử dụng của các codon đồng nghĩa khác nhau có thể ảnh hưởng đến tốc độ dịch mã gene và cũng có thể ảnh hưởng đến cấu trúc và chức năng của protein được tạo ra.

Cơ sở khoa học của việc tối ưu hóa gene dựa trên hiện tượng codon đồng nghĩa. Một codon gồm ba nucleotide có $4^3 = 64$ loại codon. Thực tế là có 61 loại codon mã hóa cho 20 loại amino acid khác nhau. Còn 3 loại codon còn lại không mã hóa cho bất kỳ amino acid nào, được gọi là các codon stop hoặc codon dừng, chúng có tác dụng làm kết thúc quá trình dịch mã gene và

tạo ra các protein đầy đủ. Vì vậy, mỗi loại amino acid có thể được mã hóa bởi nhiều loại codon khác nhau, có nghĩa là các codon đồng nghĩa có thể mã hóa cho cùng một loại amino acid. Ví dụ, Amino acid *Ala (hay A)* có bốn codon đồng nghĩa là GCA, GCC, GCG, GCU.

	CGU									UUA					UCU				
	CGC									UUG					UCC				
GCU	CGA						GGU			CUU				CCU	UCA	ACU		GUU	
GCC	CGG						GGC		AUU	CUC				CCC	UCG	ACC		GUC	UAA
GCA	AGA	AAU	GAU	UGU	CAA	GAA	GGA	CAU	AUC	CUA	AAA		UUU	CCA	AGU	ACA		UAU	GUA
GCG	AGG	AAC	GAC	UGC	CAG	GAG	GGG	CAC	AUA	CUG	AAG	AUG	UUC	CCG	AGC	ACG	UGG	UAC	GUG
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V

Hình 2.3: Minh họa các codon đồng nghĩa.

Một đại lượng quan trọng được dùng trong đánh giá độ ưu tiên sử dụng codon gọi là “**tần suất của codon tối ưu**” [19], liên quan tới độ đa dạng của thành phần tRNA trong tế bào, một số lượng lớn các đại lượng tính toán sự ưu tiên codon được phát triển cho thấy tầm quan trọng về ý nghĩa sinh học của xu hướng sử dụng codon. Có hai cách thức để tính toán đại lượng này [19] là: **chỉ số sử dụng codon đồng nghĩa** - RSCU (Relative Synonymous Codon Usage-RSCU) và CAI. Công thức được dùng để biểu diễn gene trong hệ gene bằng kỹ thuật RSCU [8].

$$r_{ac}(g) = \frac{o_{ac}}{\frac{1}{k_a} \sum_{c \in C_a} o_{ac}} = \frac{o_{ac}}{\bar{o}_{ac}} \quad (2.1)$$

Trong đó:

r_{ac} : RSCU của codon c mã hoá cho amino acid a

O_{ac} : tần suất xuất hiện của codon c trong trình tự gene

C_a : tập hợp các codon cùng mã hoá cho amino acid a

K_a : số lượng các loại codon c mã hóa cho amino acid a

Đối với các codon kết thúc (UAA, UGA, UGA) và các codon không có codon đồng nghĩa (AUG – mã hóa cho Methionine, UGG – mã hóa cho Tryptophan). Do đó, gene được biểu diễn bằng tập các RSCU là một vector RSCU 59 chiều có dạng như sau $r(g) = \{r_1, r_2, r_3 \dots r_{59}\}^T$

2.1.4. Hệ thống biểu hiện

Trong kỹ thuật tái tổ hợp (recombinant DNA technology), hệ thống biểu hiện được sử dụng để sản xuất các protein đặc biệt hoặc các sản phẩm khác từ DNA tái tổ hợp. Hệ thống biểu hiện bao gồm một loại tế bào chủ (host cell) được sử dụng để sản xuất protein từ gene được tái tổ

hợp [3]. Các tế bào chủ có thể là tế bào vi khuẩn hoặc tế bào động vật như tế bào cột sống (sử dụng trong công nghệ sinh học protein) hoặc tế bào côn trùng (sử dụng trong sản xuất protein recombinant cho dược phẩm). Các tế bào này được sử dụng để sản xuất protein từ DNA tái tổ hợp được đưa vào trong chúng. Các tế bào chủ được chọn dựa trên tính chất của protein cần sản xuất và các yếu tố khác như khả năng sinh trưởng, giá thành và tiện lợi trong việc đưa gene vào tế bào chủ. Sau khi gene được tái tổ hợp được đưa vào tế bào chủ, quá trình biểu hiện protein bắt đầu thông qua quá trình dịch mã gene và các bước xử lý protein.

Việc khảo sát các hệ thống biểu hiện cần được lựa chọn cẩn thận trước khi đưa vào thực nghiệm sản xuất protein tái tổ hợp. Bảng so sánh các tính năng đặc biệt của hệ thống biểu hiện được báo cáo tổng kết ở bảng 2.1 như sau:

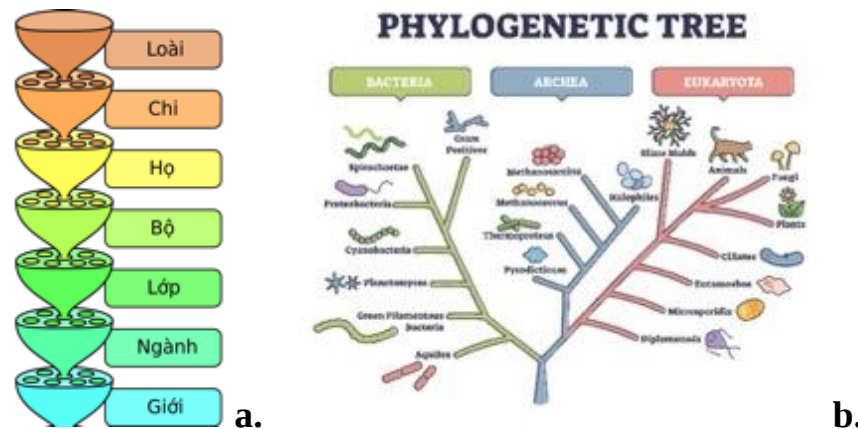
Bảng 2.1: Bảng so sánh năng lực biểu hiện của các hệ thống biểu hiện protein mục tiêu trong quá trình sản xuất sản phẩm protein tái tổ hợp

Hệ thống biểu hiện	Giá trị sản xuất	Thời gian sản xuất	Khả năng mở rộng sản xuất	Mức độ biểu hiện	Nguy cơ nhiễm bẩn	Giá thành bảo quản
Vi khuẩn	Thấp	Ngắn	Cao	Cao	Nội độc tố	Trung bình
Nấm men	Trung bình	Trung bình	Cao	Thấp -cao	Nguy cơ thấp	Trung bình
Dòng tế bào côn trùng	Cao	Trung bình	Trung bình	Thấp -cao	Cao	Đắt
Dòng tế bào động vật có vú	Cao	Dài	Rất thấp	Thấp - Trung bình	Virus, các prion	Đắt
Động vật chuyển gene	Cao	Rất dài	Thấp	Trung bình -cao	Virus, các prion	Đắt
Dòng tế bào thực vật	Thấp	Ngắn	Cao	Trung bình -cao	Nguy cơ thấp	Không đắt
Thực vật chuyển gene	Rất thấp	Dài	Rất cao	Trung bình -cao	Nguy cơ thấp	Không đắt

2.1.5. Định danh loài sinh vật

Cây phát sinh gene còn được gọi là cây tiến hóa, Cây mô tả việc phát sinh chủng loại, lịch sử tiến hóa của các loài hoặc nhóm được mô tả bằng sơ đồ đường [10]. Cây tiến hóa và các

nhánh của nó cho thấy mối quan hệ tiến hóa giữa các loài sinh vật khác nhau hoặc các nhóm khác. Phát sinh chủng loại dựa trên sự tương đồng và khác biệt về đặc điểm vật lý hoặc di truyền của chúng. Định danh loài được xây dựng dựa trên nguyên tắc phân loại nhất định, chia loài thành các nhóm phân loại như: loài, chi, họ, bộ, lớp, ngành. Các nhóm này được xếp hàng từ cấp độ thấp nhất (loài) đến cấp độ cao nhất (ngành). Hệ thống phân loại hiện đại dựa trên các khóa phân loại hình thái hoặc dựa trên nhưng phản ánh mối quan hệ tiến hóa giữa các loài. Hình 2.4 mô tả mối liên hệ giữa các loài sinh vật. Định danh loài có ý nghĩa quan trọng trong nhiều lĩnh vực. Đối với các nhà khoa học và nhà nghiên cứu, định danh loài giúp xác định và mô tả sự đa dạng sinh học, nghiên cứu quan hệ di truyền và tiến hóa, và phân tích mối tương quan giữa các loài.



Hình 2.4: a. Hệ thống phân loại sinh học các bậc phân loại chính; **b.** Cây phân loại dựa vào gene

Gene mã vạch (DNA barcode) là một đoạn gene chuẩn ngắn nằm trong bộ genome của sinh vật. Trong hầu hết các trường hợp, vùng gene không mã hóa là vùng gene rất đặc trưng so với các vùng gene khác. Ở động vật, đoạn DNA barcode được sử dụng cho phần lớn các loài là đoạn gene ở ty thể cytochrome C oxidase (CO1). Ở thực vật gene vùng nhân như ITS, ITS2 [20]. Gene mã vạch này đang được dùng trong rất nhiều nghiên cứu để phục vụ giám định loài.

2.1.6. Dữ liệu lâm sàng, cận lâm sàng

Dữ liệu lâm sàng (biomedical data) [21] bao gồm thông tin tổng quát và chi tiết hơn về sức khỏe và bệnh tật của bệnh nhân. Cụ thể các nhóm dữ liệu gồm: i) Hồ sơ y tế: các thông tin từ

việc khám bệnh, bệnh án, kết quả xét nghiệm, dấu hiệu và triệu chứng, lịch sử bệnh lý, chẩn đoán và điều trị của bệnh nhân; ii) Dữ liệu hình ảnh y tế: các hình ảnh từ các phương pháp hình ảnh y tế như tia X, siêu âm, cắt lớp vi tính (CT), hình ảnh cộng hưởng từ (MRI) và hình ảnh vô khuẩn (endoscopy); iii) Dữ liệu di truyền: các thông tin về các biến thể di truyền, gene, dấu hiệu di truyền và tiềm năng nguy cơ bệnh tật di truyền.

Dữ liệu cận lâm sàng (sub-biomedical data) là một tập con của dữ liệu lâm sàng và tập trung vào một loại bệnh trong lĩnh vực y học. Các nhóm dữ liệu cũng thường giống các dạng dữ liệu của lâm sàng.

Thu thập dữ liệu thường được gọi là thiết bị được sử dụng để lấy dữ liệu từ thử nghiệm lâm sàng, quy trình lấy mẫu dữ liệu luôn phải tuân thủ theo các nguyên tắc nghiêm ngặt khi tiến hành thực hiện. Chất lượng của dữ liệu thu được phụ thuộc hoàn toàn vào chất lượng của công cụ được sử dụng để thu thập dữ liệu. Hầu hết dữ liệu về lâm sàng trong nghiên cứu về y sinh được thu thập từ các công ty dược phẩm, các bệnh viện, hay hồ sơ khám chữa bệnh. Các phần mềm hỗ trợ chẩn đoán từ dữ liệu lâm sàng thường được dùng thường có hai nhóm: i) Phần mềm thương mại: ORACLE CLINICAL, MACRO, RAVE, CLINTRIAL, eClinicalSuite, EZ-entry; ii) Các phần dạng mở; OpenClinica, TrialIDB, openCDMS.

2.2. Các nghiên cứu liên quan có sử dụng thuật toán học máy cho dữ liệu y sinh

Tin sinh học là một lĩnh vực liên ngành, trong đó các phương pháp tính toán mới được phát triển để phân tích dữ liệu sinh học và thực hiện các khám phá sinh học. Trong di truyền học và genomics, các công cụ tin sinh học đã được áp dụng cho các quá trình giải trình tự và chú thích bộ gene. Trong tin sinh học có các loại dữ liệu đặc trưng omics như là: biểu hiện gene, biểu hiện microRNA, biến thể số lượng bản sao, đa hình nucleotide đơn (single nucleotide polymorphism – SNP). Các loại dữ liệu này thường được xem là loại Big Data với đặc điểm là loại dữ liệu phức tạp, chiều cao nên yêu cầu phải có phương pháp phân tích đặc biệt, quy trình đặc biệt và đòi hỏi thời gian xử lý nhiều hơn. Học máy liên quan đến việc phân tích và phát triển các mô hình giải quyết các vấn đề của các loại dữ liệu lớn thuộc nhiều lĩnh vực khác nhau. Do đó các vấn đề phức tạp của sinh học cũng được giải quyết bằng các thuật toán của học máy. Với thuật toán học máy có thể học từ dữ liệu để thực hiện phân tích dự đoán. “Việc học” này có thể hiểu là mô hình học có khả năng tự điều chỉnh cấu hình của chúng để giải thích tối ưu dữ liệu quan sát được. Các kỹ thuật quan trọng của học máy được áp dụng trong y sinh học

được liệt kê theo [22] như sau: Rút gọn chiều, kỹ thuật học giám sát, kỹ thuật học không giám sát, Kỹ thuật học học phối hợp. Cụ thể các kỹ thuật này được trình bày trong cụ thể sau đây

2.2.1. Rút gọn chiều

Kỹ thuật rút gọn chiều (dimensionality reduction) là phương pháp giảm số chiều của dữ liệu mà không mất đi quan trọng thông tin của nó. Kỹ thuật rút gọn chiều được sử dụng để giảm kích thước của dữ liệu và giúp giảm thiểu sự phức tạp tính toán khi xử lý dữ liệu lớn. Trong đó, phương pháp PCA (Principal Component Analysis) là một trong những phương pháp rút gọn chiều phổ biến nhất. PCA giúp tìm ra các thành phần chính (principal components) của dữ liệu, đó là những thành phần có đóng góp lớn nhất vào sự biến thiên của dữ liệu. Các thành phần chính này được sắp xếp theo độ quan trọng giảm dần, với thành phần chính đầu tiên có đóng góp lớn nhất. Phương pháp PCA được áp dụng rộng rãi trong nhiều lĩnh vực như xử lý ảnh, xử lý tín hiệu, nhận dạng giọng nói và các lĩnh vực khác. Trong sinh học, PCA được sử dụng để phân tích và rút gọn số chiều của các tập dữ liệu sinh học, chẳng hạn như dữ liệu gene và dữ liệu khảo sát lâm sàng [23]. Đặc tính dữ liệu của các bài toán sinh học này là dữ liệu chiều cao, nên việc cần giảm kích thước của tập dữ liệu ở bước trước 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. Ví dụ, trong nghiên cứu genetic, PCA có thể được sử dụng để giảm số chiều của dữ liệu gene và tìm ra những gene quan trọng nhất trong quá trình phát triển bệnh hoặc điều trị [23]. Nó cũng có thể được sử dụng để phân tích dữ liệu dịch tễ học và tìm ra các yếu tố quan trọng nhất ảnh hưởng đến sức khỏe của các cộng đồng và cá nhân. Bên cạnh kỹ thuật rút gọn chiều bằng PCA, các phương pháp rút gọn chiều bằng các kỹ thuật thống kê, học máy cũng đáp ứng tốt việc giảm kích cỡ chiều. Luận án áp các kỹ thuật rút gọn chiều bằng kỹ thuật này trong các mô hình phân lớp đề xuất cho dữ liệu y sinh.

2.2.2. Phương pháp học tập không giám sát

Phân tích cụm là một trong đại diện của kỹ thuật học tập không giám sát dùng để xác định cấu trúc từ dữ liệu mà trước đây chưa có kiến thức nào trước đây chưa nhận định về tập dữ liệu huấn luyện. Tư tưởng chính của phân cụm dựa vào kết quả đo của độ đo tương tự và phương pháp định nghĩa độ đo này. Chẳng hạn như đối với tư tưởng chính của việc Phân hoạch, phân cấp, phân cụm dựa trên mật độ. Trong các bài toán tính sinh học phân cụm có ứng dụng rất lớn từ phân loại sinh vật đến phát hiện gene gây bệnh hay dùng để phân cụm bệnh nhân. Kỹ thuật

phân cụm rất hiệu quả trong rút trích thông tin cho bộ dữ liệu. Luận án giới thiệu ở mục 3.2 hai kết quả áp dụng dựa trên kỹ thuật phân cụm phân hoạch dùng 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 (Classification) là 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. Nếu biến tập biến y_i có nhiều giá trị thì được gọi mô hình bài toán là hồi quy (Regression). Luận án tập trung vào loại mô hình phân lớp.

Mô hình dự đoán trong bài toán sinh học còn có thể phân thành hai nhóm dựa vào kiểu mô hình đã biết trước tập x_i hay kiểu mô hình chưa xác định được tập tham số x_i . Mặt khác, các mô hình không tham số không giới hạn mối quan hệ giữa các tính năng đầu vào của tập dữ liệu và biến kết quả cho một họ hàm cụ thể. Độ phức tạp của các mô hình không tham số được tự động điều chỉnh trong bước đào tạo dữ liệu. Ứng dụng về phân lớp nhiều mức với tập tham số đầu vào được định nghĩa trước dùng để dự đoán hệ thống biểu hiện phù hợp với gene mục tiêu.

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. Phương pháp học kết hợp thường được sử dụng để tăng tính tổng quát hóa và giảm thiểu overfitting của một mô hình đơn lẻ. Có nhiều phương pháp học kết hợp theo kỹ thuật *Boosting* là một trong những phương pháp phổ biến và hiệu quả nhất của học phối hợp. Ý tưởng chính của *Boosting* [24] là lặp lại quá trình học của một bộ phân lớp yếu nhiều lần. Sau mỗi bước lặp, bộ phân lớp yếu tập trung học trên các phần tử bị phân lớp sai trong các lần lặp trước. Các thuật toán áp dụng cho kỹ thuật *Boosting* như AdaBoost (Adaptive Boosting), Gradient Boosting và XGBoost (Extreme Gradient Boosting). 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. Ví dụ, 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. Ngoài ra, kỹ thuật

Boosting cũng có thể được sử dụng để xác định các yếu tố quan trọng nhất trong dữ liệu sinh học, chẳng hạn như gene quan trọng trong quá trình phát triển bệnh.

2.3. Các nghiên cứu liên quan

2.3.1. Nghiên cứu về ứng dụng các mô hình học máy trong các việc giải quyết các vấn đề trong sinh học phân tử

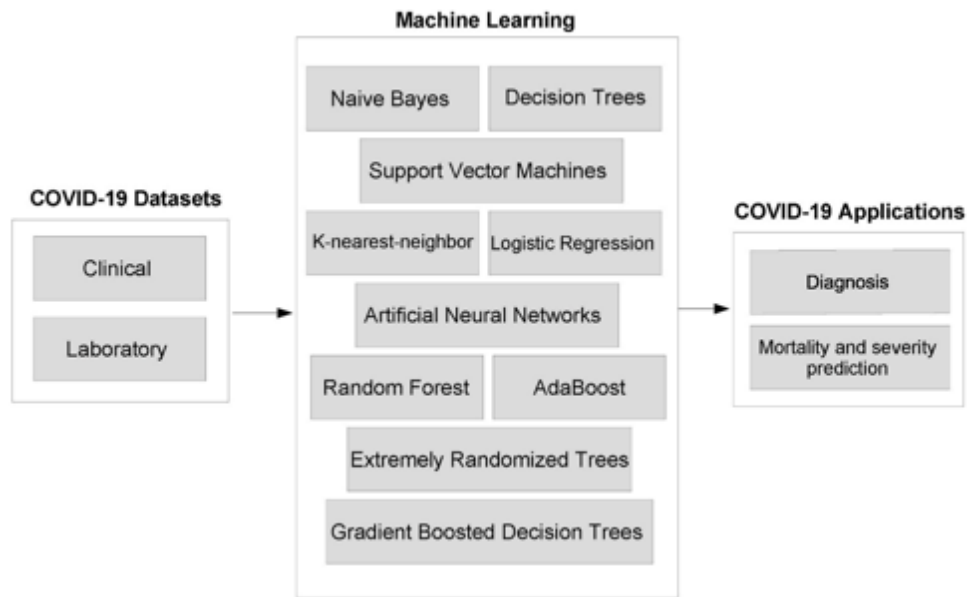
Việc nhận ra các mục tiêu thuốc mới là rất quan trọng trong khi tiến hành nghiên cứu khám phá thuốc để tìm ra các loại thuốc lâm sàng. Đã có nhiều thành công trong việc khám phá thuốc được hỗ trợ bởi ML, điều này cho thấy khả năng của các công ty tích hợp AI trong việc khám phá nhanh chóng các loại thuốc có thể ứng cử viên. Chẳng hạn như Deep genomics đã sử dụng nền tảng bàn 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].

Các nghiên cứu trước đây đã tuyên bố rằng các thuật toán khác nhau như: Extreme Learning Machines, DNN, rừng ngẫu nhiên (RF), máy vectơ hỗ trợ (SVM) có thể hỗ trợ dự báo trong độc tính và tương tác thuốc với sức khỏe con người [25] [26].

Bên cạnh đó, việc nghiên cứu các cấu trúc protein của virus (ví dụ như spike protein) cũng nhiệm vụ quan trọng của ML trong phát triển vắc-xin. Việc sử dụng ML để xác định thành phần nào có khả năng tạo ra phản ứng miễn dịch mạnh mẽ nhất [27]. Hơn nữa, sự phát triển của hệ thống ML 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]. Không những vậy điều này 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.

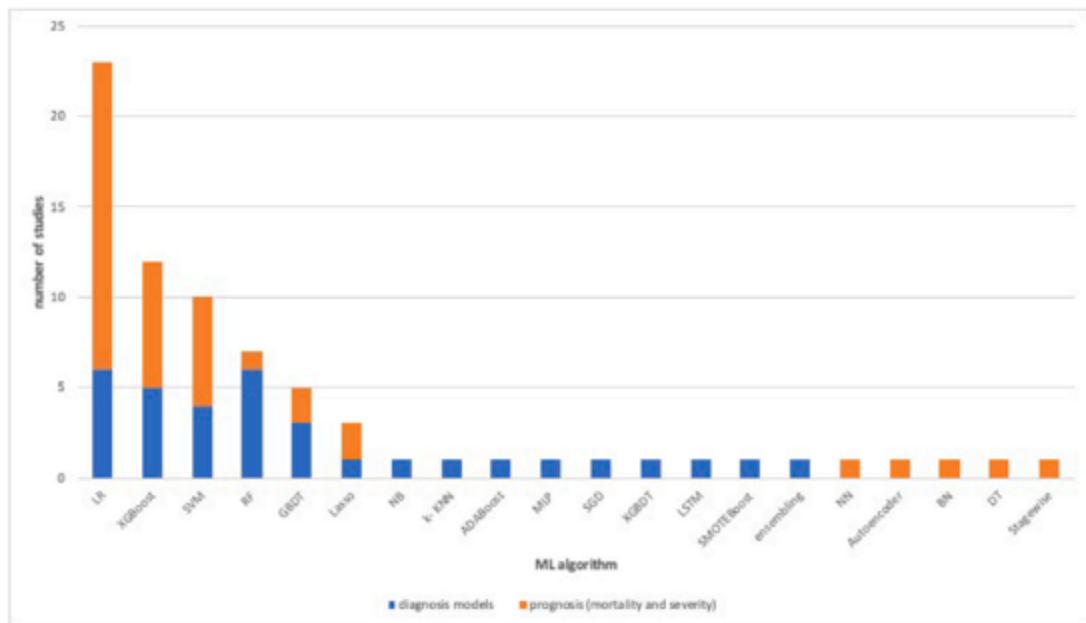
2.3.2. Nghiên cứu về việc áp dụng các mô hình học máy trong các chẩn đoán bệnh dựa trên dữ liệu lâm sàng.

Trong các đề xuất về mô ứng dụng học máy trong các chẩn đoán bệnh dựa trên dữ liệu lâm sàng được dùng gần đây, tác giả Norah Alballa [29], đã tổng kết xu hướng sử dụng cho các mô hình trong chẩn đoán bệnh CoViD-19 như hình 2.5



Hình 2.5: Quy trình tổng quát hướng tiếp cận áp dụng học máy trong chẩn đoán bệnh CoViD-19 dựa trên dữ liệu lâm sàng [29]

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].

Trong các công trình đã tham khảo [29] có cùng mục tiêu nghiên cứu với luận án nhận thấy: thuật toán cho kết quả dự đoán cao nhất là của tác giả Patrick Schwab [30] sử dụng thuật toán SVM cho mô hình học máy chẩn đoán bệnh CoViD-19 và khả năng nhập viện, độ chính xác đạt 98% trên bộ dữ liệu bệnh viện Israelita Albert Einstein ở Brazil [13]. Cùng cùng bộ dữ liệu này tác giả Maryam AlJame [31] chọn lọc các mẫu dữ liệu xét nghiệm máu và áp dụng thuật toán XGBoost cho mô hình dự đoán bệnh COVID-19 cho có kết quả mô hình tốt hơn với độ chính xác là 99,4%. Chi tiết thống kê về các nghiên cứu cùng mục tiêu với luận án được mô tả chi tiết ở bảng 2.2.

Bảng 2.2: Bảng tổng hợp các nghiên cứu về chẩn đoán bệnh CoViD-19

Đặc điểm các nghiên cứu	Tác giả của nghiên cứu	Mục đích	Kỹ thuật học máy	Kích cỡ dữ liệu sử dụng trong mô hình	Hiệu năng
Độ chính xác thấp nhất	Liang et al [32]	Chẩn đoán bệnh CoViD-19, và khả năng nhập viện.	Decision tree-DT	10,504 (1353 hospitalized, 83 ICU admission)	AUC of 76%, accuracy 68%, and recall 71%
Số mẫu lớn nhất	Levy et al. [33]	Dự đoán bệnh nhân CoViD-19 nặng.	LASSO	11,095 patients (8499 survivors, 2596 nonsurvivors)	AUCs of 86%, 82%, and 82%, respectively for internal and
Số mẫu ít nhất	Han et al [34]	Dự đoán bệnh nhân CoViD-19 nặng.	LR- logistic regression.	47 patients (24 severe)	AUC of 95.28%
Độ chính xác thấp nhất	Patrick Schwab et al [30]	Chẩn đoán bệnh CoViD-19, và khả năng nhập viện.	LR, SVM- support vector machine	5644(553 patient)	AUC of 98%;

2.3.3. Các thuật toán học máy hiệu quả của ứng dụng y sinh trong các bài toán đề xuất.

Có ba bài toán có liên quan về sử dụng học máy trong các ứng dụng y sinh chiều cao như mô tả ở mục 1.2, kết hợp các khảo sát về các nghiên cứu liên quan như đã mô tả ở trên. Một số thuật toán học máy đã áp dụng hiệu quả như sau:

❖ Thuật toán PAM (Partitioning Around Medoids)

PAM là một thuật toán phân cụm (clustering) được sử dụng để tìm các cụm (clusters) trong dữ liệu [35]. Thuật toán PAM là một phiên bản cải tiến của thuật toán k-means và nó tập trung vào việc chọn các điểm dữ liệu làm "medoids" để đại diện cho các cụm. Các bước chính của

thuật toán PAM như sau: i) Khởi tạo: Chọn ngẫu nhiên k điểm dữ liệu làm các medoids ban đầu, với k là số lượng cụm được xác định trước. ii) Gán nhãn: Gán các điểm dữ liệu còn lại vào các cụm dựa trên khoảng cách Euclid giữa các điểm dữ liệu và medoids. Mỗi điểm dữ liệu được gán vào cụm có medoid gần nhất. iii) Cập nhật medoids: Với mỗi cụm, thử thay thế medoid hiện tại bằng một điểm dữ liệu khác trong cụm và tính toán tổng khoảng cách từ tất cả các điểm dữ liệu trong cụm đến medoid mới. Medoid mới được chọn là điểm dữ liệu có tổng khoảng cách nhỏ nhất. Lặp lại bước 2 và 3: Tiếp tục quá trình gán nhãn và cập nhật medoids cho đến khi không có sự thay đổi nào trong việc gán nhãn và medoids. iv) Kết thúc: Thuật toán dừng lại khi không có sự thay đổi trong việc gán nhãn và medoids. Các điểm dữ liệu được phân vào các cụm dựa trên medoids cuối cùng. Luận án sử dụng PAM cho việc phân cụm các gene, giải thuật PAM được sử dụng cho phân cụm gene được mô tả bằng giải thuật 2.1 như sau:

Thuật toán 2.1: sử dụng PAM áp dụng cho mô hình phân cụm gene

Input:

Tập dữ liệu D Tập hợp các chuỗi *gene* $G = \{g_1, g_2, \dots, g_n\}$,
Số cụm k

Output: Tập hợp gene đã được phân vào k cụm.

Begin

(1) Chọn tùy ý k gene giữ vai trò là các g_p ban đầu;

(2) Repeat

(3) Với mỗi g_p

Lần lượt xét các gene g_i không là g_p

Tính S là độ lợi khi hoán đổi g_p với g_i

$$S = E_{g_p} - E_{g_i}$$

(4) **If** $S < 0$ **then** swap g_i with g_p

(5) **until** no change;

End

Trong đó: E_{g_p} , E_{g_i} lần lượt là giá trị hàm mục tiêu trước và sau khi thay g_p bởi g_i

$$E = \sum_{i=1}^k \sum_{p \in C_i} d(gp, gi)^2$$

Thuật toán PAM giúp tìm ra các medoids tốt nhất để đại diện cho các cụm. So với thuật toán K-means, PAM có độ phức tạp tính toán cao hơn do cần tính toán khoảng cách tới tất cả các điểm dữ liệu trong cụm trong quá trình cập nhật medoids. Tuy nhiên, PAM thường cho kết quả tốt hơn trong các trường hợp dữ liệu có nhiễu hoặc các cụm không có hình dạng cầu. Đây cũng

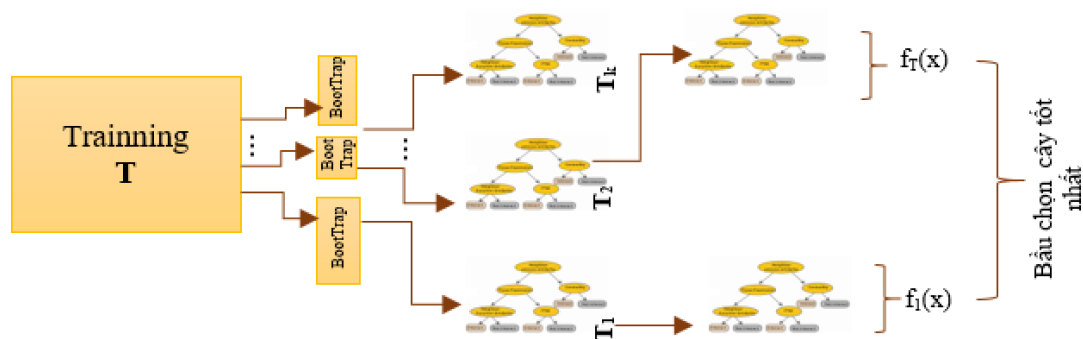
là đặc điểm nổi trội của thuật toán khi áp dụng cho các bài toán với dữ liệu trình tự gene [36]. Trong luận án đã áp dụng thuật toán này cho bài toán *tìm gene cho biểu hiện protein cao*.

❖ Thuật toán Clara (Clustering Large Applications)

Thuật toán Clara [37] là một thuật toán phân cụm dữ liệu không giám sát được sử dụng để xử lý các tập dữ liệu lớn. Thuật toán Clara bắt đầu bằng cách: i) chọn ngẫu nhiên một tập hợp con của dữ liệu ban đầu và áp dụng thuật toán PAM (Partitioning Around Medoids) trên tập hợp con này để tạo ra k cụm. Sau đó, thuật toán lặp lại hai bước cho đến khi không có sự thay đổi nào trong cụm: i) Áp dụng PAM trên một tập hợp con mới của dữ liệu, được lấy ra ngẫu nhiên từ dữ liệu ban đầu; ii) Đối với mỗi cụm, tính toán tổng khoảng cách của các điểm dữ liệu đến medoid của cụm và lưu trữ cụm có tổng khoảng cách nhỏ nhất. Trong luận án đã áp dụng thuật toán này cho bài toán *tìm gene cho biểu hiện protein cao*.

❖ Thuật toán RF (Random Forest RF)

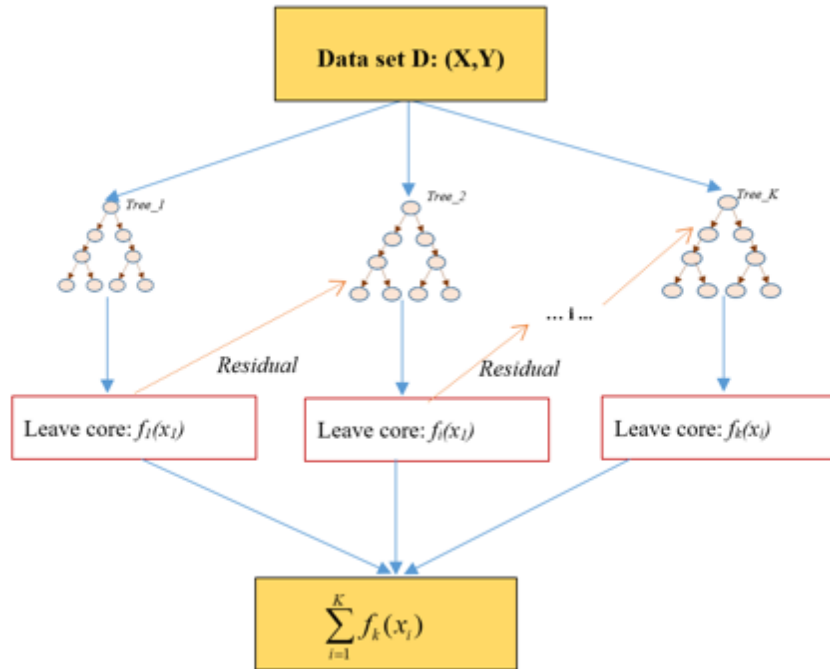
Thuật toán Random Forest [38] dựa trên phương pháp tập hợp mô hình (ensemble learning) của nhiều cây quyết định. RF sử dụng một tập hợp các cây quyết định (decision tree) để xây dựng một mô hình dự đoán. Giả sử tập dữ liệu có m phần tử $x_1, x_2, \dots, x_m$ trong không gian n chiều có lớp tương ứng là $y_1, y_2, \dots, y_m$. Mô hình cây quyết định T xây dựng cây bắt đầu từ nút gốc đến nút lá. Mỗi cây quyết định được xây dựng từ một tập dữ liệu con được lấy ngẫu nhiên từ tập dữ liệu huấn luyện ban đầu. Sau đó, các cây quyết định này được kết hợp lại với nhau để tạo ra một mô hình dự đoán cuối cùng. Phương pháp học máy này dựa trên tập hợp mô hình (ensemble learning) của nhiều cây quyết định, cách phân chia cây được như minh họa hình 2.7.



Hình 2.7: Minh họa kỹ thuật xây dựng cây quyết định thuật toán RF.

❖ Thuật toán Gradient Boosted Decision Tree (GBDT)

Thuật toán Gradient Boosted Decision Tree (GBDT) [24] là một phương pháp học máy kết hợp giữa hai kỹ thuật chính là *tăng cường độ dốc- Gradient boosting* và *cây quyết định-Decision tree*. GBDT xây dựng một loạt cây quyết định theo cách tuần tự, mỗi cây cố gắng cải thiện những sai số còn lại của mô hình trước đó. Giả sử chúng ta muốn xây dựng một mô hình GBDT với T cây và n mẫu thì quá trình dự đoán theo phương pháp GBDT được mô tả tổng quát như hình 28.



Hình 2.8: Minh họa kỹ thuật xây dựng cây quyết định thuật toán GBDT.

Quá trình huấn luyện GBDT bắt đầu bằng việc xây dựng một cây quyết định đơn giản, sau đó tính toán sai số giữa kết quả dự đoán và giá trị thực tế. Tiếp theo, một cây quyết định khác được xây dựng để dự đoán sai số còn lại. Quá trình này được lặp lại cho đến khi đạt được một số lượng cây quyết định đủ lớn hoặc đạt được một tiêu chí dừng nhất định.

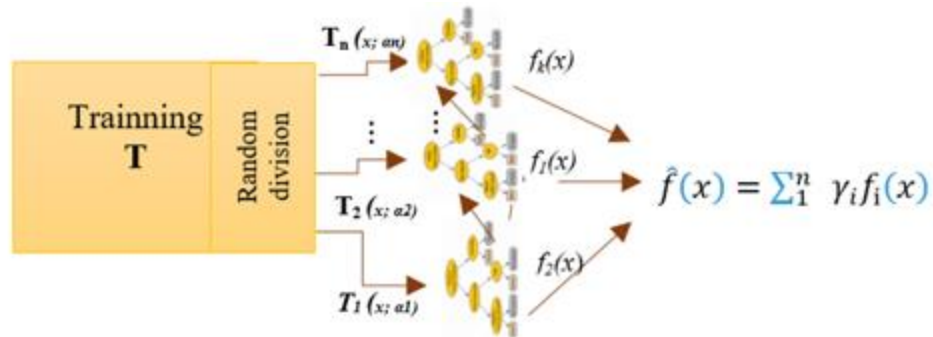
Ưu điểm của GBDT là nó có khả năng học được các mô hình phức tạp và có khả năng xử lý cả dữ liệu số và dữ liệu hạng mục. Nó cũng có khả năng xử lý các trường hợp dữ liệu thiếu và nhiễu. GBDT đã được ứng dụng rộng rãi trong nhiều lĩnh vực như dự đoán tín dụng, xử lý ngôn ngữ tự nhiên, nhận dạng hình ảnh, và nhiều bài toán khác trong lĩnh vực học máy và khoa học dữ liệu.

Các giá trị cây quyết định được trình bày theo công thức (1).

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(K)} &= \sum_{k=1}^K f_k(x_i) = \hat{y}_i^{(K-1)} + f_K(x_i)
\end{aligned} \tag{1}$$

❖ Thuật toán XGBoost (Extreme Gradient Boosting)

XGBoost là một phương pháp học máy dựa trên mô hình *Gradient Boosting Machine* của Jerome H. Friedman [24]. Tuy nhiên, XGBoost [38] [39] đã được phát triển và cải tiến để tăng tính hiệu quả và độ chính xác của mô hình. Sử dụng nhiều kỹ thuật tối ưu hóa để tăng tốc độ tính toán và giảm thiểu overfitting. Xây dựng các cây quyết định tuần tự, mỗi cây được xây dựng để tối thiểu hóa hàm mất mát cho các điểm dữ liệu bị phân loại sai bởi các cây trước đó. Quá trình xây dựng cây được mô tả tổng quát như hình 2.9.



Hình 2.9: Mô hình xây dựng cây phân loại với thuật toán XGBoost

Về cách xây dựng cây hồi quy trong XGBoost, nó sử dụng thuật toán *Split Finding* để tìm các điểm chia tốt nhất và tính toán giá trị dự đoán bằng cách thêm các giá trị dự đoán từ các cây trước đó. Các cây hồi quy được xây dựng theo kiểu tuần tự, mỗi cây được xây dựng để tối thiểu hóa hàm mất mát cho các điểm dữ liệu bị phân loại sai bởi các cây trước đó. Điều này cũng là điểm mạnh của XGBoost là tốc độ tính toán nhanh, khả năng xử lý dữ liệu lớn và độ chính xác cao. Đặc biệt với dữ liệu có đặc tính thừa thớt, Tianqi Chen đề xuất ý tưởng XGBoost cho áp dụng Gradient Boosting cho ma trận thưa với Thuật toán 2.2 *Split Finding* như sau:

Thuật toán 2.2: Split Finding**Input**

Tập hợp I , $I_k = \{i \in I | x_{ik} \neq \text{missing}\}$, X

Output: giá trị split .

Begin

(1). Đặt: $G \leftarrow \sum_{i \in I} g_i$ $H \leftarrow \sum_{i \in I} h_i$

$\hat{f}(x) = 0$, $r_i = y_i$ cho tất cả các i trong tập training

(2). **For** $k = 1$ **to** m **do**

//xử lý dữ liệu missing bên nhánh trái

$G_L \leftarrow 0, H_L \leftarrow 0$

For j **in** *sorted*(I_k , *ascent order* x_{jk}) **do**

$G_L \leftarrow G_L + g_j$, $H_L \leftarrow H_L + h_j$

$G_R \leftarrow G - G_L$, $H_R \leftarrow H - H_L$

$score \leftarrow \max\left(score, \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}\right)$

//xử lý dữ liệu missing bên nhánh phải

$G_R \leftarrow 0, H_R \leftarrow 0$

For j **in** *sorted*(I_k , *descent order* x_{jk}) **do**

$G_R \leftarrow G_R + g_j$, $H_R \leftarrow H_R + h_j$

$G_L \leftarrow G - G_R$, $H_L \leftarrow H - H_R$

$score \leftarrow \max\left(score, \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}\right)$

EndFor

EndFor

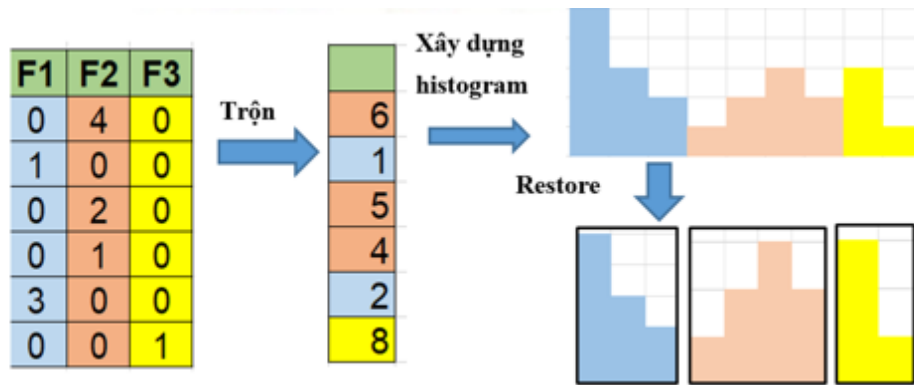
(3). Output hướng và giá trị Split

End

❖ Thuật toán LightGBM (Light Gradient Boost Machine)

LightGBM [40] là một phương pháp học máy được tạo ra bởi Microsoft Research Asia vào năm 2017. LightGBM được thiết kế để tăng tốc độ tính toán và giảm thiểu sử dụng tài nguyên so với các phương pháp học máy khác. LightGBM đã đạt được nhiều kết quả ấn tượng và đạt được giải thưởng trong các cuộc thi khoa học dữ liệu và học máy. Để có được hiệu quả ấn tượng như vậy là do LightGBM sử dụng một số kỹ thuật tối ưu hóa đặc biệt như Gradient-based One-Side Sampling (GOSS) và Exclusive Feature Bundling (EFB) để tăng tốc độ tính toán và giảm thiểu sử dụng bộ nhớ trong quá trình huấn luyện mô hình. GOSS là một kỹ thuật tối ưu hóa được sử dụng để lấy mẫu các điểm dữ liệu và giảm thiểu số lượng điểm dữ liệu được sử dụng trong quá trình huấn luyện mô hình. EFB là một kỹ thuật tối ưu hóa được sử dụng để

gom nhóm các đặc trưng tương tự với nhau để giảm thiểu sử dụng bộ nhớ trong quá trình huấn luyện mô hình. EFB được hiểu như *Gói tính năng độc quyền/Gói tính năng loại trừ lẫn nhau*. Đối với dữ liệu có các đặc trưng thừa chiều cao, nhiều đặc điểm loại trừ lẫn nhau (nghĩa là nhiều nhất một đặc điểm trong số nhiều đặc điểm có giá trị khác 0), EFB tạo thành một "đặc trưng lớn" bằng cách gộp nhiều đặc điểm loại trừ lẫn nhau, từ đó giảm đáng kể số lượng tính năng tương đương với phương pháp giảm kích thước dữ liệu. Như mô tả ở hình 2.10, giá trị của tính năng F1 là 0 ~ 10, giá trị của tính năng F2 là 0 ~ 20, F1 và F2 là các tính năng loại trừ lẫn nhau, sau đó F1/F2 được nhóm lại để tạo thành tính năng F3 và giá trị của tính năng F3 là 0 ~ 30, vậy F1=5 tương đương với F2=15.



Hình 2.10: Minh họa kỹ thuật EFB trong thuật toán LightGBM

Việc tìm số lượng kết hợp gói tối ưu trong LightGBM đã được giải quyết áp dụng thuật toán xấp xỉ tham lam. Nghĩa là, nó được thực hiện như một bài toán tô màu đồ thị. Các điểm trong đồ thị là các đặc điểm. Các đặc điểm không loại trừ lẫn nhau được kết nối bởi một cạnh. Trọng số của cạnh là tổng giá trị xung đột của các đặc điểm được kết nối. Khi đó, các điểm có cùng màu là các điểm (đặc điểm) loại trừ lẫn nhau.

LightGBM được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm cả xử lý ngôn ngữ tự nhiên, xử lý ảnh, dự đoán giá trị bất động sản, và nhiều lĩnh vực khác. Trong luận án LightGBM được dùng để toán ưu hóa dữ liệu đầu vào cho mô hình phân loại bệnh dựa trên dữ liệu cận lâm sàng.

❖ Thuật toán CatBoost (Categorical Boosting - CatBoost)

CatBoost [41] cũng là một thuật toán được phát triển dựa trên cây quyết định. Đây là thuật toán làm việc với tập dữ liệu mà có số lượng lớn thuộc tính đầu vào có dạng phân loại

(categorical type). Trong các bài toán sinh tin học đã sử dụng CatBoost cho nhiều mục đích như xác định vi khuẩn ở cấp độ gene trong trình tự 16S rRNA [42] hay dùng cho xây dựng gói trích xuất tính năng cho chuỗi DNA, RNA và protein [43]. Đề xuất của luận án đã sử dụng thuật toán CatBoost trong quá trình xây dựng mô hình định danh loài sinh vật.

2.4. Đánh giá mô hình mô hình học máy

Độ đo đánh giá (evaluation metrics): trong học máy là quá trình đánh giá độ chính xác và hiệu quả của một mô hình dự đoán trên tập dữ liệu kiểm tra. Mục tiêu của đánh giá mô hình là xác định độ chính xác và tính tổng quát hóa của mô hình trên các tập dữ liệu mới.

Các phương pháp đánh giá mô hình được sử dụng phụ thuộc vào bài toán cụ thể và tính chất của dữ liệu. Việc sử dụng nhiều phương pháp đánh giá khác nhau cũng giúp đảm bảo tính chính xác và độ tin cậy của kết quả đánh giá.

Với một bộ dữ liệu cụ thể khi sử dụng để phân lớp cần xác định đâu là dữ liệu dương tính (positive) mang tính khẳng định tích cực, đâu là dữ liệu âm tính (negative). Chẳng hạn, trong mùa dịch bệnh CoViD-19, dương tính để chỉ đó là kết quả xét nghiệm bị nhiễm virus SAR-CoV-2. Trong trường hợp này kết quả xét nghiệm của người bị bệnh thì dữ liệu này được coi là positive. Nhưng khi dịch bệnh trở thành bình thường, nên khi chọn người để giao công việc, thì yếu tố tích cực (positive) phải là không bị nhiễm SARS-CoV-2 virus; nên dương tính (positive) phải là không bị nhiễm virus.

Chính vì vậy, cần phải phân biệt rõ ràng đâu là dữ liệu positive, đâu là dữ liệu negative. Từ đó các ký hiệu viết tắt TP, FN, TN, FP được lý giải chính xác như sau khi sử dụng trong công thức tính hiệu năng của mô hình học máy:

- TP (True Positive): dương tính thực sự, hay kết quả dự đoán đúng là dương tính
- FN (False Negative): âm tính giả, hay dữ liệu là âm tính nhưng đoán nhầm là dương tính (negative nhưng đoán nhầm là positive)
- TN (True Negative): âm tính thực sự, hay kết quả dự đoán đúng là âm tính
- FP (False Positive): dương tính giả, hay dữ liệu là dương tính nhưng dự đoán là âm tính (positive nhưng đoán nhầm thành negative)

Để đánh giá hiệu suất của mô hình đề xuất, luận án đã sử dụng các phương pháp đo đặc độ chính xác của mô hình học máy bao gồm:

- ❖ **Accuracy:** Cho biết tỷ lệ các trường hợp được dự đoán đúng (TP và TN) trong tổng số cần dự đoán (cả positive và negative) của tập dữ liệu, Accuracy được tính theo công thức:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- ❖ **Sensitivity (Độ nhạy):** Hay còn gọi là True Positive Rate (TPR – Tỷ lệ dương tính thật), hoặc còn được gọi là Recall (PR), Hit Rate (Tỷ lệ trúng đích) nhằm xác định trong tất cả các trường hợp thực tế positive (TP + FN), thì có bao nhiêu dự đoán positive là đúng (TP).

$$TPR = Sensitivity = \frac{TP}{TP + FN}$$

- ❖ **Specificity (Độ đặc hiệu):** Hay True Negative Rate (TNR - Tỷ lệ âm tính thật). Đây là độ đo ngược với TPR: nhằm xác định trong tất cả các trường hợp thực tế negative (TN + FP), thì có bao nhiêu phần dự đoán negative là đúng (TN)

$$Specificity = \frac{TN}{TN + FP}$$

- ❖ **FPR (False Positive Rate/Fall-out):** Là tỷ lệ dương tính giả, nhằm xác định trong tất cả các trường hợp thực tế negative (TN + FP), thì có bao nhiêu phần dự đoán positive là sai (FP).

$$FPR = \frac{FP}{TN + FP} = \frac{TN + FP - TN}{TN + FP} = 1 - Specificity$$

- ❖ **Precision:** Hay Positive Predictive Value (PPV - trị số tiên lượng dương tính): nhằm xác định trong tất cả các dự đoán về positive (TP + FP), thì có bao nhiêu phần dự đoán positive là đúng (TP). Công thức tính được tính như sau:

$$Precision = \frac{TP}{TP + FP}$$

- ❖ **F1-score:** là một độ đo kết hợp giữa Precision và Recall của một mô hình phân loại. F1-score giúp đánh giá hiệu suất của mô hình phân loại bằng cách kết hợp cả khả năng dự đoán chính xác các trường hợp Positive (Precision) và khả năng tìm ra tất cả các trường hợp Positive trong tập dữ liệu (Recall). F1-score giá trị càng cao, mô hình phân loại càng tốt.

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

- ❖ **ROC (Receiver Operating Characteristics):** Nhằm mô tả mối liên hệ giữa Độ nhạy và Độ đặc hiệu (specificity). Biểu đồ ROC này chính là đồ thị của hàm $y = f(x)$ tạo nên bởi các giá trị tương ứng của giữa tỷ số dương tính giả với $x = FPR$ (là $1 - Specificity$) và tỷ số dương tính thật $y = TPR$ (Sensitivity). Vùng diện tích giới hạn trong vùng gồm: đồ thị ROC, trục hoành $y = 0$ và đường thẳng $x = 1$ gọi là chỉ số AUC (Area Under the Curve). Chỉ số AUC thể hiện hiệu suất phân loại của mô hình. Như vậy, biểu đồ ROC để hiển thị từng cặp (TPR, FPR) cho các ngưỡng khác nhau với mỗi điểm trên đường cong biểu diễn 1 cặp (TPR, FPR) cho 1 ngưỡng, từ đó suy ra chỉ số AUC cho đường cong này. Chỉ số AUC được tính bằng diện tích phía dưới đường cong ROC. Giá trị AUC càng gần 1, mô hình phân loại càng tốt. Đường cong ROC và chỉ số AUC thường được sử dụng trong các bài toán phân loại nhị phân, đặc biệt là khi tập dữ liệu bị mất cân bằng giữa các lớp.
- ❖ **Đường cong Precision-Recall:** là một biểu đồ biểu diễn quan hệ giữa Precision và Recall trong các bài toán phân loại. Precision (độ chính xác) là tỷ lệ giữa số lượng dự đoán đúng thuộc một lớp cụ thể và tổng số dự đoán thuộc lớp đó, trong khi Recall (độ phủ) là tỷ lệ giữa số lượng dự đoán đúng thuộc một lớp cụ thể và tổng số mẫu thực tế thuộc lớp đó. Đường cong Precision-Recall thể hiện mức độ cân bằng giữa Precision và Recall khi ngưỡng quyết định thay đổi.
- ❖ **K-fold cross-validation (CV):** Là một phương pháp phân tích thống kê đã được các nhà nghiên cứu sử dụng rộng rãi để đánh giá hiệu suất của bộ phân loại học máy. K-Fold Cross Validation (CV) là một phương pháp đánh giá mô hình và đo lường độ chính xác của mô hình trong học máy. K-Fold CV chia tập dữ liệu thành K phần bằng

nhau, gọi là fold, và thực hiện quá trình huấn luyện và kiểm tra trên K lần. Cụ thể, quá trình K-Fold CV có các bước sau:

- (i) Chia tập dữ liệu thành K phần bằng nhau, mỗi phần được gọi là fold;
- (ii) Lần lượt chọn một fold làm tập kiểm tra và sử dụng các fold còn lại để huấn luyện mô hình;
- (iii) Đánh giá độ chính xác của mô hình trên fold kiểm tra bằng cách tính toán giá trị của một số thông số độ chính xác như độ chính xác (accuracy), độ phân loại chính xác (precision), độ phủ (recall) hoặc F1-score;
- (iv) Lặp lại quá trình 2 và 3 K lần, mỗi lần chọn một fold khác làm tập kiểm tra.
- (v) Tính trung bình của các giá trị độ chính xác được tính toán trong K lần để đánh giá độ chính xác của mô hình.

Quá trình K-Fold CV giúp đánh giá độ chính xác của mô hình một cách chính xác hơn và giảm thiểu tác động của việc chọn ngẫu nhiên các tập huấn luyện và kiểm tra. Ngoài ra, K-Fold CV cũng giúp phát hiện ra một số vấn đề như học quá dư thừa (overfitting) hoặc học chưa đủ (underfitting) của mô hình.

2.5. Dữ liệu thực nghiệm

Hiểu về cơ chế hoạt động của DNA, gene, protein giúp quá trình đạt hiệu quả cao. Trong quá trình điều chế thuốc việc tối ưu hóa protein để thiết kế protein có các hoạt tính sinh học cụ thể. Ở đây luận án tập trung cho việc tối ưu hoá protein. Còn tối ưu hóa protein là quá trình thiết kế và tạo ra protein với các đặc tính cụ thể, chẳng hạn như chức năng, độ ổn định hoặc độ hòa tan. Cụ thể luận án đã thực hiện nhiệm vụ này bằng cách chọn lựa gene có khả năng cho biểu hiện Protein cao, Tìm được tế bào vật chủ phù hợp với gene mục tiêu để nâng cao hiệu quả biểu hiện protein của gene mục tiêu. 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 mốc. 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. Chi tiết cụ thể về nhóm dữ liệu được trình bày trong bảng 2.3.

Bảng 2.3: Dữ liệu y sinh được sử dụng trong các nhiệm vụ của luận án

Stt	Tên	Số lượng	Nhiệm vụ
1	<i>B.Subtilus</i>	4100	Tìm gene bằng kỹ thuật phân cụm với thuật toán PAM, CLARA [CT2]
2	<i>Ecoli (Class 1)</i>	4288	Tìm hệ thống biểu hiện phù hợp với tế bào vật chủ [CT3]
3	<i>B.Subtilus (Class 2)</i>	4100	
4	<i>Latococcus (Class 3)</i>	2264	
5	<i>Uncultured Termitomyces</i>	799	Định danh loài sinh vật: dữ liệu thực nghiệm dựa vào 17 loài nấm mối với số lượng trình tự tối thiểu cho mỗi lớp là 10 [CT4][CT7]
6	<i>Termitomyces sp.</i>	483	
7	<i>Termitomyces intermedius</i>	94	
8	<i>Termitomyces symbiont</i>	60	
9	<i>Termitomyces microcarpus</i>	34	
10	<i>Termitomyces clypeatus</i>	33	
11	<i>Termitomyces cylindricus</i>	30	
12	<i>Termitomyces striatus</i>	29	
13	<i>Termitomyces DKA-2007</i>	24	
14	<i>Termitomyces heimii</i>	24	
15	<i>Termitomyces bulborhizus</i>	17	
16	<i>Termitomyces fuliginosus</i>	16	
17	<i>Termitomyces eurrhizus</i>	15	
18	<i>Termitomyces albuminosus</i>	14	
19	<i>Termitomyces sp. symbiont of Macrotermes bellicosus</i>	12	
20	<i>Termitomyces sp. symbiont of Macrotermes subhyalinus</i>	10	
21	Bệnh giống cúm CoViD-19	5564	Nhiệm vụ dự đoán mắc bệnh CoViD-19 [CT5].
22	Bệnh CoViD-19	553	

23	Bệnh Cúm H1N1	1072	Nhiệm vụ phân biệt bệnh Cúm H1N1, mắc bệnh CoViD-19 [CT6].
24	Bệnh CoViD-19	413	

Nâng cao hiệu quả của các mô hình học máy trong lĩnh vực y sinh là mục tiêu chính của luận án, các đề xuất chính tập trung và hai nhóm dữ liệu quan trọng là dữ liệu genome và dữ liệu lâm sàng. Chi tiết về dữ liệu được sử dụng trong các mô hình đề xuất được mô tả trong bảng 2.4. Qua đó cho thấy các thách thức tiềm ẩn trong các số liệu trong các mẫu của các nhóm, dữ liệu không đồng đều trong các nhiệm vụ phân loại. Trong các chương tiếp theo của luận án giải quyết các nhiệm vụ nghiên cứu từ tập dữ liệu này cũng như các đặc tính chi tiết của dữ liệu

2.6. Kết chương

Trong chương này luận án đã trình bày các khái niệm liên quan trong nghiên cứu cũng như các vấn đề tổng quan về đến các bài toán nghiên cứu. Trong các chương sau luận án trình bày chi tiết về các công trình đã nghiên cứu và cách thức giải quyết các thách thức bài toán đã nêu.

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

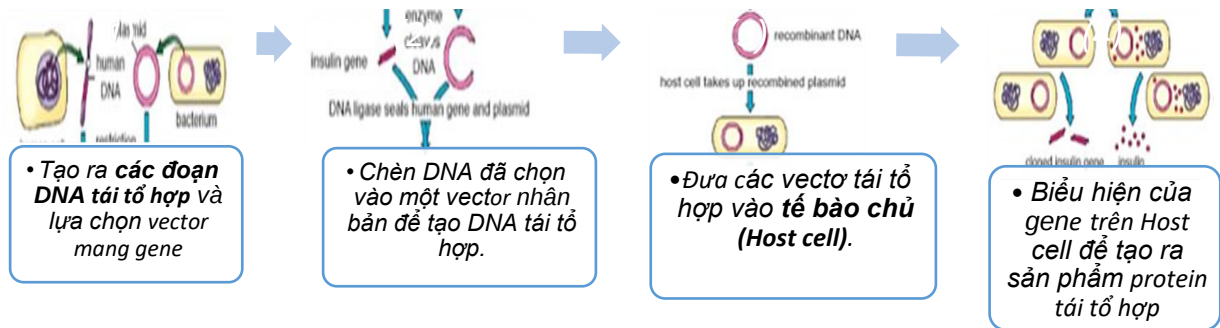
Chương này trình bày cách xây dựng mô hình học máy hiệu quả cho dữ liệu Genomic. Triển khai hai phương pháp học máy không giám sát và giám sát cho dữ liệu trình tự cho hai bài toán bài toán tìm gene có khả năng biểu hiện protein cao và dự đoán gene phù hợp với hệ thống biểu hiện. Nội dung trình bày trong chương này cũng đã được công bố trong các công trình [CT1] [CT2][CT3].

3.1. Giới thiệu.

Trong những năm gần đây, kỹ thuật tái tổ hợp DNA đã có sự tiến bộ đáng kể và đã được sử dụng rộng rãi trong sản xuất thuốc, vaccine, giống cây trồng và vật nuôi [3, 4]. Kỹ thuật tái tổ hợp DNA, cho phép các nhà nghiên cứu can thiệp vào cấu trúc gen của một sinh vật và tạo ra các biến thể gen mới. Trong lĩnh vực sản xuất thuốc, vaccine các nhà nghiên cứu sử dụng kỹ thuật này để chèn các gen mã hóa các phân đoạn protein của vi khuẩn, virus hoặc vi sinh vật khác vào các tế bào vi khuẩn hoặc tế bào thực vật. Qua đó, các tế bào này có khả năng sản xuất các protein đặc hiệu cung cấp khả năng phòng ngừa hoặc điều trị bệnh. Cụ thể kỹ thuật này phối hợp các từ hai nguồn DNA của sinh vật khác nhau để tạo ra các protein đích hay gene đích. Quá trình tạo ra các sản phẩm của DNA tái tổ hợp có thể tóm tắt theo các bước như sau: (i) Tạo ra các đoạn DNA tái tổ hợp và lựa chọn vector mang gene; (ii) Chèn DNA đã chọn vào một vector nhân bản để tạo DNA tái tổ hợp. (iii). Đưa các vector tái tổ hợp vào tế bào chủ. (iv) Biểu hiện của gene trên tế bào vật chủ (Host cell) để tạo ra sản phẩm protein tái tổ hợp.

Hình 3.1 minh họa giai đoạn quan trọng của quá trình sản xuất protein tái tổ hợp. Trong giai đoạn đầu của việc tiến hành kỹ thuật tái tổ hợp cần phải chọn lựa một gene mục tiêu sau cho

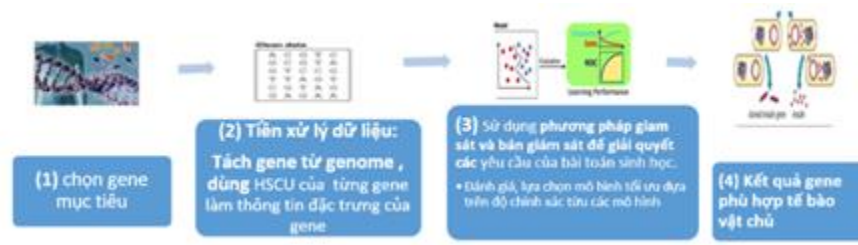
thích hợp với các thao tác ở các giai đoạn tiếp theo để tạo ra sản phẩm cuối cùng đạt chất lượng nhất [3]. Tổng quan quá trình tạo ra các sản phẩm protein của DNA tái tổ hợp sau:



Hình 3.1: Các giai đoạn tiến hành thực kỹ thuật tái tổ hợp

Hệ gene sinh vật dùng trong kỹ thuật tái tổ hợp nhỏ nhất cũng chứa bốn ngàn gene. Dữ liệu của từng gene mục tiêu có độ dài khác biệt dao động từ 20 nucleotides đến 6 ngàn nucleotides. Dữ liệu gene dùng trong kỹ thuật này số chiều rất lớn (hàng chục nghìn chiều) và số mẫu cũng khá lớn (vài ngàn mẫu). Các giải pháp thường được lựa chọn để tìm hệ thống biểu hiện phù hợp với gene mục tiêu là thống kê chọn lọc từ thực nghiệm [8]. Áp dụng giải pháp học máy cho việc tìm gene mục tiêu phù hợp là một đề xuất đầu tiên của luận án. Với mục tiêu tìm gene biểu hiện cao phù hợp với hệ thống biểu hiện gia tăng sản lượng cũng như chất lượng của protein tái tổ hợp, luận án thực hiện nhiệm vụ này với công việc: i) tìm nhóm gene cho biểu hiện; ii) kế tiếp là tìm Host cell phù hợp với gene mục tiêu.

Quá trình tạo ra DNA tái tổ hợp cần phải tìm được **nhóm gene biểu hiện protein cao (Highly Expressed Gene, HEG), cũng như môi trường tế bào vật chủ phù hợp** nhất cho việc tạo ra sản phẩm tái tổ hợp để nâng cao chất lượng sản phẩm protein tái tổ hợp [5] [3]. Theo Pere Puigbò và cộng sự để tìm được HEG thì họ dựa vào chỉ số thích nghi codon CAI (Codon Adaptation Index) và dùng thống kê để xác định HEG [8]. Hệ gene sinh vật dùng trong kỹ thuật tái tổ hợp nhỏ nhất cũng chứa 4 ngàn gene. Dữ liệu của từng gene mục tiêu có độ dài khác biệt dao động từ 20 nucleotides đến 6 ngàn nucleotides. Đây là hai công việc khó, tốn nhiều thời gian và chi phí để thực hiện. Luận án đã đề xuất hai mô hình học máy để giải quyết vấn đề tìm nhóm **gene cho biểu hiện protein cao và lựa chọn môi trường tế bào vật chủ** phù hợp với gene mục tiêu để nâng cao hiệu quả sản xuất được phẩm bằng công nghệ tái tổ hợp. Phương pháp học máy tổng quát để giải quyết nhiệm vụ này được trình bày như hình 3.2.



Hình 3.2: Quy trình tổng quát của bài toán tìm gene biểu hiện cao, và tìm host cell phù hợp với gene mục tiêu.

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

3.2.1. Bài toán tìm HEG

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. Các HEG 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. Tìm được tập gene HEG cho hệ thống biểu hiện bằng phương pháp học máy là một giải pháp

3.2.2. Phương pháp giải quyết

Phương pháp phân tích cụm thường được thực hiện bằng các kỹ thuật học không giám sát. Nguyên tắc cơ bản của phân cụm là tìm cách tạo ra các nhóm dữ liệu có tính đồng nhất cao trong cùng một nhóm và có độ khác nhau giữa các nhóm là lớn nhất. Do đó, luận án sử dụng phương pháp phân hoạch được sử dụng để phân cụm nhằm lựa chọn các nhóm gene có xu hướng sử dụng codon tương tự nhau trong cùng một nhóm và có thể chọn được các HEG có khuynh hướng gần nhất đến trung tâm của mỗi nhóm. 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.

Giai đoạn (1): Chỉ số RSCU (Relative Synonymous Codon Usage) là một chỉ số được sử dụng để phân tích sự sử dụng của các codon đồng nghĩa trong mã di truyền. Công thức được dùng để biểu diễn gene trong hệ gene bằng kỹ thuật RSCU [8].

$$r_{ac}(g) = \frac{O_{ac}}{\frac{1}{K_a} \sum_{c \in C_a} O_{ac}} = \frac{O_{ac}}{O_a} \quad (3.1)$$

Trong đó:

r_{ac} : RSCU của codon c mã hoá cho amino acid a

O_{ac} : tần suất xuất hiện của codon c trong trình tự gene

C_a : tập hợp các codon cùng mã hoá cho amino acid a

K_a : số lượng các loại codon c mã hóa cho amino acid a

Đối với các codon kết thúc (UAA, UGA, UGA) và các codon không có codon đồng nghĩa (AUG – mã hóa cho Methionine, UGG – mã hóa cho Tryptophan). Do đó, gene được biểu diễn bằng tập các RSCU là một vector RSCU 59 chiều có dạng như sau $r(g) = \{r_1, r_2, r_3 \dots r_{59}\}^T$

Bộ dữ liệu hệ gene được biểu diễn lại có hình dạng như sau:

ID_gen	GCU	GCC	GCA	GCG	CGU	CGC	CGA	CGG	AGA	AGG	AAU	AAC	GAU	GAC	UGU	UGC	CAA	CAG	GAA	GAG	GGU	GGC	GGA	GGG	CAU	CAC	UUA	UUG	CUU	CUC	CUA	CUG	AUU	AUC	AUA	AAA	
NC_000964.3_cds_YP_054573.1_864		0	1	1	0	0	3	1	2	2	0	0	0	0	0	0	3	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	6
NC_000964.3_cds_NP_389485.1_1634	0	3	6	2	0	2	1	1	5	1	0	2	0	5	0	0	6	4	15	10	0	9	2	1	0	5	0	0	0	0	0	0	0	0	0	0	12
NC_000964.3_cds_NP_390868.1_3045	0	3	2	2	0	3	1	3	2	0	0	6	0	6	0	0	3	2	11	6	0	7	2	2	0	2	0	0	0	0	0	0	0	0	0	0	15

Hình 3.3: . Minh họa cho biểu diễn gene với RSCU

Giai đoạn (2): phân cụm PAM, CLARA

Phương pháp phân hoạch được dùng để gom cụm các nhóm gene có cùng xu hướng sử dụng codon vào cùng một nhóm và có thể chọn HEG có khuynh hướng gần nhân của nhóm nhất. Cách xác định giá trị khoảng cách giữa các gene được tính như sau:

Giả sử hệ gene của tế bào có n gene được biểu diễn bằng p chiều là 59, và tập dữ liệu này là cấu trúc có dạng bảng quan hệ, hay ma trận $n \times p$ với n gene; p chiều là tên các codon. Tập dữ liệu gene được biểu diễn lại bằng ma trận dữ liệu có dạng:

$$\begin{bmatrix} r_{11} & r_{12} & \dots & r_{159} \\ \vdots & \vdots & \ddots & \vdots \\ r_{i1} & r_{i2} & \dots & r_{i59} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \dots & r_{n59} \end{bmatrix}$$

Sử dụng khoảng cách Euclid

$$d(i, j) = \sqrt{(|r_{i1} - r_{j1}|^2 + |r_{i2} - r_{j2}|^2 + \dots + |r_{ip} - r_{jp}|^2)} \quad (3.2)$$

Xây dựng ma trận khoảng cách giữa các gene trong không gian dữ liệu gồm n gene theo thuộc tính RSCU ta dùng ma trận phân biệt với $d(i, j)$ là khoảng cách giữa gene i và gene j ; khoảng cách này được tính theo công thức (3.2)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Áp dụng các thuật toán phân hoạch lên bộ dữ liệu vừa xây dựng, một trong các thuật toán sử dụng là PAM, CLARA.

Thuật toán PAM (Partitioning Around Medoids) là một thuật toán phân cụm dữ liệu không giám sát, được sử dụng để tìm các cụm dữ liệu dựa trên các điểm trung tâm (medoids) [35]. Thuật toán PAM bắt đầu bằng cách chọn ngẫu nhiên k điểm dữ liệu ban đầu làm các medoids của k cụm. Sau đó, thuật toán lặp lại hai bước cho đến khi không có sự thay đổi nào trong cụm: i) Gán mỗi điểm dữ liệu vào cụm có medoid gần nhất; ii) Tính toán tổng khoảng cách của các điểm dữ liệu đến medoid của cụm và chọn một điểm dữ liệu khác như là medoid mới cho cụm đó, sao cho tổng khoảng cách này là nhỏ nhất; iii) Các bước trên được lặp lại cho đến khi không có sự thay đổi nào trong cụm. Kết quả của thuật toán là các cụm dữ liệu được xác định bởi các medoids. Luận án đã áp dụng Thuật toán 2.1 vào bài toán phân cụm gene với dữ liệu như minh họa hình 3.3.

Thuật toán CLARA (Clustering LARge Applications) là một thuật toán phân cụm dữ liệu không giám sát được sử dụng để xử lý các tập dữ liệu lớn [37]. Thuật toán Clara bắt đầu bằng cách: i) chọn ngẫu nhiên một tập hợp con của dữ liệu ban đầu và áp dụng thuật toán PAM (Partitioning Around Medoids) trên tập hợp con này để tạo ra k cụm. Sau đó, thuật toán lặp lại hai bước cho đến khi không có sự thay đổi nào trong cụm: i) Áp dụng PAM trên một tập hợp con mới của dữ liệu, được lấy ra ngẫu nhiên từ dữ liệu ban đầu; ii) Đối với mỗi cụm, tính toán tổng khoảng cách của các điểm dữ liệu đến medoid của cụm và lưu trữ cụm có tổng khoảng cách nhỏ nhất. Thuật toán 3.2 được áp dụng được mô tả dưới dạng mã giả tương ứng

Thuật toán 3.2: CLARA

Gọi S là kích thước mẫu được trích từ tập gene $G = \{g_1, g_2, \dots, g_n\}$,

k : số cụm,

n : số gene.

S : tập hợp các gene được đưa vào cụm

Input:

Tập hợp các chuỗi gene $G = \{g_1, g_2, \dots, g_n\}$,

Số cụm k

Output: Tập hợp gene đã được phân vào k cụm.

Begin:

(1). **For** $i = 1$ to S **do**

(2). Lấy một mẫu có S_j gene ngẫu nhiên từ tập dữ liệu G . Áp dụng thuật toán PAM cho mẫu dữ liệu này nhằm để tìm các gene medoid đại diện cho các cụm.

(3). Đối với mỗi đối tượng trong tập dữ liệu ban đầu, xác định gene medoid tương tự nhất trong số k đối tượng medoid.

(4). Tính độ phi tương tự 2 trung bình cho phân hoạch các đối tượng thu được ở bước trước, Nếu giá trị này bé hơn giá trị tối thiểu hiện thời thì sử dụng giá trị này thay cho giá trị tối thiểu ở trạng thái trước, như vậy, tập k đối tượng medoid xác định ở bước này là tốt nhất cho đến thời điểm này.

(5). Quay về bước 1.

End

Kỹ thuật phân cụm K-means [44] dựa trên việc tính toán khoảng cách Euclidean giữa các điểm dữ liệu và trung tâm cụm. Kỹ thuật K-means không yêu cầu tính toán ma trận khoảng cách hoàn toàn mà chỉ tính toán khoảng cách giữa các điểm dữ liệu. Thuật toán này nhạy cảm với các giá trị ngoại lệ. Bài toán *Tìm gene biểu hiện cao* là tìm chính xác tên gene hay điểm medoid (điểm dữ liệu đại diện cho cluster). Trong bài toán *Tìm gene biểu hiện cao*, có tồn tại các điểm dữ liệu có các giá trị ngoại lệ có thể ảnh hưởng đáng kể đến kết quả phân tích. Kỹ thuật phân cụm CLARA/PAM đáp ứng được điều này bằng cách tính độ tương đồng giữa các gene bằng so sánh giá trị mô tả gene bằng ma trận khoảng cách. Do đó các kỹ thuật phân cụm bằng PAM, CLARA có khả năng xử lý dữ liệu ngoại lệ tốt hơn. Bên cạnh đó CLARA/PAM hiệu quả hơn khi xử lý dữ liệu lớn đặc biệt là dữ liệu genomic.

Giai đoạn (3): Đánh giá phân cụm

Phương pháp phân hoạch để phân cụm thường sử dụng chỉ số Silhouette để đo chất lượng của các cụm được tạo ra [45]. Chỉ số Silhouette là một phương pháp đánh giá cụm dựa trên các giá trị khoảng cách giữa các điểm trong cùng một cụm và giữa các điểm của các cụm khác nhau. Chỉ số Silhouette được sử dụng để đo độ tách biệt giữa các cụm và độ tương đồng giữa các điểm trong cùng một cụm. Giá trị của chỉ số Silhouette dao động từ -1 đến 1, với giá trị càng gần 1 thì cụm càng tốt, giá trị càng gần -1 thì cụm càng không tốt và giá trị 0 cho thấy các điểm không rõ ràng thuộc về cụm nào.

Công thức (1.2) tính chỉ số Silhouette cho một điểm dữ liệu i như sau [45]:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (1.2)$$

Trong đó:

$a(i)$: là khoảng cách trung bình giữa gene i và tất cả các gene trong cùng một cluster.

$b(i)$: là khoảng cách trung bình giữa gene i và tất cả các gene trong cluster lân cận gần nhất

Chỉ số Silhouette có thể được tính toán cho từng chỉ số RSCU của gene trong tập dữ liệu và sau đó có thể lấy trung bình để có một chỉ số tổng thể cho toàn bộ phân cụm. Giá trị càng gần 1 cho thấy phân cụm tốt, trong khi giá trị gần 0 hoặc âm cho thấy phân cụm không tốt.

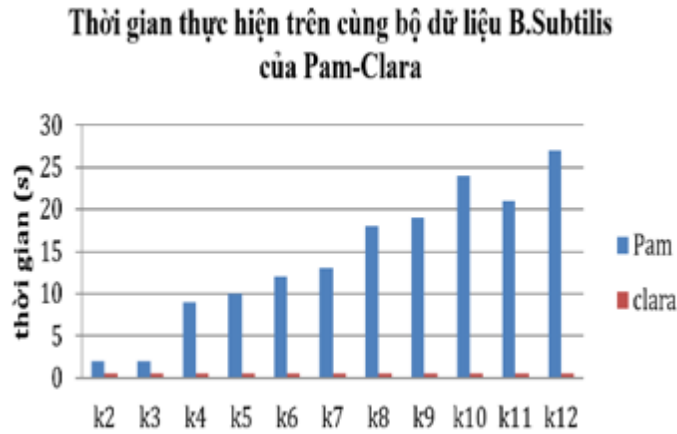
Chỉ số Silhouette có thể được sử dụng để so sánh hiệu quả của các phương pháp clustering khác nhau hoặc để tìm số lượng clusters tối ưu trong thuật toán phân cụm.

3.2.3. Kết quả thực nghiệm

Thuật toán học máy không giám sát được sử dụng là: thuật toán PAM, và CLARA để phân cụm dữ liệu để tìm HEG trên bộ gene *B.Subtilis*. Đánh giá hiệu năng của hai thuật toán PAM-CLARA

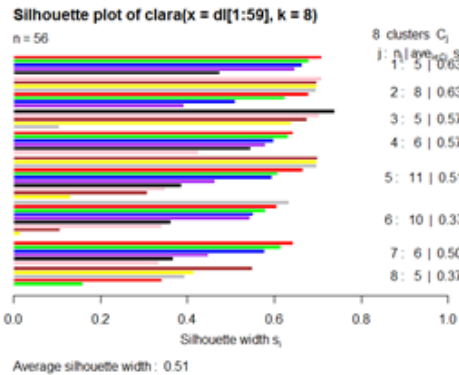
❖ Thời gian thực thi PAM, CLARA trên bộ dữ liệu *B.subtilis*

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. Thống kê thời gian thực thi của PAM và CLARA trên cùng tập gene *B.Subtilis*



Hình 3.4: . Thống kê thời gian thực thi PAM, CLARA

- ❖ **Chỉ số Silhouette (Silhouette Index):** được dùng để đánh giá chất lượng phân cụm. Trong thực nghiệm, thực toán CLARA cho kết quả tốt nhất như hình 3.5. Thông tin mã số gene HEG được chọn như mô tả như hình 3.6.



Hình 3.5: Hình ảnh phân cụm bằng CLARA

objective function: 83.84559
clustering vector: int [1:3543] 1 2 3 4 5 6 7 3 8 1 9 10 11 9 11 9 9 12 .
cluster sizes: 86 53 60 72 56 66 75 33 65 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 71 27 59 60
99 70 29 54 48 44 11 30 26 41 56 36 13 40 40 10 50 21 26 20 80 46 57 50 5 35 55
52 76 30 5
best sample:
[1] 17 45 51 79 82 84 89 138 152 191 219 242 259 1277
[15] 281 285 287 300 306 329 358 368 397 450 461 473 529 540
[29] 563 606 633 638 647 649 676 685 712 744 751 769 781 786
[43] 797 803 837 866 883 897 914 955 1004 1020 1028 1033 1065 1071
[57] 1081 1089 1139 1153 1159 1172 1174 1195 1196 1205 1206 1210 1235 1244
[71] 1246 1279 1282 1301 1306 1320 1321 1322 1328 1355 1379 1417 1451 1469
[85] 1482 1485 1490 1502 1528 1550 1561 1621 1640 1651 1669 1675 1685 1687
[99] 1739 1774 1786 1790 1820 1829 1832 1838 1841 1937 1944 1959 1975 1983
[113] 2063 2087 2098 2137 2146 2147 2148 2150 2170 2253 2260 2300 2326 2347
[127] 2352 2360 2366 2438 2438 2454 2455 2462 2510 2532 2553 2567 2582 2599
[141] 2621 2626 2645 2669 2692 2710 2717 2762 2765 2784 2785 2836 2838 2898
[155] 2906 2916 2944 2953 2996 2997 3006 3020 3025 3088 3102 3109 3161 3164
[169] 3213 3214 3254 3285 3347 3357 3376 3378 3381 3400 3425 3458 3461 3490
[183] 3492 3497 3521 3524 3528 3536 3540 3543

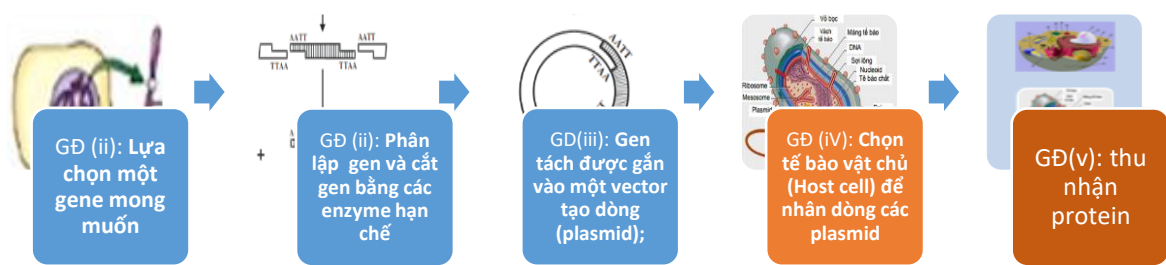
Hình 3.6: 5% HEG được chọn

3.3. Bài toán tìm hệ thống biểu hiện phù hợp với gene mục tiêu.

3.3.1. Phát biểu bài toán

Quy trình sản xuất protein tái tổ hợp thông thường bao gồm 5 giai đoạn được mô tả ở hình 3.6. Host cell được chọn ở giai đoạn 4 với mục đích là nhân dòng gene mục tiêu thành nhiều bản nhất có thể và dễ cho thao tác tách dòng. Đối với giai đoạn (v) thì tế bào vật chủ được chọn phải đáp ứng đủ các yếu tố hỗ trợ cho việc dịch mã mà quan trọng nhất là là cần phải đáp ứng

đủ xu hướng sử dụng của gene mục tiêu. Vấn đề then chốt trong kỹ thuật biểu hiện chính là khả năng mã (codon usage). Trước khi bắt tay vào thí nghiệm, cần so sánh khả năng mã của vật chủ với thành phần codon trong gene mã hóa protein mục tiêu. Kiểm tra lại trình tự protein quan tâm và sau đó thì lựa chọn hệ thống biểu hiện phù hợp cho nó. Rất nhiều protein từ sinh vật eukaryote chỉ có hoạt tính khi ở dạng đường hóa (glycosylated). Thông thường, các hệ thống biểu hiện ở vi khuẩn không có khả năng đường hóa protein. Nếu protein quan tâm đòi hỏi phải được chế biến sau dịch mã mới có hoạt tính thì chắc chắn hệ biểu hiện ở vi khuẩn không thể là giải pháp. Do đó việc tìm Host Cell phù hợp với gene mục tiêu là bài toán phân lớp quan trọng được luận án chọn áp dụng 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 như hình 3.7.



Hình 3.7: Các giai đoạn trong quá trình sản xuất protein tái tổ hợp;

(i) Lựa chọn một gene mong muốn, (ii) phân lập gene và cắt gene bằng các enzyme hạn chế, (iii) Gene tách được gắn vào một vector tạo dòng (plasmid); (iv) Chọn tế bào vật chủ để nhân dòng các plasmid; (v) Sử dụng gene đã nhân dòng đưa vào một ‘hệ thống biểu hiện’ thích hợp để thu protein mong muốn.

3.3.2. Phương pháp giải quyết

Phương pháp phân lớp gene được thực hiện bằng các kỹ thuật học giám sát trong học máy. Nguyên tắc cơ bản của phân lớp là xác định lớp hoặc nhãn của một mẫu dữ liệu mới dựa trên thông tin từ các mẫu dữ liệu huấn luyện đã biết lớp hoặc nhãn của chúng. Trong phân lớp gene, mỗi gene được coi như một đặc trưng của mẫu dữ liệu, và mục tiêu là phân loại các mẫu dữ liệu thành các lớp khác nhau dựa trên các giá trị của các đặc trưng này. Các kỹ thuật học giám sát được sử dụng để xây dựng các mô hình phân lớp từ các mẫu dữ liệu huấn luyện, và sau đó áp dụng mô hình này để phân lớp các mẫu dữ liệu mới.

Việc xây dựng một mô hình phân lớp chính xác và tin cậy đòi hỏi các bước tiền xử lý dữ liệu và lựa chọn đặc trưng thích hợp, cùng với việc đánh giá và điều chỉnh các tham số của mô hình để đạt được hiệu suất tốt nhất. Có thể khái quát phương pháp này như sau:

(1). Ở giai đoạn tiền xử lý dữ liệu: mã hóa gene dùng kỹ thuật RSCU cho toàn bộ gene trong hệ gene .

(2). Áp dụng các thuật toán phân lớp Random Forest, XGBoost để xây dựng bộ dữ liệu đã mã hóa ở bước (1). Đánh giá chọn lựa mô hình

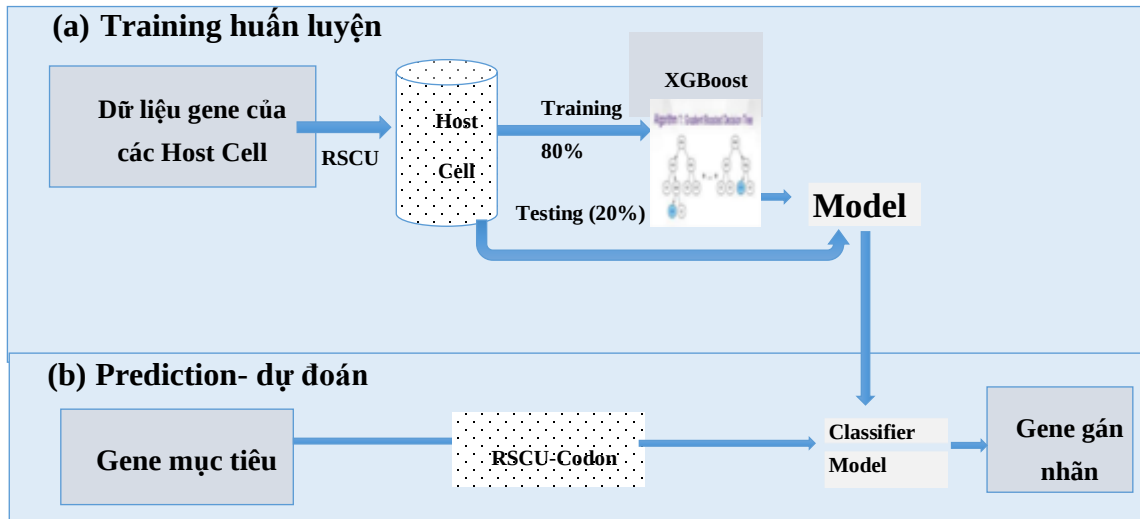
Cụ thể các công việc triển khai cho đề xuất này như sau:

Giai đoạn (1): Sử dụng phương pháp biểu diễn tập gene của các tế bào vật chủ bằng chỉ số các codon đồng nghĩa RSCU. Các ở công thức (3.1). Do đó, gene được biểu diễn bằng tập các RSCU là một vector RSCU 59 chiều có dạng như sau $r(g)=\{r_1, r_2, r_3...r_{59}\}^T$, Bộ dữ liệu hệ gene của từng Host cell được biểu diễn lại có dạng như hình 3.8.

ID_gen	GCU	GCC	GCA	GCG	CGU	CGC	CGA	CGG	AGA	AGG	AAU	AAC	GAU	GAC	UGU	UGC	CAA	CAG	GAA	GAG	GGU	GGC	GGA	GGG	CAU	CAC	UUA	UUG	CUU	CUC	CUA	CUG	AUU	AUC	AUA	AAA	class	
NC_000964.3_cds_YP_054573.1_864	0	1	1	0	0	3	1	2	2	0	0	0	0	0	0	3	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	6	1
NC_000964.3_cds_NP_389485.1_1634	0	3	6	2	0	2	1	1	5	1	0	2	0	5	0	6	4	15	10	0	9	2	1	0	5	0	0	0	0	0	0	0	0	0	0	12	2	
NC_000964.3_cds_NP_390868.1_3045	0	3	2	2	0	3	1	3	2	0	0	6	0	6	0	3	2	11	6	0	7	2	2	0	2	0	0	0	0	0	0	0	0	0	0	15	2	

Hình 3.8: Minh họa cho biểu diễn tập gene của các HostCell với RSCU

Giai đoạn (2): Các thuật toán học máy phân lớp truyền thống như SVM, K-NN, Naive Bayes, thường dùng cho các bài toán phân loại. Tuy nhiên với đặc điểm đặc thù liệu y sinh thường có đặc điểm chiều cao, và hạn chế xóa dữ liệu trống thì hiệu năng từ các thuật toán này không đáp ứng tốt. Luận án đã đề sử dụng hai thuật toán điển hình của phương pháp học máy này là Random forest (RF) và XGBoost để xây dựng mô hình phân loại. Quy trình đề xuất được mô tả ở hình 3.9.



Hình 3.9: Sơ đồ xây dựng mô hình dự đoán gene tương quan với tế bào vật chủ.

Như vậy với tập dữ liệu đầu vào là các trình tự n của các hostcell R , có m thuộc tính thì $D = \{(x_i, y_i)\} (|D| = n, x_i \in R, y_i \in R)$ tập gene biểu diễn thành tập dữ liệu đầu vào cho bài toán phân loại có dạng:

$$\mathcal{L}(X_i, Y_i)_{i=1}^n \quad (3.3)$$

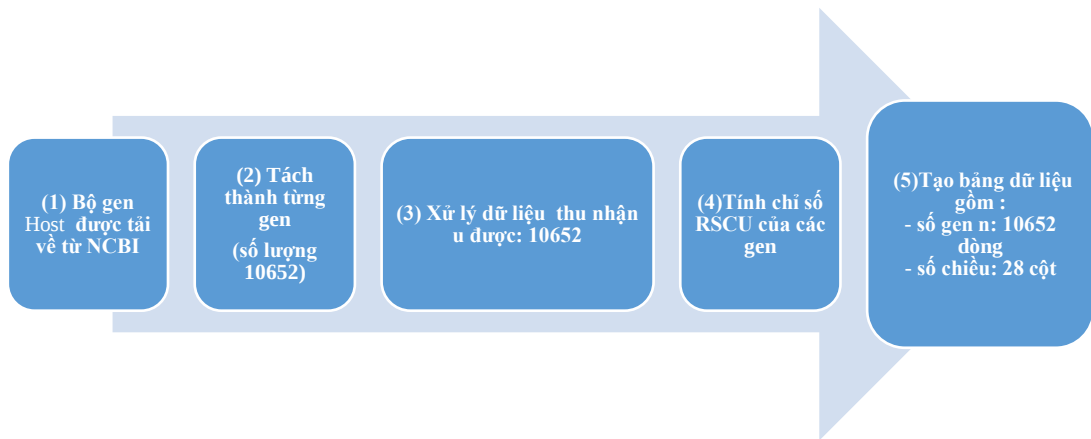
Trong đó X_i : là predictor features có $X_i = \{x_1, x_2, \dots, x_p\}$ là các giá trị RSCU của từng loại codon, Y_i là response features, $Y_i = \{y_1, y_2, \dots, y_k\}$ là tập biến đích gán nhãn cho bộ dữ liệu.

Xây dựng cây quyết định là giải pháp cốt lõi của bài toán phân lớp gene tương quan với tế bào vật chủ.

Luận án sử dụng phần mềm Rstudio cùng gói Random Forest, XGBoost có chứa các thuật toán về các mô hình hồi quy. Quy trình thực nghiệm triển khai như hình 3.9, Sử dụng từ ba hệ gene thường được dùng làm Host cell là: E.coli, B.subtilis, Latococcus Lactis MG1363, sau đó dùng mô hình này để dự đoán gene tương quan phù hợp với các host cell này.

3.3.3. Xử lý dữ liệu thực nghiệm

Bên cạnh E. coli, vi khuẩn chủng Bacillus subtilis (gọi là B.subtilis) và Latococcus Lactis cũng là một trong số các hệ thống biểu hiện đang được quan tâm nghiên cứu và sử dụng hiện nay trong lĩnh vực sản xuất protein tái tổ hợp [3]. Quá trình chuẩn bị dữ liệu được mô tả như hình 3.10



Hình 3.10: Quy trình xử lý dữ liệu tổng quát [46]

Dữ liệu sau khi tải về từ ngân hàng gene quốc tế NCBI, tách thành từng gene, Số lượng mẫu thu được là 10652 gene. Tuy nhiên, chỉ hơn một nửa số loại codon của từng vật chủ có hỗ trợ cho quá trình dịch mã, tức là có một số codon không tham gia vào quá trình dịch mã. Do đó, số chiều của tập dữ liệu huấn luyện nhỏ hơn 59, tập dữ liệu huấn luyện được chia thành ba loại vật chủ E.coli, B.subtilus, Latococcus Lactis MG1363. Tập dữ liệu huấn luyện này có 29 chiều (29 loại codon được sử dụng trong quá trình dịch mã) và có thể được biểu diễn dưới dạng ma trận kích thước 10652 x 29, trong đó mỗi hàng tương ứng với một gene và mỗi cột tương ứng với một loại codon.

3.3.4. Thực nghiệm mô hình dự đoán gene tương quan

Áp dụng gói phần mềm XGBoot tiến hành thực nghiệm trên môi trường R Statistic với giải pháp XGBoot dùng cho dữ liệu thưa như đã mô tả ở phần II, và gói phần mềm Random Forest dùng để kiểm chứng kết quả cho mô hình đề xuất. Luận án đã thực nghiệm từ tập gene 10652 và 28 tiêu chí RSCU Codon với 80% tập gene này dùng làm dữ liệu huấn luyện- Training, và 20% dùng để làm dữ liệu kiểm thử- testing mô hình. Quy trình xây dựng mô hình hồi quy được mô tả trong câu hỏi bao gồm các bước sau:

1. Sử dụng kỹ thuật kiểm tra chéo 5-folds để đánh giá mô hình. Trước khi thực hiện kiểm tra chéo, ta phải chọn giá trị n_round (số lượng cây trong ensemble) để sử dụng trong mô hình. Luận án đã chọn giá trị $n_round = 10$ và thực hiện kiểm tra chéo để đánh giá hiệu suất của các mô hình với các giá trị hàm loss khác nhau.

2. Thực nghiệm mô hình với XGB.CV trên tập training để liệt kê các giá trị hàm loss. XGB.CV là một phương thức trong thư viện XGBoost để thực hiện kiểm tra chéo và tìm giá trị hàm loss tối ưu.
3. Chọn giá trị hàm loss thấp nhất từ bước trên và thực hiện điều chỉnh n_round để tìm giá trị nhỏ nhất với hàm loss này.
4. Thực hiện kiểm tra mô hình trên tập dữ liệu testing và xây dựng ma trận hỗn độn (confusion matrix) để đánh giá hiệu suất của mô hình.

Sau khi tìm được mô hình tốt nhất, luận án đã sử dụng nó như một classifier model để gán nhãn cho tập gene mục tiêu gồm khoảng 21 gene.

Luận án đã tiến hành thực nghiệm quy trình trên 10 lần với số n-round dao động từ 10 đến 400 để tìm tập các giá trị hàm Loss nhỏ nhất và tìm mô hình tốt nhất cho mô hình dự đoán, Các kết quả về hiệu năng của mô hình lần lượt trình bày như các hình 3.11 và hình 3.12

Confusion Matrix

Prediction	Reference		
	1	2	3
1	3281	12	9
2	4	3282	23
3	22	36	1852

Hình 3.11: Confusion Matrix của mô hình đề xuất [46]

Độ chính xác mô hình

Accuracy :	0.9876
95% CI :	(0.985, 0.9898)
No Information Rate :	0.3908
P-Value [Acc > NIR] :	< 2.2e-16
Kappa :	0.9808
McNemar's Test P-Value :	0.006375

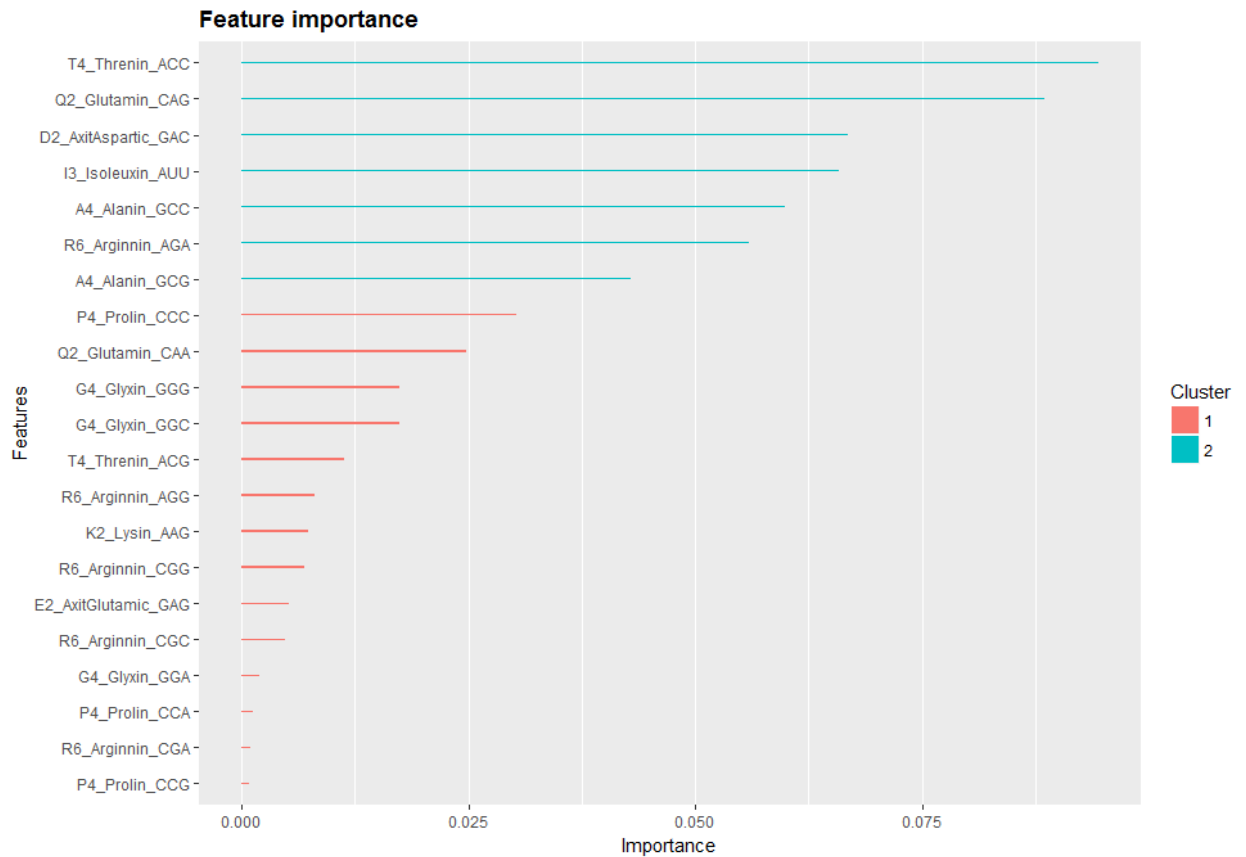
Hình 3.12: Hiệu năng về độ chính xác của mô hình đề xuất [46]

Các chỉ số quan trọng về các lớp dự báo của mô hình đề xuất như độ nhạy-Sensitivity, độ đặc hiệu – Specificity. Các thông tin này đều cho thấy mức độ đáng tin cậy của mô hình đề xuất. Các thông tin chi tiết về các chỉ số này được trình bày trong bảng 3.1

Bảng 3.1: Thống kê hiệu năng của mô hình đề xuất [47]

	Class1	Class2	Class3
Độ nhạy (Sensitivity)	0.9921	0.9856	0.9830
Độ đặc hiệu ((Specificity)	0.9960	0.9948	0.9913

Bên cạnh đó luận án còn đánh giá độ quan trọng của các đặc trưng trong dữ liệu, chi tiết về các được trình bày trong hình 3.13.

**Hình 3.13:** Kết quả đặc trưng quan trọng dùng trong mô hình dự đoán [46]

❖ Thực nghiệm với thuật toán Random Forest

Luận án đã tiến hành thực thi Random Forest với cùng tập dữ liệu huấn luyện và kiểm tra như đã thực nghiệm với XGBoost, cũng đã áp dụng với cùng số lượng cây là 100. Sử dụng mô hình dự đoán thu được từ việc thực thi này (rfhost) áp dụng dự đoán nhãn cho tập 21 gene mục tiêu đã dùng với mô hình đã đề xuất bằng thuật toán XGBoost đã nêu trên. Kết quả dự đoán nhận được từ hai thuật toán giống nhau hoàn toàn.

❖ Đánh giá hiệu năng của hai thuật toán trên từng loại dữ liệu Host cell

Luận án đã thực nghiệm và đánh giá thời gian thực thi của hai thuật toán Random Forest và XGBoost trên tập dữ liệu huấn luyện gồm 8520 gene. Bộ dữ liệu huấn luyện được tổng hợp từ ba bộ tế bào vật chủ là: Host Ecoli (Class 1); B.Subtilus (Class 2); Latococcus (Class 3). Kết quả hai mô hình thu được từ hai thuật toán trên được đánh giá về độ chính xác và thời gian thực thi. Thông tin chi tiết về hiệu năng được trình bày trong bảng 3.2.

Bảng 3.2: Độ chính xác của hai thuật toán tương ứng với các mô hình dự báo [46]

	Host cell	Random Forest	XGBoost
Độ chính xác	Ecoli (Class 1)	0,97	0,99
	B.Subtilus (Class 2)	0,94	0,99
	Latococcus (Class 3)	0,88	0,97
Thời gian thực thi (s)		30,22	22,37

Như vậy các giải pháp đề xuất sử dụng loại hệ gene của hệ thống biểu hiện là vi khuẩn như E.coli, B.subtilus, Latococcus Lactis MG1363. Đây là các loại hệ thống biểu hiện có giá thành rẻ và thường được dùng trong các nghiên cứu về protein tái tổ hợp. Kết quả thực nghiệm cho thấy kết quả dự đoán của các mô hình đều trùng khớp nhau tuy nhiên thời gian thực thi của thuật toán XGBoost nhanh hơn. Nên XGBoost được chọn làm mô hình đề xuất chính cho bài toán dự đoán host cell phù hợp với gene mục tiêu. Hơn nữa với thời gian thực thi thuật toán dao động từ 22 đến 30 giây, đây là khoảng thời gian có thể chấp nhận được cho các ứng dụng thời gian thực, do đó việc áp dụng mô hình đề xuất này trong thực tế là rất khả thi.

3.4. KẾT LUẬN

Để tạo ra sản phẩm protein tái tổ hợp chất lượng cao, cần tìm ra gene mục tiêu tốt và môi trường tế bào vật chủ phù hợp nhất để đạt được năng suất sản xuất cao. Dữ liệu của từng gene mục tiêu có độ dài khác nhau và hệ gene của sinh vật dùng trong kỹ thuật tái tổ hợp có thể chứa hàng ngàn gene. Đây là một thách thức khi cần chọn lọc gene cho sản xuất protein tái tổ hợp, Vấn đề này có thể được giải quyết bằng sử dụng các phương pháp học máy để tìm ra nhóm gene có biểu hiện protein cao, cũng như lựa chọn môi trường tế bào vật chủ phù hợp với gene mục tiêu. Luận án đã đề xuất hai phương pháp học máy như học máy không giám sát (unsupervised learning) và học giám sát đã được áp dụng để phân tích dữ liệu gene và tìm ra nhóm gene có biểu hiện protein cao. Các thuật toán phân cụm (clustering) bằng PAM và Clara và phân loại (classification) RF, XGBoost cũng được sử dụng để xác định môi trường tế bào

vật chủ phù hợp cho từng gene mục tiêu. Các kết quả từ mô hình đề xuất có độ chính xác tổng thể trên 98%, với thời gian tối thiểu là 30 giây. Tóm lại, các phương pháp học máy như đã đề cập đã chứng minh là hiệu quả đối với dữ liệu genomic trong nhiệm vụ tìm nhóm gene có biểu hiện protein cao và lựa chọn môi trường tế bào vật chủ phù hợp. Dữ liệu genomic của dữ liệu y sinh trong hai nhiệm vụ này đều dựa trên kỹ thuật mã hóa codon đồng nghĩa dùng chỉ số RSCU. Trong nhiệm vụ tiếp theo của luận án đề xuất phương pháp học máy phân loại gene dựa trên trình tự gene dạng thô. 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 để xây dựng mô hình học máy hiệu quả hơn cho dữ liệu genomic.

CHƯƠNG 4. MÔ HÌNH ĐỊNH DANH LOÀI SINH VẬT

Biết rõ nguồn gốc loài sinh vật có thể giúp nhà nghiên cứu xác định mối quan hệ di truyền giữa các loài, xác định các mô hình tiến hóa và phân loại chúng. Công việc này giúp bảo tồn loài, phát hiện loài mới, hay trong y sinh thì hỗ trợ chẩn đoán và điều trị. Trong chương này, luận án đề xuất mô hình học tập kết hợp (Ensemble learning) cho nhiệm vụ định danh loài sinh vật dựa trên dữ liệu genomic. Nội dung trình bày trong chương này được công bố trong các công trình [CT4], [CT7].

4.1. Giới thiệu

4.1.1. Định danh loài sinh vật

Định danh loài sinh vật là quá trình gán cho một loài sinh vật một cái tên duy nhất và chuẩn xác dựa trên các đặc điểm di truyền, hình thái, hệ thống học và đặc điểm sinh học khác của loài đó. Quá trình này được gọi là phân loại sinh học (biological classification) và được thực hiện bởi các nhà sinh học và các chuyên gia định danh loài. Định danh loài sinh vật rất quan trọng trong nghiên cứu và bảo tồn đa dạng sinh học. Nó giúp xác định và phân loại các loài sinh vật khác nhau, đo lường sự đa dạng sinh học của các khu vực và giúp quản lý và bảo vệ các loài hoang dã.

Trong việc định danh loài sinh học, các công cụ tính toán như BLAST (Basic Local Alignment Search Tool) và MEGA (Molecular Evolutionary Genetics Analysis) được sử dụng để so sánh các chuỗi gene và phát hiện các mối quan hệ giữa các loài sinh vật. Các phương pháp tính toán đặc hiệu cho sinh học phân tử như UPGMA (Unweighted Pair Group Method with Arithmetic Mean), Maximum Parsimony, Neighbor-Joining và Maximum Likelihood được sử dụng để xây dựng cây phân loại sinh học. Các cây phân loại này giúp hiểu rõ hơn về mối quan hệ giữa các loài và cung cấp thông tin quan trọng cho các nghiên cứu về đa dạng sinh học và tiến hóa. Đã có rất nhiều giải pháp hỗ trợ cho việc định danh loài sinh học tại các ngân hàng gene lớn trên thế giới như BLAST, MEGA. Các giải pháp này thường dùng dữ liệu về gene hiện có kết hợp dùng các kỹ thuật tính toán đặc hiệu cho sinh học phân tử như UPGMA, Maximum Parsimony, Neighbor-Joining, và Maximum Likelihood để xây dựng cây phân loài sinh học. Giải pháp học máy cũng đã có nhiều giải pháp hiệu quả ứng dụng cho việc định danh loài. Một số nghiên cứu nổi bật có thể kể đến như Prabina Kumar Meher và các cộng sự để xây dựng một mô hình phân loại thực vật [47] dựa vào kỹ thuật Random Forest để định danh loài

sinh vật. Đối với mô hình phân loại động vật thì Mulyati cũng áp dụng mô hình học máy thuật toán SVM [48]. Kết quả nghiên cứu cho thấy rằng phương pháp học máy có thể tạo ra các mô hình phân loại loài sinh vật với độ chính xác cao. Trong luận án đã kết hợp nhiều phương pháp học máy để xây dựng mô hình dự đoán tên loài sinh vật dựa trên dữ liệu sinh học phân tử.

4.1.2. Giới thiệu định danh loài nấm

Nấm là một nhóm sinh vật thuộc về giới nấm (Fungi) trong hệ thống phân loại sinh vật. Chúng là một nhóm đa dạng, bao gồm hàng triệu loài khác nhau trên khắp thế giới. Nấm không thuộc về giới thực vật do chúng không có khả năng tổng hợp năng lượng từ ánh sáng mặt trời thông qua quá trình quang hợp như cây cỏ và cây cối. Nấm có vai trò rất quan trọng trong tự nhiên như: Chúng đóng vai trò trong quá trình phân hủy và tái sinh chất hữu cơ trong môi trường; Nấm cũng có giá trị kinh tế cao trong lĩnh vực thực phẩm, y học và công nghiệp; Một số loại nấm được sử dụng làm thực phẩm, như nấm mèo, nấm rơm, nấm hương, và nấm bào ngư; Nấm còn được sử dụng trong y học để điều trị một số bệnh và làm nguyên liệu cho việc sản xuất thuốc; trong công nghiệp, nấm cũng được sử dụng để sản xuất men, enzyme và các hợp chất sinh học khác. Xác định được đúng tên loài nấm có ý nghĩa quan trọng trong việc phân loại và xác định đặc tính ứng dụng quan trọng trong tự nhiên. Định danh loài sinh vật bằng kỹ thuật sinh học phân tử là quá trình xác định và phân loại các loài sinh vật dựa trên thông tin di truyền được mã hóa trong phân tử DNA hoặc RNA của chúng.

4.1.3. Định danh loài nấm mới bằng phương pháp học máy

Phương pháp định danh loài truyền thống thường dựa trên các khóa phân loại còn gọi là đặc điểm hình thái của từng loài, tuy nhiên với các loài mà mẫu vật thu thập không nguyên vẹn, hay bảo quản không đúng cách thì rất khó phân định chính xác. Phương pháp sinh học phân tử là một kỹ thuật được sử dụng để xác định loài dựa trên phân tích các đặc trưng di truyền của chúng. Phương pháp này dựa trên việc phân tích các đoạn DNA hoặc RNA của các mẫu nghiên cứu, từ đó đưa ra các kết quả liên quan đến sự khác biệt giữa các loài. Một trong những phương pháp phổ biến nhất để định danh loài bằng phương pháp sinh học phân tử là sử dụng kỹ thuật mã vạch DNA (DNA barcoding). Trong phương pháp này, một đoạn nhỏ của DNA được chọn để mã hóa các loài khác nhau, và sau đó được sử dụng để so sánh với các cơ sở dữ liệu đã biết để xác định loài. Ngoài các phương pháp trên, còn có các phương pháp khác như sử dụng kỹ thuật học máy để phân loại và xác định các loài. Trong phạm vi nghiên cứu của

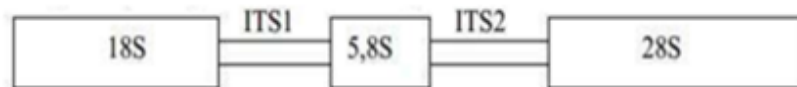
luận án đề xuất một giải pháp mới dùng học máy để định danh loài nấm mối. Phương pháp này đều cung cấp các kết quả đáng tin cậy để xác định và phân loại các loài nấm, nhằm hỗ trợ nghiên cứu và quản lý bảo tồn các loài nấm quý hiếm.

Nấm mối thuộc chi *Termitomyces* là một loại nấm có giá trị dinh dưỡng cao và hương vị ngon, chỉ mọc trong môi trường tự nhiên. Ngoài giá trị dinh dưỡng cao, loại nấm này còn được biết đến với tính chất dược phẩm ở nhiều quốc gia trên thế giới. *Termitomyces clypeatus* cũng hỗ trợ trong điều trị bệnh thủy đậu [49]. Các hợp chất quý giá của các loài nấm đặc biệt này được thu được thông qua quá trình nuôi cấy sinh khối. *Termitomyces Albuminosus* để kiểm tra hiệu quả của nó trong giảm đau và chống viêm trong khi *Termitomyces striatus* được sử dụng để chiết xuất các hợp chất khác. *Termitomyces heimii* và *Termitomyces microcarpus* được sử dụng trong điều trị sốt, cảm lạnh và nhiễm nấm và trong việc thúc đẩy liệu pháp ung thư [50]. Có khoảng 30 loài nấm mối trên toàn thế giới, và 10 loài ở Việt Nam, trong đó *Termitomyces clypeatus* và *Termitomyces microcarpus* phổ biến ở Bình Dương. Mặc dù rất hiệu quả kinh tế, sản lượng tự nhiên của những loài nấm này đang giảm đáng kể, và chúng vẫn chưa được nuôi cấy một cách bền vững, vì chúng chỉ mọc theo mùa.

Xác định chính xác tên của một loài nấm mối là một công việc quan trọng trong nghiên cứu sinh học. Thông thường các chuyên gia sử dụng khóa phân loại truyền thống để xác định các loài nấm mối dựa trên hình thái của chúng. Và các đặc điểm này gồm: núp nấm, thịt, màng và thân, có thể có vòng [51]. Tuy nhiên, cấu trúc của nấm thay đổi từ loài này sang loài khác, đặc biệt là khi xảy ra đột biến. Hơn nữa, việc xác định các mẫu không có đặc điểm hình thái có thể khó khăn [52]. Một phương pháp để xác định các loài mới của các sinh vật thường được sử dụng để xác định các loài nấm ăn được và dược phẩm dựa trên các kỹ thuật phân tử. Trong phương pháp này, các kỹ thuật phân tử như mã vạch DNA đã được sử dụng thành công trong những năm gần đây để xác định các loài [53]. Các phương pháp phân tử này dựa trên phân tích các đánh dấu di truyền và đã chứng tỏ rất hiệu quả trong việc xác định các loài, đặc biệt là khi kết hợp với các phương pháp hình thái truyền thống. Nhìn chung, kết hợp các kỹ thuật phân tử vào quá trình xác định các loài nấm mối có thể cung cấp việc xác định chính xác và hiệu quả hơn, đặc biệt là trong những trường hợp phương pháp hình thái truyền thống không đủ.

Một nhóm gene thường được sử dụng trong việc xác định phân tử là nhóm mã hóa rRNA. Nhóm này rất hiệu quả trong việc tìm sự tương đồng và khác biệt khi so sánh các sinh vật khác

nhau do tính bảo toàn tương đối cao của hầu hết các phân tử rRNA (Van de Peer 1996). Đối với nấm, vùng rDNA ITS (Internal Transcribed Spacer), bao gồm hai chuỗi ITS1 và ITS2, kề hai bên chuỗi 5,8S, được chấp nhận rộng rãi là vùng phân tử để xác định loài bởi hầu hết các nhà nghiên cứu nấm [54], như được thể hiện trong Hình 4.1. Vùng ITS cũng được sử dụng để dự đoán các loài nấm bằng cách sử dụng các phương pháp học máy. Phương pháp này bao gồm sử dụng dữ liệu chuỗi ITS để huấn luyện một mô hình học máy, sau đó có thể được sử dụng để phân loại và xác định các loài nấm khác nhau một cách chính xác. Bằng cách kết hợp các kỹ thuật phân tử như học máy với các phương pháp xác định hình thái truyền thống, các nhà nghiên cứu có thể đạt được việc xác định các loài nấm chính xác và hiệu quả hơn, góp phần trong cả nghiên cứu và bảo tồn.



Hình 4.1: Cấu trúc của gene ITS [54]

Với một mẫu nấm nhỏ qua quá trình phân tích trình tự có thể được sử dụng để xác định được loài. Các đoạn DNA này được gọi là DNA Barcode để định danh loài nấm. Các đoạn gene là trình tự ITS cho nấm hay thực vật thì sử dụng gene rRNA 18S, 5S và 16S. Các gene này thường được dùng để đánh giá mối quan hệ tiến hoá giữa các sinh vật. So với chỉ thị hình thái và chỉ thị hoá học, chỉ thị DNA cho độ chính xác cao hơn mà không lệ thuộc vào bất cứ yếu tố khách quan nào [11]. Quy trình để có được đoạn gene ITS của nấm thường được tiến hành theo 5 bước [11]:

1. Lấy mẫu DNA từ mẫu nấm thuần đã được phân lập.
2. Sử dụng DNA để thực hiện phản ứng PCR để khuếch đại đoạn DNA đặc hiệu trong khoảng 900-1200 bp trên vùng rDNA ITS1, 5.8S, ITS2, 28S của nấm bằng thiết bị PCR Thermal Cycler (Bio-Rad).
3. Sản phẩm PCR được điện di trên gel agarose 2%, sau đó gel được chụp hình lại bằng thiết bị Gel Doc (Bio –Rad).
4. Sản phẩm PCR được tinh sạch bằng bộ kit QIA quick PCR Purification Kit của QIAGEN, sau đó điện di trên bộ điện di Agilent 2100 Bioanalyzer (Agilent Technologies).

PCR sản phẩm đã tinh sạch trước khi giải trình tự trên hệ thống máy ABI 3130XL (Applied Biosystems).

5. Kết quả được phân tích bằng phần mềm Sequencing analysis 5.3 và so sánh với các trình tự đã biết trên cơ sở dữ liệu NCBI. Các trình tự về nhận dạng loài thường được sử dụng trình tự ITS là vùng DNA nằm giữa các gene hay còn gọi là DNA chỉ thị loài. Các trình tự ITS này thường được các nhà nghiên cứu công bố lên các ngân hàng gene quốc tế phục vụ cho việc nghiên cứu chung. Các dữ liệu chuỗi ITS cho nấm có thể được truy cập từ hai bộ sưu tập lớn, BOLD (Barcode of Life Data) và Trung tâm Thông tin Sinh học Quốc gia (NCBI). Cả hai ngân hàng gene này đều chứa một bộ sưu tập lớn các chuỗi ITS cho tất cả các loài nấm.

Phân loại các loài nấm dựa trên học máy sử dụng chuỗi ITS đã được đề xuất bởi một số nhà nghiên cứu như: Schloss PD 2009; Delgado-Serrano L 2016; Deshpande V 2016 và Prabina Kumar Meher 2019. Một danh sách toàn diện về các kỹ thuật và dữ liệu được sử dụng trong các nghiên cứu phân loại nấm được cung cấp trong Bảng 4.1.

Bảng 4.1: Các nghiên cứu về định danh nấm bằng phương pháp Học máy

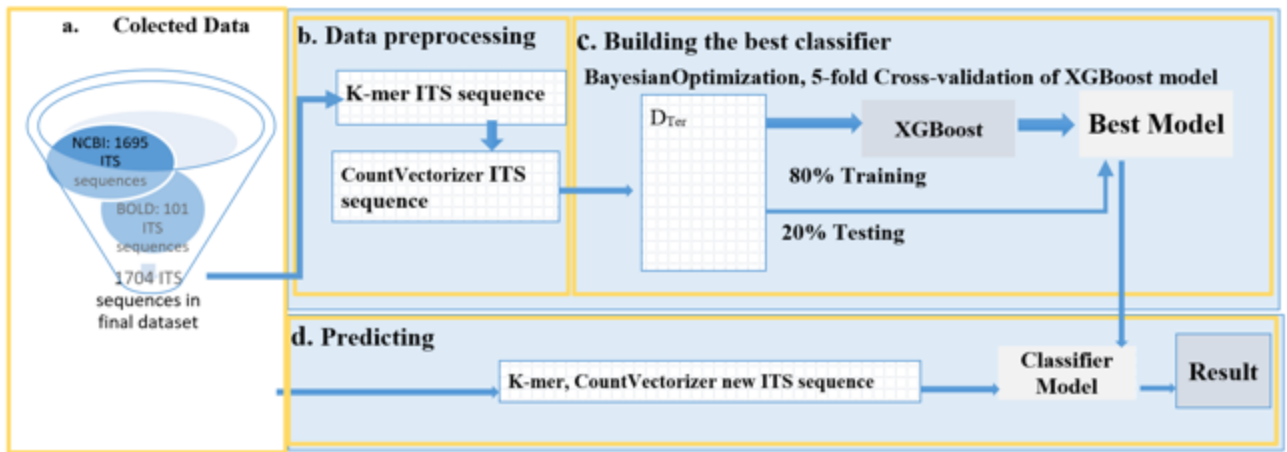
Tham khảo	Phần mềm	Nguồn dữ liệu\ số mẫu	Kỹ thuật tích thuộc tính, thuật toán học máy	Độ chính xác tốt nhất của mô hình
[55]	MOTHUR	The SILVA Database Project, Bremen, March 2009	K-mer (k=5), The k-nearest neighbor (kNN) algorithm, and PGMA (unweighted-pair group method using average linkages) algorithms	0.86
[56]	Mycofar	NCBI GeneBank	K-mer (k=5), Naïve Bayes classifier	0.87
[57]	RDP	The Warcup dataset (18878 chuỗi thuộc về 8551 loài)	K-mer (k=8) Bayesian regression.	0.87
[58]	SINTAX	RDP Warcup ITS (18878 chuỗi thuộc về 8551 loài)	K-mer (k=8) Naive Bayesian Classifier	0.87
[59]	funbarRF	BOLD systems	K-mer (k=4)	0.89

			Random Forest.	
[60]	CNN_FunBar	UNITE + INSDC (4504529 chuỗi thuộc về 44167 loài)	K-mer (k=6), CNN	0.86

Các nghiên cứu đã thành công trong việc sử dụng các phương pháp học máy giám sát như phân loại Naive Bayes, kNN và các mô hình hồi quy Bayesian để phân loại các loài nấm. Tuy nhiên, chỉ có Delgado-Serrano [56] xác định được các loài nấm ở mức chi, trong khi các nghiên cứu khác chỉ xác định được tên loài. Đa số các nghiên cứu sử dụng dữ liệu chuỗi gene ITS nấm từ NCBI để làm tập dữ liệu chính cho việc huấn luyện tập học, tuy nhiên nguồn dữ liệu này không đầy đủ. Ví dụ, chuỗi ITS của chi nấm mốc *Termitomyces euripus* trong NCBI gene bank chỉ có một chuỗi, trong khi có sáu nhãn thuộc chi nấm này trong BOLD. Ngoài ra, độ dài của chuỗi ITS có sự khác biệt lớn, dao động từ 200 đến 2000 base, và số lượng chuỗi giữa các chi nấm khác nhau cũng có sự khác biệt lớn, dao động từ 1 mẫu đến 500 mẫu. Do những hạn chế này với dữ liệu ITS cho nấm mốc, các thuật toán học máy cổ điển gặp khó khăn trong việc phân loại chính xác các nhãn của các loài nấm mốc. Nghiên cứu của luận án tập trung vào xác định các nhãn của các chi nấm mốc bằng cách sử dụng dữ liệu chuỗi ITS được thu thập từ hai ngân hàng gene NCBI và BOLD. Luận án kết hợp kỹ thuật K-mer và xử lý ngôn ngữ tự nhiên (NLP) để trích xuất đặc trưng và thử nghiệm các phương pháp phân loại hiện đại như XGBoost (Extreme Gradient Boosting), Random Forest và CatBoost để xây dựng một bộ phân loại loài nấm mốc tự động.

4.2. Mô hình định danh loài nấm bằng kỹ thuật học kết hợp

Luận án đã phát triển một quy trình tự động gồm bốn giai đoạn để dự đoán tên loài của một loài nấm mốc mới. Giai đoạn đầu tiên liên quan đến việc thu thập dữ liệu nấm mốc từ các kho lưu trữ chuỗi gene ITS. Ở giai đoạn thứ hai, các đặc trưng chuỗi được trích xuất và mã hóa. Giai đoạn thứ ba liên quan đến việc xây dựng một bộ phân loại bằng cách xây dựng và điều chỉnh các thông số để tìm bộ phân loại tối ưu. Cuối cùng, ở giai đoạn thứ tư, bộ phân loại được sử dụng để dự đoán các mẫu nấm mốc mới. Hình 4.2 cung cấp một mô tả chi tiết về quy trình này. Quy trình xây dựng mô hình dự đoán tên loài nấm mốc được tiến hành theo 7 bước được liệt kê như sau:



Hình 4.2: Quy trình dự đoán tên loài nấm mốc.

- Thu thập dữ liệu (collected data):** Nghiên cứu đã thu thập tổng cộng 1796 chuỗi ITS của các loài nấm từ Gene Bank NCBI và BOLD. Sau khi lọc các chuỗi loài nấm mốc có ít hơn 10 chuỗi, số lượng chuỗi ITS cuối cùng là 1704.
- Tiền xử lý dữ liệu (Data preprocessing):** Các chuỗi ITS được chia thành các chuỗi nhỏ hơn, theo các quy tắc được mô tả trong Hình 2, bằng cách sử dụng K-mer với kích thước 7. Chuỗi ITS dài nhất có độ dài 2470 cơ sở, tương ứng với độ dài vector là 14425 khi được mã hóa.
- Xây dựng mô hình phân lớp với tham số đã được tối ưu hóa (Building the best classifier):** Quá trình huấn luyện sử dụng tỷ lệ chia 80:20 và tối ưu hóa siêu tham số cho mô hình huấn luyện. Mô hình được tối ưu hóa bằng cách sử dụng kỹ thuật Cross-Validation với $k = 5$, và Bayesian Optimization được thực hiện để điều chỉnh các thông số sau: 'max_depth': (5,10), 'gamma': (0,1), 'learning_rate': (0,1), 'n_estimators': (100,400). Mô hình có độ chính xác cao nhất được chọn để phân loại.
- Dự đoán (predicting):** Các mẫu nấm được thu thập tại tỉnh Bình Dương, Việt Nam, và tải xuống từ NCBI được sử dụng làm tập kiểm tra. Những mẫu này được chia thành các K-mer với kích thước 7 và sau đó được áp dụng CountVectorizer. Cuối cùng, mô hình tốt nhất từ giai đoạn c được áp dụng để dự đoán loài của các loài nấm mốc mới.

4.3. Các thuật toán hiệu quả cho mô hình định danh loài nấm

Thuật toán đề xuất xây dựng mô hình định danh loài dựa trên Học phối hợp (Ensemble learning). Luận án đã tiến hành khảo sát trên ba thuật toán Random Forrest, CatBoost,

XGBoost. CatBoost là một lựa chọn khả thi cho phân tích dữ liệu chuỗi gene, như đã chỉ ra trong nghiên cứu gần đây bởi Robson P. Bonidia [61]. Trong các thí nghiệm với dữ liệu nấm, CatBoost hoạt động tương đương với XGBoost về độ chính xác dự đoán. Tuy nhiên, thời gian huấn luyện của CatBoost khá lâu hơn so với XGBoost để đạt được mức độ hiệu suất tương đương. Do đó, XGBoost được sử dụng làm thuật toán chính cho mô hình dự đoán.

Hiệu suất của mô hình XGBoost phụ thuộc vào một số tham số quan trọng như '*max_depth*', '*gamma*', '*n_estimators*' và '*learning_rate*'. Những tham số này được gọi là siêu tham số và có thể được điều chỉnh thủ công trong quá trình huấn luyện hoặc tự động. Mô hình được đề xuất sử dụng kỹ thuật Tối ưu Bayesian [62] và cụ thể là tìm kiếm ngẫu nhiên (Random search) để điều chỉnh siêu tham số. Luận án đã áp dụng Tối ưu Bayesian để điều chỉnh bốn tham số chính của bộ phân loại XGBoost: '*max_depth*', '*gamma*', '*n_estimators*', và '*learning_rate*', 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

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

Đầu ra: Best_Model

Begin

- 1: **Khởi tạo:** 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: **Tính** the roc_auc_score, accuracy_score, precision_score, recall_score, and f1_score over iterations
- 6: **Chọn** best_model mô hình tốt nhất ở bước 4
- 7: **Trực quan** the mean value ở bước 4
- 8: **Return** the Best_Model from step 4

End

Để cải thiện hiệu suất dự đoán của mô hình, đã thực hiện xác thực chéo với $k = 5$ để lựa chọn mô hình phân loại tốt nhất, bên cạnh sử dụng Tối ưu Bayesian để điều chỉnh siêu tham

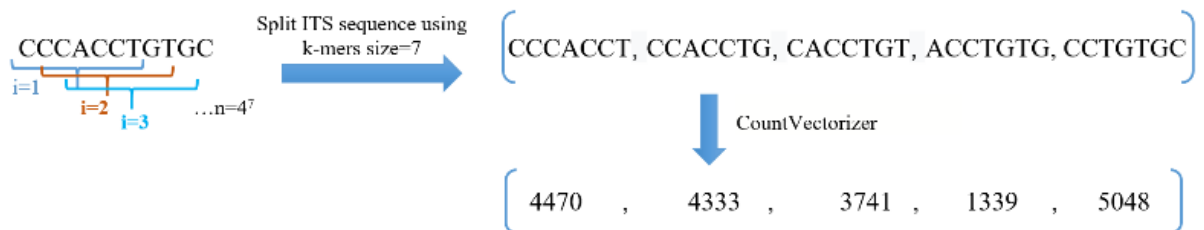
số. Bộ dữ liệu mới từ kết quả giai đoạn 1 được sử dụng, bao gồm n mẫu dữ liệu và m đặc trưng. Sau đó, một tham số tối ưu được sử dụng làm đầu vào cho Thuật toán 4.1 để xây dựng một mô hình phân loại tối ưu.

4.4. Kết quả thực nghiệm

4.4.1. Xây dựng tập dữ liệu cho mô hình định danh loài nấm

Việc chuẩn bị dữ liệu trong học máy là giai đoạn rất quan trọng, luận án đã tiến hành khảo thu thập và khảo sát dữ liệu trình tự nấm mốc từ các ngân hàng gene quốc tế. Dữ liệu sau khi tải về từ ngân hàng gene quốc tế NCBI theo dạng Fasta, tách thành từng gene, trích chọn các đặc trưng chuỗi ITS nấm mốc bằng kỹ thuật K-mer.

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. Với 4 base cơ bản (A, G, T, C) có thể có $4k$ vị trí cho chuỗi trình tự. Hình 4.3 minh họa quá trình trích chọn thông tin bằng kỹ thuật K-mer



Hình 4.3: Minh họa cách tính k-mer cho chuỗi trình tự với $k=7$.

Như đã phân tích như trên cho thấy, 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 để có được 101 chuỗi ITS từ BOLD, với số lượng chuỗi cho mỗi chi dao động từ 1 đến 12. Với ngân hàng dữ liệu từ NCBI, luận án thu được 1740 chuỗi ITS, với số lượng chuỗi cho mỗi loài dao động từ 1 đến 799. Sau khi tổng hợp dữ liệu chuỗi ITS từ hai ngân hàng gene này và loại bỏ các loài nấm mốc có ít hơn 7 chuỗi, có được 1704 chuỗi thuộc về 17 loài nấm mốc. Các nhãn của mỗi loài nấm mốc được trình bày chi tiết trong Bảng 4.2.

Bảng 4.2: Các nhân của mỗi loài nấm mối

TT	Tên loài	Số lượng trình tự	Lable
1	<i>Uncultured Termitomyces</i>	799	16
2	<i>Termitomyces</i> sp.	483	10
3	<i>Termitomyces intermedius</i>	94	8
4	<i>Termitomyces symbiont</i>	60	14
5	<i>Termitomyces microcarpus</i>	34	9
6	<i>Termitomyces clypeatus</i>	33	3
7	<i>Termitomyces cylindricus</i>	30	4
8	<i>Termitomyces striatus</i>	29	13
9	<i>Termitomyces</i> DKA-2007	24	0
10	<i>Termitomyces heimii</i>	24	7
11	<i>Termitomyces bulborhizus</i>	17	2
12	<i>Termitomyces fuliginosus</i>	16	6
13	<i>Termitomyces eurhizus</i>	15	5
14	<i>Termitomyces albuminosus</i>	14	1
15	<i>Termitomyces</i> sp. symbiont of <i>Macrotermes bellicosus</i>	12	11
16	<i>Termitomyces</i> sp. symbiont of <i>Macrotermes subhyalinus</i>	10	12
17	Uncultured Ascomycota	10	15

Tập dữ liệu được sử dụng làm tập kiểm tra là các mẫu nấm được thu thập tại tỉnh Bình Dương, Việt Nam, và tải xuống từ NCBI được mô tả chi tiết ở bảng 4.3.

Bảng 4.3: Tập dữ liệu được sử dụng làm tập kiểm tra là các mẫu nấm được thu thập tại tỉnh Bình Dương

ID_sequences	Tên loài nấm mối ở Bình Dương	Website	Chiều dài của trình tự
KU569480	<i>Termitomyces clypeatus</i>	https://www.ncbi.nlm.nih.gov/search/all/?term=KU569480	980
MF163136-BD5	<i>Termitomyces clypeatus</i>	https://www.ncbi.nlm.nih.gov/search/all/?term=MF163136	720
MF163152.1	<i>Termitomyces clypeatus</i>	https://www.ncbi.nlm.nih.gov/search/all/?term=MF163152	938
MF163445-BD3	<i>Termitomyces</i> sp.	https://www.ncbi.nlm.nih.gov/search/all/?term=MF163445	669
MF163446-BD6	<i>Termitomyces</i> sp.	https://www.ncbi.nlm.nih.gov/search/all/?term=MF163446	1020
MT672480.1	<i>Termitomyces microcarpus</i>	https://www.ncbi.nlm.nih.gov/search/all/?term=MT672480.1	721
MT730584.1	<i>Termitomyces clypeatus</i>	https://www.ncbi.nlm.nih.gov/search/all/?term=MT730584	608
MF163149-BD4	<i>Termitomyces</i> sp.	https://www.ncbi.nlm.nih.gov/search/all/?term=MF163149	812

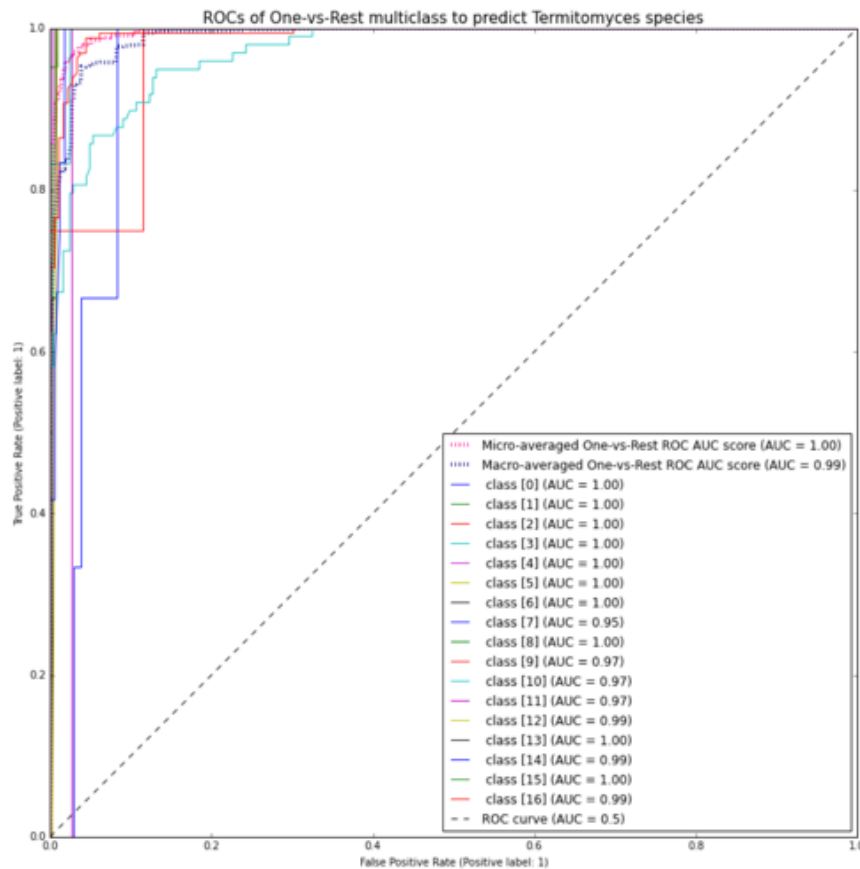
4.4.2. Đánh giá hiệu năng của mô hình đề xuất

Trong quá trình thực nghiệm được sử dụng Python 3.9 và các thư viện Scikit-learn, Biopython, XGBoost, CatBoost và Bayesian optimization để xây dựng một mô hình phân loại nấm theo quy trình được đề xuất như được mô tả trong Hình 4.2. 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 như: AUROC, Accuracy, Precision, Recall, F1 score. Kết quả chi tiết được trình bày trong bảng 4.4.

Bảng 4.4: 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
XGBboost	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 [63]. Ở đây sử dụng lớp cuối (là lớp 16) làm lớp tích cực các lớp còn lại là tiêu cực. Và kết quả AUCROC của mỗi lớp được trình bày trong hình 4.4.



Hình 4.4: Đường cong ROC sử dụng phương pháp OvR macro-average cho mỗi lớp trong phương pháp XGBoost với kích thước K-mer=7.

❖ So sánh hiệu năng với các bộ phân lớp về nấm cũng sử dụng trình tự ITS.

Độ chính xác của các mô hình dự đoán loài nấm trước đây đã dùng phương pháp K-mer và các kỹ thuật toán học máy như: k-NN, Naïve Bayes, Random Forest. Các kết quả lần lượt được trình bày trong bảng 4.5. Kết quả đề xuất 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.

Bảng 4.5: So sánh hiệu năng với các bộ phân lớp về nấm cũng sử dụng trình tự ITS.

Nghiên cứu	Phương pháp	Độ chính xác
[55]	K-mer (k=5), The k-nearest neighbor (kNN) algorithm and PGMA algorithms	0.86
[56]	K-mer (k=5), Naïve Bayes classifier model	0.87
[57]	K-mer (k=8) Bayes regression.	0.87
[59]	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 mới được thu thập tại tỉnh Bình Dương đã được công bố trên NCBI và được mô tả chi tiết ở bảng 4.3. Các trình tự này được dự đoán bằng mô hình phân loại được đề xuất 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 mới MF163445-BD3, MF163446-BD6, MF163149-BD4 nhưng chưa rõ thuộc loài nào. Bộ phân loại này đã định danh được hai trình MF163445-BD3, MF163446-BD6 thuộc loài *Termitomyces striatus* giống kiểu hình của mẫu nấm được thu thập khi giải trình tự. Còn với mẫu nấm MF163149-BD4 thì có kết quả giống như định danh loài trên NCBI. Chi tiết về kết quả định danh loài được trình bày trong Bảng 4.6.

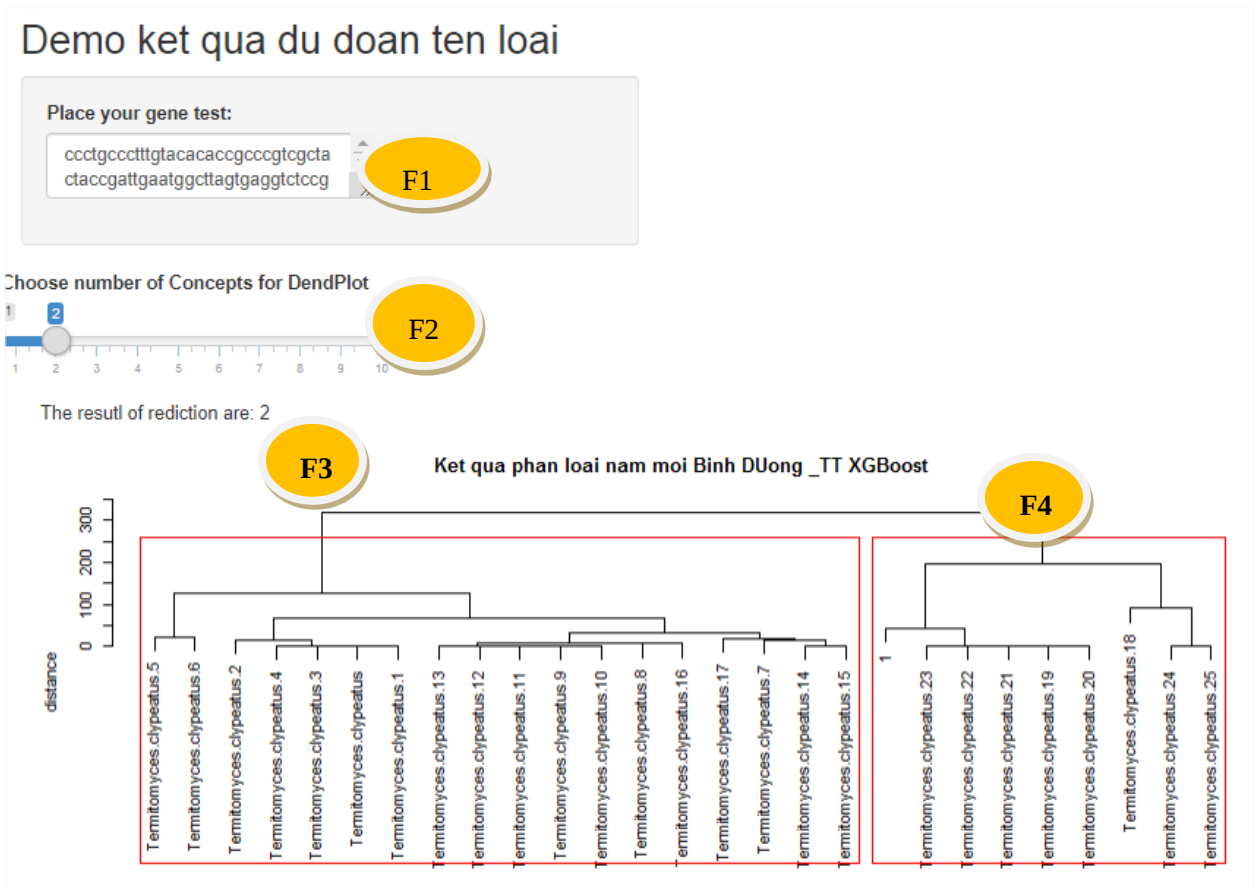
Bảng 4.6: Kết quả so sánh việc định danh loài của trình tự ITS của nấm mới thu thập tại tỉnh Bình Dương, Việt Nam với định danh trên NCBI.

ID_sequences	Bình Dương <i>Termitomyces</i> species in NCBI	Bình Dương <i>Termitomyces</i> species in Our proposal
KU569480	<i>Termitomyces clypeatus</i>	<i>Termitomyces clypeatus</i>
MF163136-BD5	<i>Termitomyces clypeatus</i>	<i>Termitomyces clypeatus</i>
MF163152.1	<i>Termitomyces clypeatus</i>	<i>Termitomyces clypeatus</i>
MF163445-BD3	<i>Termitomyces sp</i>	<i>Termitomyces striatus</i>

MF163446-BD6	<i>Termitomyces sp</i>	<i>Termitomyces striatus</i>
MT672480.1	<i>Termitomyces microcarpus</i>	<i>Termitomyces microcarpus</i>
MT730584.1	<i>Termitomyces clypeatus</i>	<i>Termitomyces clypeatus</i>
MF163149-BD4	<i>Termitomyces sp</i>	<i>Termitomyces sp</i>

4.5. Giao diện mô hình dự đoán tên loài

Để thuận lợi cho việc thực nghiệm, luận án xây dựng một Website để hiển thị kết quả Training dữ liệu được lưu trữ bởi hệ quản trị PostPreSQL Server. Chức năng cơ bản là cho phép nhập trình tự cần định danh và chọn cách xem dữ liệu theo cụm. Hình 4.5 hiển thị minh họa các sử dụng phần mềm dự đoán tên loài nấm mối.



Hình 4.5: Minh họa giao diện định hỗ trợ người dùng sử dụng thuật toán [64]

Các chức năng chính của phần mềm hỗ trợ

F1: Nhập trình tự nấm mối cần phân định loài

F2: Chọn số cụm hiển thị quan sát độ tương đồng

F3: Hiển thị kết quả nhóm loài định danh

F4: Hiện thị cây phân loài của nấm cần định danh với các trình tự gene nấm cùng loài gần với kết quả tập mẫu của CSDL TerDB nhất.

4.6. Kết luận

Với dữ liệu được sử dụng là 17 loài nấm mới được tổng hợp từ hai ngân hàng gene NCBI và BOLD. Dữ liệu được chọn là Nấm mới với số lượng mẫu khá hạn chế phải thu nhận từ nhiều nguồn khác nhau nhưng hiệu năng của mô hình đề xuất có kết quả vượt trội và có ý nghĩa triển khai trong thực tiễn. Cho thấy phương pháp này không chỉ đạt hiệu quả trên dự đoán tên loài nấm mới, mà có thể triển khai trên loài sinh học, y học nói chung. Nên đây cũng là giải pháp có thể áp dụng chung cho các loài khác trong sinh, y học.

Tóm lại, các phương pháp học máy như đã đề cập đã chứng minh là hiệu quả đối với dữ liệu genomic trong nhiệm vụ đề xuất phương pháp học máy phân loại loài sinh vật dựa trên trình tự gene dạng thô. 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. Trong phần tiếp theo của luận án, trình bày phương pháp học máy hiệu quả cho dữ liệu y sinh thuộc loại cận lâm sàng. Đây là loại dữ liệu quan trọng trong lĩnh vực y sinh, cũng với đặc điểm thưa thớt và cũng như số mẫu dữ liệu mất cân bằng nghiêm trọng. Mô hình học máy đề xuất giải quyết các thách thức này của dữ liệu này một cách tự động và mô hình phân loại đề xuất có hiệu năng tốt hơn các nghiên cứu công bố mới nhất trên cùng bộ dữ liệu y sinh này.

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

Bên cạnh dữ liệu genomic, dữ liệu y sinh còn bao gồm dữ liệu cận lâm sàng và lâm sàng. Loại dữ liệu y sinh này được thu thập từ kết quả xét nghiệm sàng lọc khi khám bệnh của các cơ sở y tế. Dữ liệu này thường chứa lỗi, dữ liệu bị thiếu, mất cân bằng nghiêm trọng đối với lớp bệnh hiếm. Cần có công cụ mới hiệu quả hơn hỗ trợ cho việc dự đoán bệnh. Trong chương này luận án trình bày hai đề xuất mô hình dự đoán bệnh dựa trên dữ liệu lâm sàng và mô hình phân biệt bệnh nhân bệnh CoViD-19 hay bệnh Cúm. Nội dung trình bày trong chương này được công bố trong các công trình [CT5], [CT6]

5.1. Giới thiệu

Dữ liệu lâm sàng, cận lâm sàng bao gồm các thông tin như triệu chứng, tiền sử bệnh, kết quả khám lâm sàng. Dữ liệu cận lâm sàng bao gồm các kết quả xét nghiệm máu, nước tiểu, xét nghiệm hình ảnh (như X-quang, siêu âm, CT scan, MRI), xét nghiệm vi sinh, xét nghiệm gene và nhiều phương pháp khác. Thông tin từ dữ liệu lâm sàng và cận lâm sàng giúp cho bác sĩ và chuyên gia y tế để đưa ra quyết định chẩn đoán và điều trị bệnh. Dữ liệu này giúp xác định loại bệnh, đánh giá mức độ nghiêm trọng, phân loại và phân biệt các bệnh có một số triệu chứng tương tự nhau, và theo dõi sự tiến triển của bệnh. Bên cạnh đó dữ liệu này còn được sử dụng để xác định các yếu tố nguy cơ, mô hình dự báo, phân tích kết quả điều trị và đánh giá hiệu quả của các phương pháp điều trị mới. Việc sử dụng một công cụ phân tích dữ liệu, chẩn đoán tự động để phân tích kết quả kiểm tra của bệnh nhân là cần thiết. Hiện nay các phương pháp học máy được sử dụng trong các chẩn đoán y tế mang lại hiệu quả cao cho các cơ sở y tế, bệnh viện. Mục đích của luận án này là phát triển các mô hình dự đoán khả năng mắc bệnh, và phân biệt các loại bệnh có triệu chứng gần giống nhau dựa trên phương pháp học máy. Cụ thể, luận án trình bày hai mô hình: i) Mô hình dự đoán mắc bệnh CoViD-19 dựa trên dữ liệu cận lâm sàng; ii) Mô hình phân biệt hai loại bệnh có triệu chứng gần giống nhau là CoViD-19 và Cúm H1N1.

5.2. Mô hình dự đoán bệnh dựa trên dữ liệu cận lâm sàng

5.2.1. Giới thiệu.

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 nhồi máu cơ tim, suy tim, nhồi máu cơ tim, và rối loạn nhịp tim), bệnh về gan (viêm gan, xơ gan). Với số cỡ mẫu dữ liệu đủ lớn và cùng với sự kiểm tra và xác minh bởi các chuyên gia y tế trên tập mẫu tốt cho kết quả dự đoán tốt nhất. Cũng xuất phát từ mục tiêu xây dựng mô hình hỗ trợ chẩn đoán bệnh dựa trên bộ dữ liệu đã được gán nhãn bởi các chuyên gia y tế. Luận án này là phát triển một mô hình dự đoán khả năng bệnh nhân nhiễm bệnh CoViD-19 dựa trên phương pháp học máy. Cụ thể luận án cũng trình bày về tập dữ liệu được sử dụng để đào tạo mô hình, mô hình tổng quát cho dự đoán, các phương pháp học máy được dùng khảo sát thực nghiệm và cách thức đánh giá mô hình thực nghiệm.

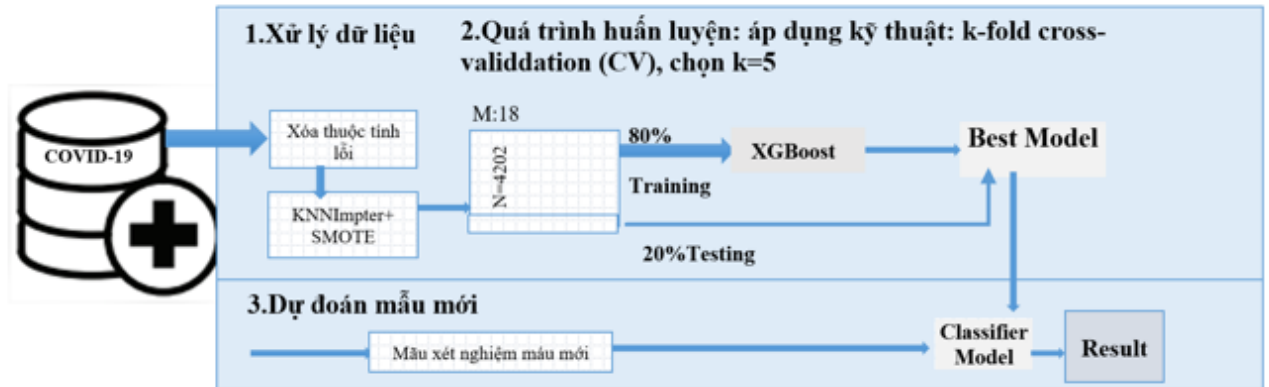
Bệnh do virus SARS_CoV-2 gây ra được báo cáo lần đầu tiên vào tháng 12 năm 2019 tại Trung Quốc. Ban đầu của các đợt bùng phát bệnh, bệnh nhân CoViD-19 thường gặp các triệu chứng giống như cúm như: sốt, ho khan, mệt mỏi, khó thở. Tất cả các triệu chứng lâm sàng và các kết quả xét nghiệm máu từ bệnh nhân mắc bệnh CoViD-19 hoặc tương tự như CoViD-19 được ghi nhận lại, dữ liệu kết quả khẳng định mắc bệnh bằng phương pháp PCR. Các bộ dữ liệu này được xem là bộ dữ liệu cận lâm sàng được các nhà nghiên cứu sử dụng cho nhiều hỗ trợ chẩn đoán bệnh trong bệnh viện và các cơ sở y tế. Nhiều nghiên cứu đã sử dụng với mục tiêu dự đoán mắc bệnh CoViD-19 từ dữ liệu lâm sàng của bệnh nhân đã đạt được kết quả tích cực. Nhiều nghiên cứu tập trung vào phân tích và xác định bệnh CoViD-19 dựa trên kết quả của các thử nghiệm máu; họ thường chọn các tham số từ các chỉ số kết quả xét nghiệm máu của bệnh nhân mắc bệnh làm dữ liệu đầu vào cho các mô hình dự đoán. Độ chính xác của các mô hình này tương đối thấp, dao động từ 81% đến 99,4%. Trong vài nghiên cứu sử dụng bộ dữ liệu lâm sàng cho các mô hình dự đoán của họ, độ chính xác chỉ trong khoảng 84-92%. Việc tổng hợp các nghiên cứu cùng lĩnh vực cũng như hiệu năng của các công bố được trình bày ở bảng 5.1. Các ứng dụng phương pháp học máy trong chẩn đoán bệnh cũng được áp dụng có chính xác cao [65]. Bên cạnh đó cũng có nghiên cứu cho thấy có thể phát hiện bệnh nhân nhiễm CoViD-19 từ xét nghiệm máu định kỳ bằng học máy [66] [65].

Bảng 5.1: Tổng hợp các nghiên cứu cùng mục tiêu dự đoán bệnh nhân CoViD-19 bằng mẫu máu [67]

Tham khảo	Dữ liệu nguồn	Kích cỡ bộ dữ liệu (CoViD-19)	Số lượng thuộc tính	Thuật toán dùng trong các nghiên cứu	Độ chính xác
[65]	IRCCS Ospedale San Raffaele, Italian Scientific Institute for Research, Hospitalization and Healthcare	279 (279)	14 Blood features are available from the clinical data.	Decision Tree (DT); Extremely Randomized Trees (ET); K-nearest neighbors (KNN); Logistic Regression (LR) ; Naive Bayes (NB); Random Forest (RF);	LR = 85%
[31]	Albert Einstein Hospital, Brazil	5564/(553)	19 Blood features	XGBoost.	99,4%
[68]	Albert Einstein Hospital, Brazil	5564/(553)	14 Blood features	RF, LR, GLMNET, ANN.	86%
[69]	Hospitals in Wuhan, China	294 (208)	15 Blood features are available from the clinical data.	RF, SVM , ANN.	84%

5.2.2. Mô hình dự đoán bệnh dựa trên dữ liệu lâm sàng

Với tập dữ liệu đầu vào T với n mẫu bệnh nhân đã được xác định là bị bệnh giống bệnh CoViD-19 (5644 mẫu) và bệnh CoViD-19 (553 mẫu), và $M=18$ thuộc tính thì. Mô hình được áp dụng được mô tả như hình 5.1.



Hình 5.1: Mô hình tổng quan đề xuất cho bộ phân phân lớp bệnh CoViD-19

Trong hình 5.1, mô hình đề xuất chi tiết đã thực hiện qua các bước như sau: (i) Xử lý trước các thuộc tính của số liệu, bao gồm khảo sát các giá trị dữ liệu còn thiếu, phân tích tương quan của các thuộc tính và lọc các dữ liệu nhiễu. (ii) Sử dụng kỹ thuật KNNImputer và SMOTE để xử lý dữ liệu trống và giảm mất cân bằng dữ liệu. (iii) Xây dựng bộ phân loại bệnh dựa trên thuật toán XGBoost với mô hình tối ưu, đã được đánh giá về độ chính xác. Các mẫu thử nghiệm mới có cùng thuộc tính như dữ liệu mẫu máu có tỷ lệ chia cho quá trình huấn luyện là 80:20, trong đó 80% dữ liệu dùng cho việc huấn luyện, 20% dùng cho việc testing. Trong quá trình huấn luyện, luận án đã sử dụng phương pháp k-fold cross-validation (CV). K-Fold CV là một phương pháp phân tích thống kê đã được các nhà nghiên cứu sử dụng rộng rãi để đánh giá hiệu suất của bộ phân loại học máy [70]. Quá trình này được tiếp diễn tự động và lặp lại năm lần nhằm mục tiêu tối ưu hóa các dữ liệu. Cụ thể về giải pháp được trình bày qua các giai đoạn sau:

❖ *Giai đoạn 1: Chuẩn bị dữ liệu cho mô hình học*

Bộ dữ liệu được dùng trong quá trình thử nghiệm được thu bệnh viện Israelita Albert Einstein ở Brazil đây là bộ dữ liệu do Teich tổng hợp từ các bệnh nhân nhập viện tháng 4 đến tháng 5 năm 2020 và được xuất bản công khai trên tạp chí Einstein Journal [31]. Theo các nghiên cứu lâm sàng về bệnh CoViD-19 [71] [31] [72] các thuộc tính về mẫu máu có thể khẳng định khả năng mắc bệnh CoViD-19. Dựa trên công bố [71] [31] [72] [73] luận án đã trích xuất 19 thuộc tính về mẫu máu cho mô hình dự đoán bệnh CoViD-19, thông tin chi tiết về các đặc tính lâm sàng được mô tả ở bảng 5.2.

Bảng 5.2: Thông tin chi tiết các thuộc tính lâm sàng của các mẫu máu từ dữ liệu bệnh viện Israelita Albert Einstein [67]

Tên thuộc tính	Mô tả	Số lượng mẫu
diagnosis	Chẩn đoán	5644
patient_age_quantile	Tuổi	5644
hemoglobin	huyết sắc tố	603
platelets	số lượng tiểu cầu	602
leukocytes	bach cầu Lympho	602
lymphocytes	bach cầu Lympho	602
basophils	bach cầu đa múi ưa kiềm	602

eosinophils	bạch cầu đa múi ưa axit	602
neutrophils	bạch cầu trung tính	513
monocytes	bạch cầu MONO	601
urea	Acid Uric trong máu	397
proteina_c_reativa_mg/dl	protein C	506
creatinine	Chỉ số đào thải creatin phosphat ở cơ	424
potassium	Kali trong máu	371
sodium	Natri huyết thanh	370
alanine_transaminase	Mức độ tổng thương gan alanine_transaminase	225
aspartate_transaminase	Mức độ đông máu aspartate_transaminase	226
international_normalized_ratio_(inr)	international_normalized_ratio	133
albumin	Mức độ thâm thấu albumin	13

❖ **Giai đoạn 2: Xử lý dữ liệu lỗi, các thuộc tính có mức độ liên quan cao trong tập dữ liệu**

Trong dữ liệu y sinh thường chứa nhiều thuộc tính có giá trị lỗi do thuộc tính chứa giá trị null tin trong kết quả xét nghiệm máu của các bệnh nhân [67]. Những thuộc tính có độ lỗi cao cũng ảnh hưởng đến kết quả dự đoán. 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 [67]

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

Đầu vào: set $D_{oar} = X_{oar} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, $x \in R^m$, $y \in \{0, +1\}$ 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* $\leftarrow Dog.drop[data_missing]$

End

Mối tương quan giữa các thuộc tính là một yếu tố quan trọng trong quá trình đánh giá và chẩn đoán bệnh cũng như có thể ảnh hưởng trực tiếp đến kết quả của mô hình dự đoán. Các thuộc tính có mối tương quan cao không có nhiều thông tin hữu ích (hoặc chỉ rất ít), nhưng chúng làm tăng độ phức tạp của thuật toán [74]. Để loại bỏ các thuộc tính này, luận án đã tính toán độ tương quan cho toàn bộ tập dữ liệu định lượng thuộc tính trong tính toán, sau đó loại bỏ các thuộc tính có độ tương quan cao. Phép đo hiệu năng của mối quan hệ giữa hai biến liên tục được sử dụng bởi mô hình Pearson. Giá trị mối tương quan cung cấp thông tin về cả bản chất và mức độ liên quan giữa các mối quan hệ của các đặc trưng. Các giá trị tương quan này luôn nằm trong khoảng từ 0 đến $|\pm 1|$. Ngưỡng được chọn cho việc loại bỏ các thuộc tính tương quan là 0.95 và các tương quan lớn hơn ngưỡng đều bị loại bỏ.

❖ **Giai đoạn 3: Xử lý mất cân bằng dữ liệu của dữ liệu huấn luyện:**

Sử dụng kỹ thuật KNNImputer và SMOTE để xử lý dữ liệu trống và giảm mất cân bằng dữ liệu. Cụ thể đề xuất thực hiện quá trình xử lý mất cân bằng dữ liệu qua hai bước:

Bước 1: Kỹ thuật KNNImputer để điền giá trị bị thiếu: **i)** xác định số lượng láng giềng gần nhất (k) cần tìm cho mỗi điểm dữ liệu bị thiếu; **ii)** k láng giềng gần nhất của điểm dữ liệu bị thiếu bằng các đo khoảng cách Euclidean hoặc các phép đo tương tự; **iii)** Tiếp theo, chúng ta sử dụng giá trị của các láng giềng gần nhất để ước lượng và điền vào giá trị bị thiếu.

Bước 2: sử dụng kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) để tạo ra các mẫu nhân tạo trong các lớp thiểu số để giảm mất cân bằng dữ liệu, cụ thể: **i)** xác định lớp thiểu số và lớp đa số trong tập dữ liệu; **ii)** Chọn một mẫu ngẫu nhiên từ lớp thiểu số và tìm k láng giềng gần nhất của nó trong cùng lớp; **iii)** Tạo ra các mẫu nhân tạo bằng cách lấy một tỷ lệ ngẫu nhiên của độ chênh lệch giữa các đặc trưng của mẫu ban đầu và láng giềng gần nhất, và thêm chúng vào tập dữ liệu.

❖ **Giai đoạn 4: Xây dựng bộ phân loại bằng kỹ thuật Gradient Boosting**

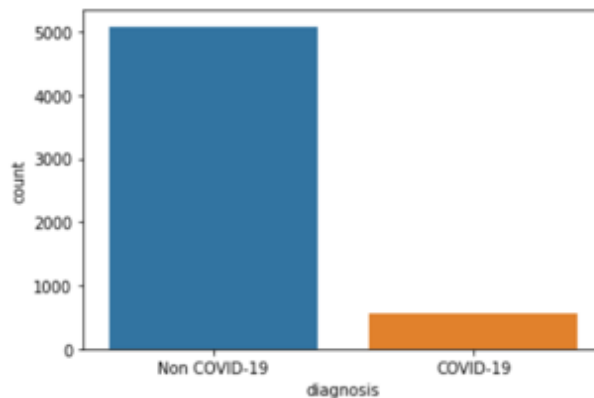
Luận án giải thuật cải tiến vượt trội của Gradient Boosting XGBoost. Khi xây dựng mô hình dự đoán, luận án sử dụng kỹ thuật kiểm tra chéo 5-folds với quy tắc: (i) lần đầu cho ngẫu nhiên số lần lặp n-round=30; (ii) thực nghiệm mô hình với bộ phân lớp (Classifier) trên tập training và liệt kê các giá trị hàm Loss; (iii) chọn giá trị hàm loss thấp nhất; (iv) thực nghiệm điều chỉnh n_round về giá trị nhỏ nhất vừa tìm được, tìm được mô hình hoàn chỉnh.

5.2.3. Kết quả thực nghiệm

Luận án đã ứng dụng mô hình thực nghiệm bằng phần mềm Python và kết hợp sử dụng các gói Sklearn, Missingno, XGBoost. Hiệu suất của phương pháp được kiểm tra bằng các thí nghiệm trên các bộ dữ liệu lâm sàng là mẫu máu được, kết quả thực nghiệm của nghiên cứu cũng được trình bày và được đánh giá so sánh với các nghiên cứu liên quan.

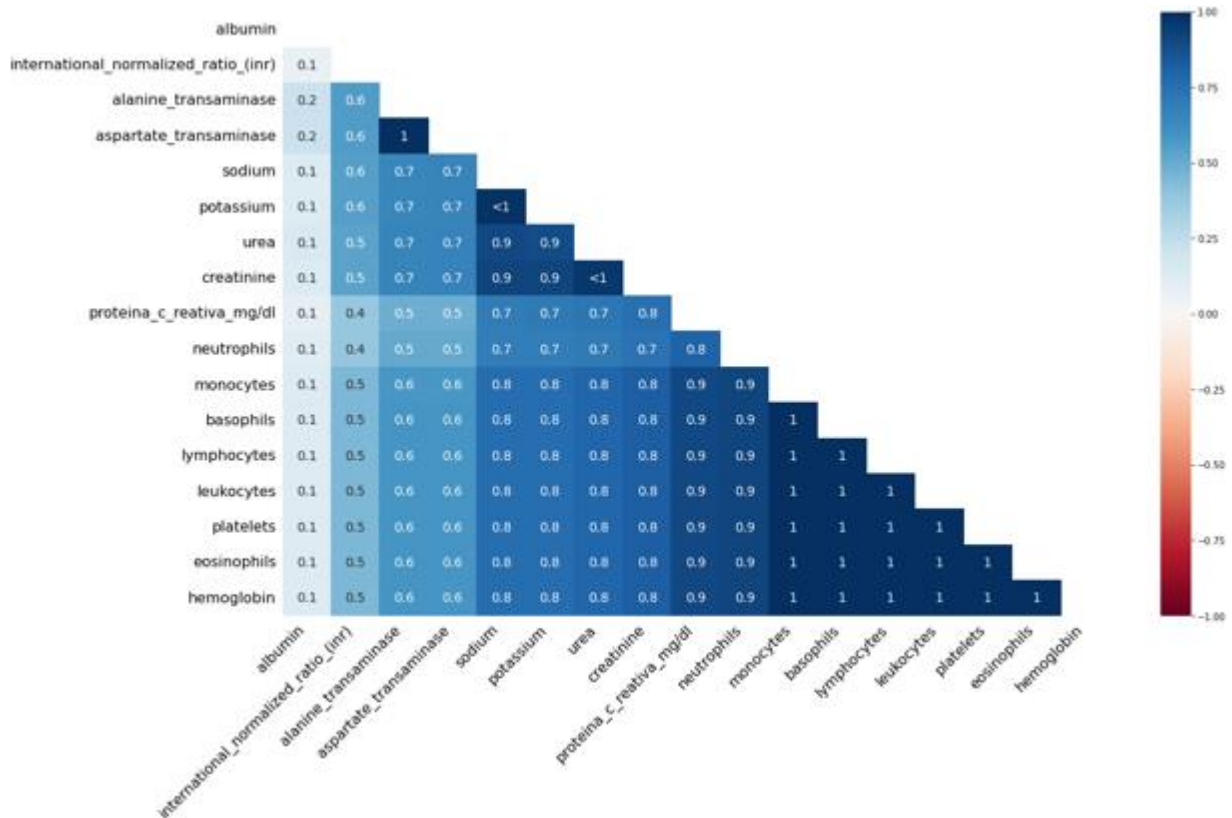
❖ Kết quả xử lý dữ liệu lỗi.

Dữ liệu được sử dụng trong thực nghiệm là của bệnh viện Israelita Albert Einstein ở Brazil về bệnh nhân CoViD-19. Dữ liệu này được thu thập tháng 4 đến tháng 5 năm 2020, gồm 5644 ca nhập viện trong đó có 553 trường hợp mắc bệnh CoViD-19 [31] bao gồm 110 thuộc tính mô tả về các ca bệnh. Số lượng mẫu của bộ dữ liệu được phân bổ như hình 5.2.



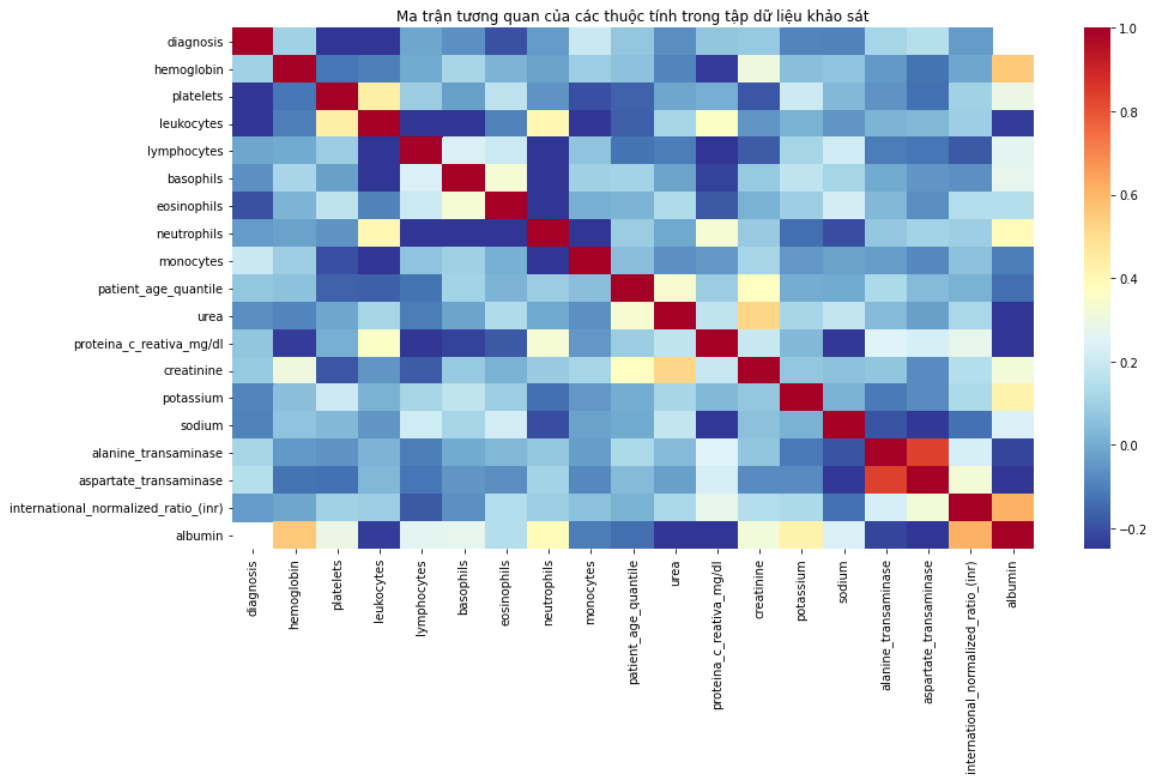
Hình 5.2: Số mẫu dữ liệu của hai lớp dữ liệu [67]

Luận án tiến hành loại bỏ dữ liệu lỗi bằng thuật toán 5.1. Kết quả của giai đoạn này là kết quả bỏ 1 thuộc tính “albumin”. Để có thể khẳng định thuộc tính này có thật sự là nên xóa được hay không, luận án còn khảo sát mức độ tương quan lỗi của toàn bộ tập thuộc tính bằng kỹ thuật kiểm tra tương quan lỗi của thư viện Missingno [75], kết quả kiểm tra thuộc tính “albumin” là thuộc tính có mức tương quan lỗi thấp nhất và không ảnh hưởng khi loại bỏ, kết quả kiểm tra được trình bày trong hình 5.3.



Hình 5.3: Mối tương quan giữa các thuộc tính khảo sát

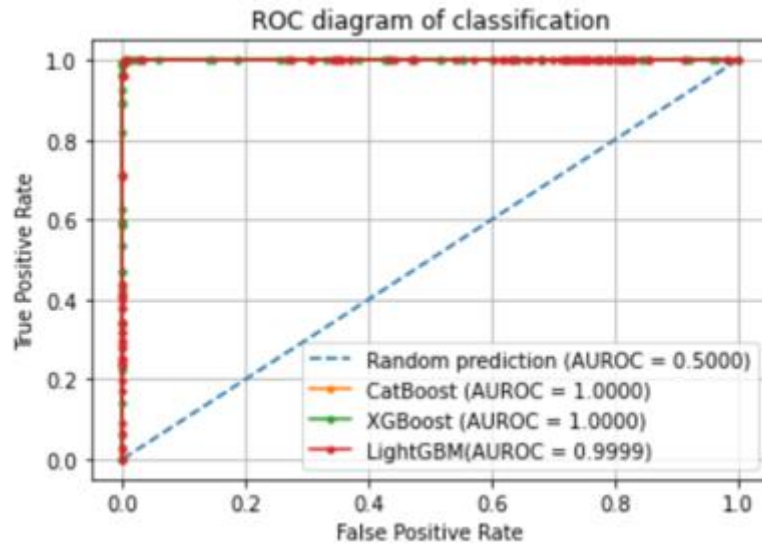
Ngoài ra luận án còn kiểm tra mức độ tương quan của toàn bộ thuộc của tập dữ liệu và lựa chọn các thuộc tính có độ tương quan cao trên 95% để xóa, kết quả là không có thuộc tính tương quan đạt mức độ đã nêu, nên không có dữ liệu cần loại bỏ cùng với kỹ thuật so sánh mối tương quan giữa các thuộc tính. Mối quan hệ tương quan giữa các thuộc tính được minh họa trong hình 5.4.



Hình 5.4: Mối tương quan giữa các thuộc tính khảo sát.

❖ **Kết quả mô hình phân loại từ dữ liệu lâm sàng**

Tập dữ liệu sau khi loại bỏ dữ liệu lỗi được gọi là Datasmall và được chia thành hai phần với 80% tập dữ liệu này dùng làm dữ liệu huấn luyện và 20% dùng để làm dữ liệu kiểm tra mô hình. Luận án đã sử dụng nhiều thuật toán khác để xây dựng bộ phân lớp dựa trên Datasmall như: XGBoost [39], LGBMClassifier [76], CatBoost [41]. Kết quả đánh giá AUROC của từng thuật toán được mô tả như hình 5.5. Nhận thấy XGBoost, cho giá trị AUROC tốt nhất nên dùng chọn dùng để xây dựng bộ phân lớp. Các kết quả chi tiết về hiệu năng của mô hình được xây dựng bằng XGBoost được mô tả bởi các hình 5.5



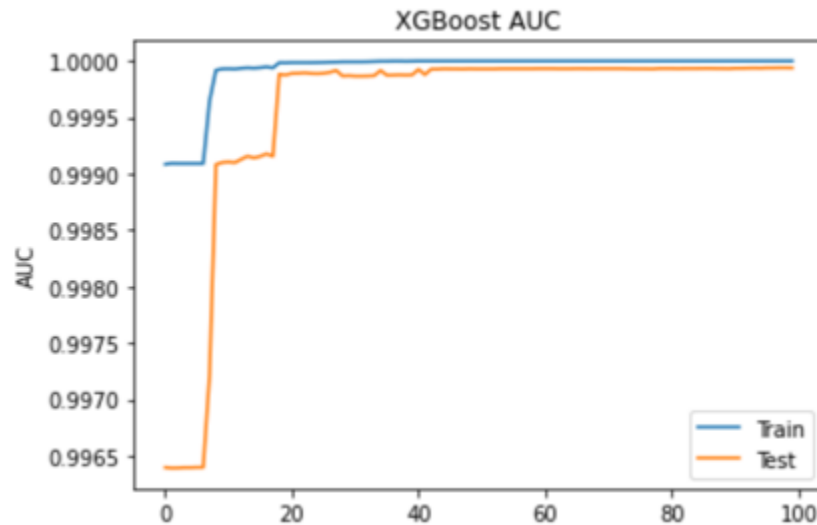
Hình 5.5: So sánh hiệu năng ROC của các thuật toán tăng cường độ dốc

❖ *So sánh hiệu năng mô hình thực nghiệm với các nghiên cứu khác.*

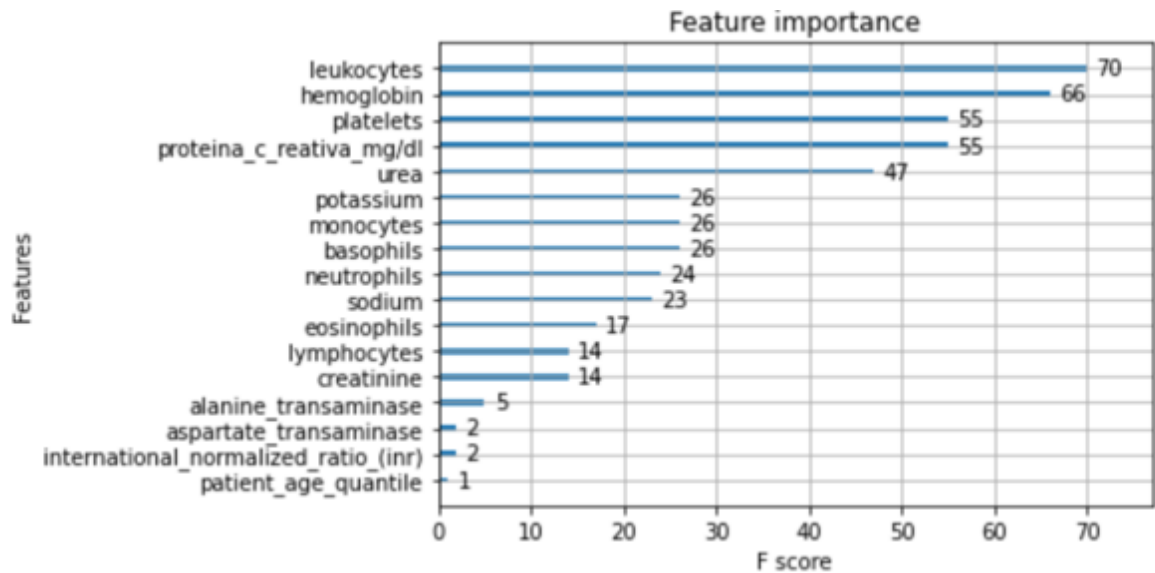
Các kết quả của mô hình đề xuất còn được so sánh với kết quả dự đoán cùng chức năng như nghiên cứu của tác giả Maryam AlJame [31]. Các độ đo hiệu năng của mô hình đề xuất được liệt kê chi tiết ở bảng 5.3.

Bảng 5.3: Bảng so sánh hiệu năng của mô hình đề xuất với tác giả Maryam AlJame

Bộ phân lớp	Kỹ thuật sử dụng	AUC	Precision (Specificity)	Recall (Sensitivity)	F1 score
Mô hình đề xuất của luận án	XGBoost	0.998	1.0	1.0	1.0
Mô hình đề xuất Maryam AlJame [31]	XGBoost	0.994	0.98	0.99	-



Hình 5.6: Hiệu năng ROC của các thuật toán đề xuất



Hình 5.7: Mức độ quan trọng chi tiết của các thuộc tính từ của mô hình đề xuất

Nhận thấy mô hình đề xuất với các cách xử lý dữ liệu đầu vào như loại bỏ thuộc tính có độ lỗi lớn (hơn 99%), điền khuyết 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.

5.3. Mô hình phân loại bệnh CoViD-19 và bệnh cúm mùa

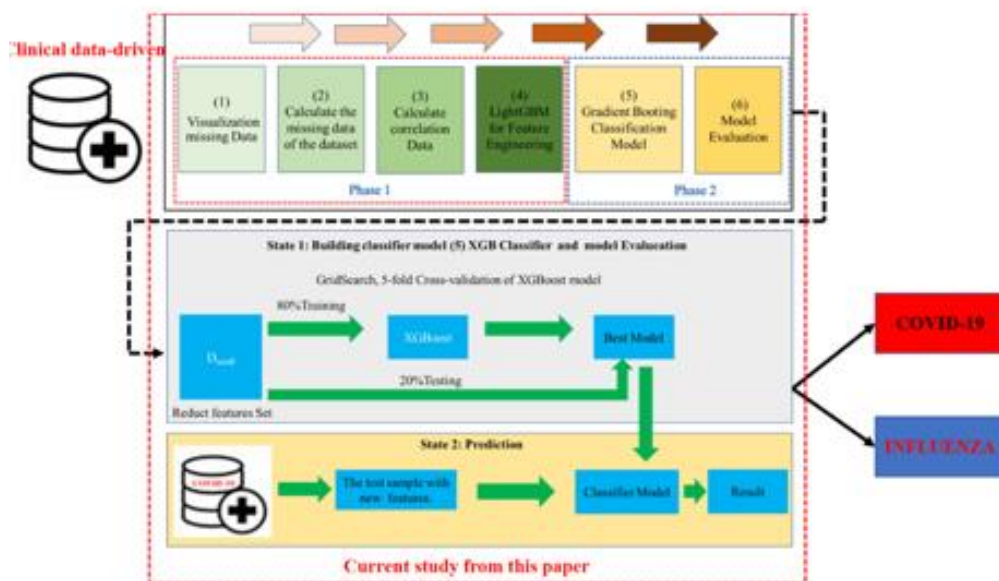
5.3.1. Giới thiệu

Dữ liệu lâm sàng và cận lâm sàng khi được thu thập tại các bệnh viện hay hồ sơ bệnh án luôn tìm ẩn chứa đựng rất nhiều thuộc tính bị thiếu, lỗi. Trong khi đó hầu hết các thuật toán

trên các nền tảng học máy đòi hỏi dữ liệu đầu vào có dạng số học và không được trống. Có thể sử dụng các kỹ thuật KNNImputer để xử lý dữ liệu bị thiếu, hoặc phương pháp mã hóa cho tập thuộc tính đó sử dụng các giá trị trung bình cho dữ liệu và có thể xóa dữ liệu với tỉ lệ lỗi cao. Với cách xử lý dữ liệu này giúp giải quyết vấn đề dữ liệu trống và chuyển đổi dữ liệu thành định dạng số phù hợp cho các thuật toán học máy. Tuy nhiên, việc xử lý trước này có thể ảnh hưởng đến kết quả dự đoán bằng cách bỏ qua dữ liệu chứa các thuộc tính quý giá 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 có thể học trực tiếp từ dữ liệu thô mà không cần sự can thiệp để xóa dữ liệu trống.

5.3.2. Mô hình phân biệt bệnh CoViD-19 và Cúm H1N1

Phương pháp và kỹ thuật hoạt động trong 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. Kết quả nghiên cứu cho thấy việc kết hợp hai mô hình gradient-boosting cả LightGBM và XGBoost để tạo ra các bộ phân loại tự động CoViD-19 và cúm từ dữ liệu lâm sàng đã cho kết quả tốt và hiệu suất vượt trội hơn so với một mô hình đơn lẻ, với độ chính xác tổng thể trên 99,96%. Mô hình đề xuất tổng quát như mô tả hình 5.8



Hình 5.8: Mô hình tổng quan đề xuất phân biệt COVID-19 và Cúm H1N1

5.3.3. Các thuật toán dùng trong mô hình đề xuất

Trong nghiên cứu này, luận án đã trích xuất và sử dụng 1485 mẫu kết quả xét nghiệm lâm sàng từ bệnh nhân nhiễm cúm mùa hoặc CoViD-19. Theo đó, tập dữ liệu chứa 51 thuộc tính được liệt kê trong dữ liệu lâm sàng CoViD-19 và cúm mùa. Quá trình phát triển cho nghiên cứu này bao gồm bốn giai đoạn chính:

- **Giai đoạn 1:** Tiền xử lý các thuộc tính của dữ liệu, bao gồm xác định giá trị dữ liệu bị thiếu, thực hiện phân tích tương quan của các thuộc tính và lọc dữ liệu nhiễu. Thuật toán được sử dụng trong giai đoạn này là thuật toán 5.2,

- **Giai đoạn 2:** Sử dụng mô hình LightGBM để phân tích các thuộc tính nổi bật có thể ảnh hưởng đến kết quả dự đoán. Thuật toán được sử dụng trong giai đoạn này là thuật toán 5.3,

- **Giai đoạn 3:** Xây dựng mô hình dự đoán với tham số được tối ưu hóa bằng GridSearch. Thuật toán được áp dụng trong giai đoạn này là 5.4.

- **Giai đoạn 3:** Đánh giá mô hình.

Các thuật toán được dùng trong các đề xuất lần lượt trình bày như sau:

❖ Thuật toán 5.1 loại bỏ thuộc tính có độ lỗi cao.

Dữ liệu y sinh thường chứa nhiều thuộc tính có giá trị lỗi do qui trình xét nghiệm hoặc do nhập liệu các kết quả xét nghiệm lâm sàng của bệnh nhân. Luận án đã sử dụng một thuật toán để phát hiện và loại bỏ tất cả các thuộc tính của dữ liệu đầu vào có tỷ lệ lỗi lớn hơn 99,9%. Luận án đã áp dụng Thuật toán 5.1 để loại bỏ thuộc tính có độ lỗi cao.

❖ Thuật toán trích chọn các thuộc tính có giá trị khác không dùng kỹ thuật LightGBM

Đây là một phương pháp tổng hợp để tích hợp cây quyết định vào kỹ thuật rút trích đặc trưng. Kỹ thuật LightGBM huấn luyện nhiều mô hình cây theo cách thêm vào mỗi mô hình cây mới được huấn luyện để dự đoán các sai số (tức là lỗi) của các mô hình trước đó. Phương pháp này tăng độ chính xác của tính toán lợi ích thông tin so với các phương pháp lấy mẫu đồng nhất như thuật toán GBDT. Chức năng đóng gói đặc trưng độc quyền (EFB) của LightGBM là hiệu quả đối với dữ liệu chiều cao và thưa. Giá trị khác không cho các thuộc tính được gọi là các thuộc tính độc quyền (EF) vì chúng không bao giờ có giá trị khác không đồng thời. Đối với nghiên cứu này, tập dữ liệu cho bệnh nhân CoViD-19 và cúm chứa dữ liệu cao và thưa, vì vậy 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. Quá trình trích xuất các thuộc tính quan trọng được thực hiện bởi thuật toán 5.2, 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

Tập dữ liệu cho bệnh nhân CoViD-19 và cúm mùa chứa hai loại thuộc tính: phân loại và định lượng. Cả hai thuộc tính này đều quan trọng cho phân loại và dự đoán nhóm bệnh của bệnh nhân khi thực hiện các thử nghiệm lâm sàng. Trong mô hình, luận án mã hóa các giá trị phân loại bằng kỹ thuật mã hoá nhãn: cụ thể luận án chuyển đổi mỗi giá trị trong một cột trong sơ đồ cây thành một số. Các nhãn số luôn nằm trong khoảng 0- n danh mục. Sau đó, luận án sử dụng tính năng categorical_feature của LightGBM để xác định tập thuộc tính quan trọng cho dữ liệu lâm sàng của bệnh nhân.

❖ Thuật toán Xây dựng bộ phân loại dùng kỹ thuật XGboost

Luận án sử dụng phương pháp tìm kiếm lưới để tìm tham số tối ưu cho mô hình đề xuất. Quá trình tìm kiếm lưới trên tập huấn luyện, sử dụng kỹ thuật cross-validation với $k = 5$ để chọn ra mô hình phân loại tốt nhất. Một tập dữ liệu mới D_{small} được thu được từ kết quả của

giai đoạn 1, trong đó n là số lượng mẫu dữ liệu và m là số lượng thuộc tính. Ba tham số chính được sử dụng cho quá trình điều chỉnh: số cây K (`n_estimators`), độ sâu của cây quyết định d (`colsample_bytree`) và hệ số co giãn (`learning_rate`). Bộ tham số được tối ưu được sử dụng như đầu vào cho thuật toán để phân loại và đánh giá toàn bộ quá trình. Chi tiết thuật toán 5.3 được mô tả như sau:

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

Với phương pháp XGBoost không sử dụng biến mục tiêu y trực tiếp như giá trị dự báo, mà thay thế nó bằng sai số của mô hình trước đó. Các sai số được cập nhật dần dần theo hệ số co giãn, để chuyển đổi chuỗi mô hình thành các cây quyết định đa dạng hơn. Các mô hình ngừng được thêm vào khi đạt đến số lượng mô hình dự đoán tối đa hoặc tất cả các quan sát được phân loại hoặc dự đoán đúng.

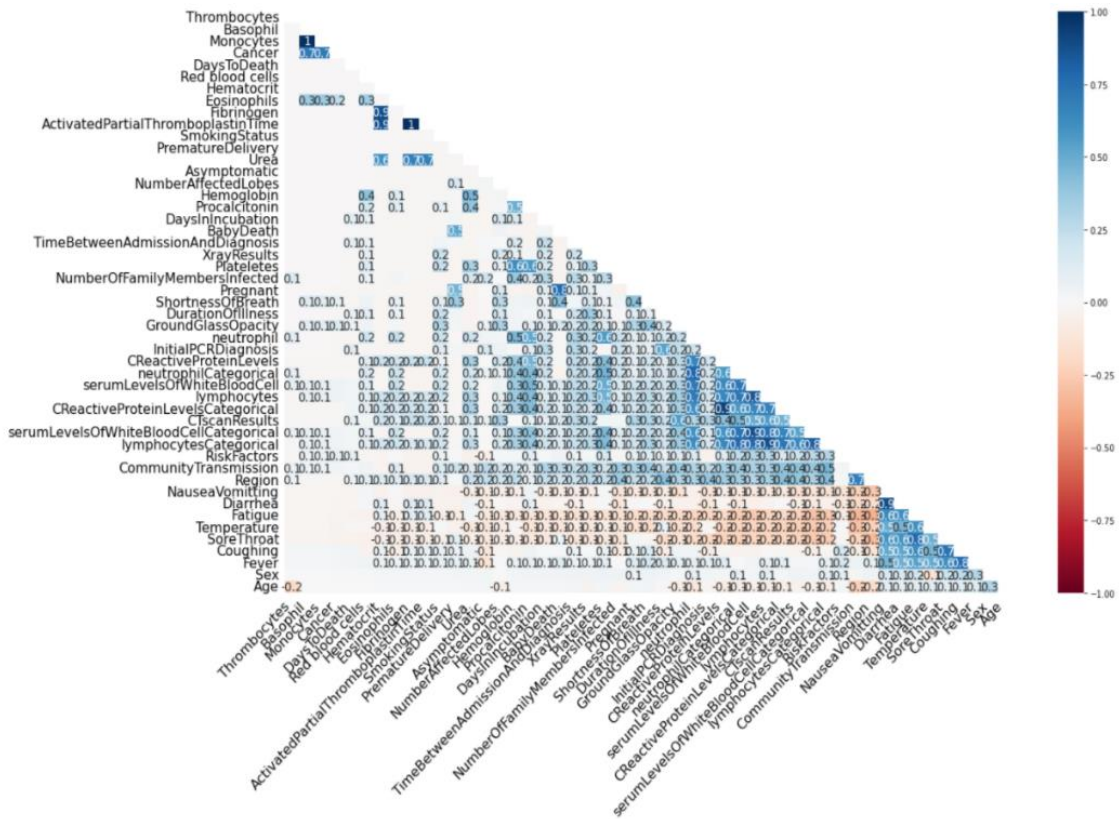
5.3.4. Kết quả thực nghiệm

❖ Khảo sát dữ liệu lỗi

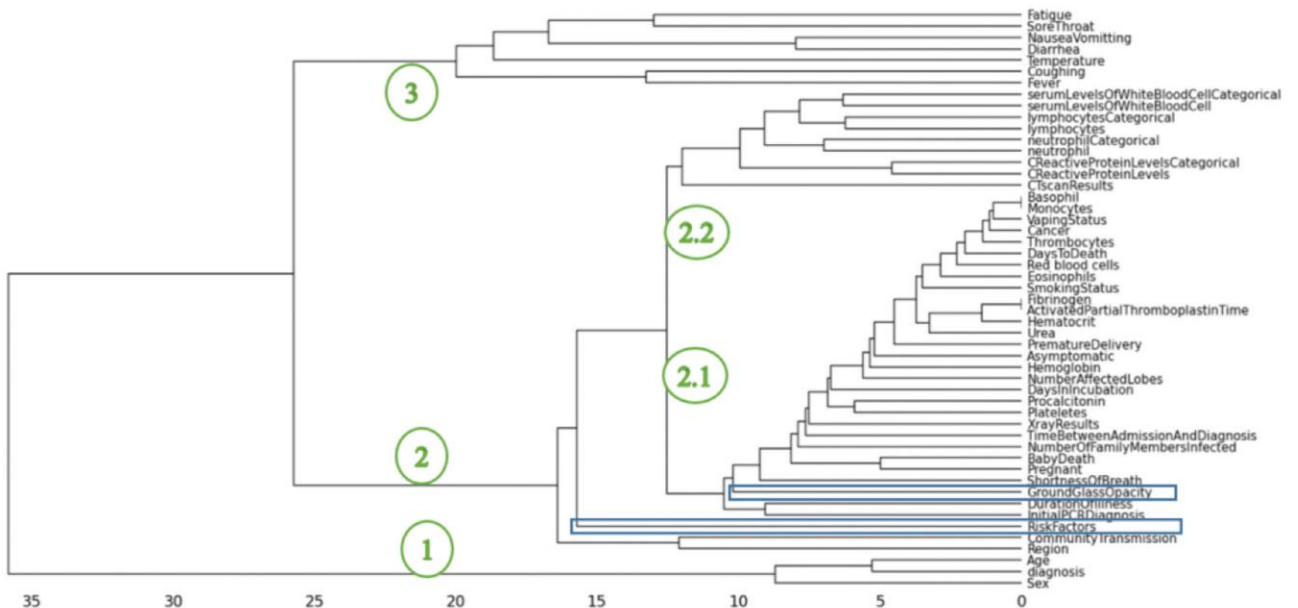
Dữ liệu lỗi trong y khoa rất thường xuất hiện vì nhiều lý như do dữ liệu có thể đến từ việc thu thập; lấy mẫu từ các thử nghiệm lâm sàng khác nhau, trong phòng thí nghiệm hoặc theo cơ chế truyền thống lấy mẫu khác nhau. Hiện thị toàn cảnh về dữ liệu lỗi giúp việc chọn lựa đặc trưng tốt hơn cho mô hình học máy và không làm giảm độ chính xác của mô hình. Thông thường khi

xử lý dữ liệu lỗi cho mô hình học máy thì việc loại bỏ missing values là rất thông dụng. Tuy nhiên đối với dữ liệu dùng trong y tế thì việc này không nên thực hiện vì dữ liệu có thể có độ lỗi cao nhưng có thể chứa nhiều thông tin quan trọng cho quá trình chuẩn đoán. Việc trực quan hóa các mối quan hệ giữa các cột bị lỗi giúp đánh giá tổng quan toàn cảnh dữ liệu trước khi tiến hành các quá trình xử lý dữ liệu tránh xóa nhầm các đặc trưng quan trọng. Luận án đã sử dụng kỹ thuật trực quan hóa dữ liệu bằng gói thư viện Missingno [75] để trực quan hóa dữ liệu lỗi.

Tập dữ liệu Dogr được đề cập trong nghiên cứu [66] gồm 1072 bệnh nhân mắc cúm mùa hoặc cúm H1N1 và 413 bệnh nhân mắc CoViD-19. Tập dữ liệu chứa 50 thuộc tính, được coi là 50 đặc trưng, và một biến chuẩn đoán được đánh dấu được sử dụng như biến phân loại. Trong hình 5.9, ma trận tương quan thể hiện chi tiết mức độ lỗi của từng đặc trưng theo thang đo từ thấp đến cao. Kết quả ở ma trận tương quan ở hình này cho thấy các đặc trưng có mức độ lỗi cao nhất là: 'Cancer', 'SmokingStatus', 'Monocytes', 'Thrombocytes', 'DaysToDeath', 'VapingStatus', 'Eosinophils', 'Hematocrit', 'ActivatedPartialThromboplastinTime', 'Red blood cells', 'Basophil', 'Fibrinogen'. Tuy có độ lỗi cao nhưng cũng cần phải xem xét cẩn thận trước khi xóa vì có thể các chi tiết có độ lỗi cao nhưng nếu chúng có mối liên hệ với nhiều nhiều đặc trưng khác thì không nên xóa vì đây có thể là những thuộc tính có giá trị quan trọng quyết định việc chuẩn đoán bệnh Covid-19 hay bệnh cúm. Luận án đã xem xét tổng quát các mối quan hệ lỗi này bằng biểu đồ dendrogram như hình hình 5.10, biểu đồ dendrogram thể hiện mối quan hệ giữa các cột dựa trên mối quan hệ giữa chúng dạng phân cấp. Việc này giúp nhóm các cột có cùng mức độ lỗi lại với nhau. Kết quả hiển thị cho thấy đặc trưng RiskFactor có độ độ lỗi 84,78% nhưng có liên quan đến liên quan đến hai nhóm thuộc tính khác ở nhóm 2.1 và 2.2. Nên thuộc tính này không xóa được. Hay đối với thuộc tính “GroundGlassOpacity” có tỷ lệ lỗi đến 93,6%, nhưng lại có liên quan đến tất cả các thuộc tính của nhóm 2.1.



Hình 5.9: Hiển thị giá trị lỗi của các cột và mối liên quan của chúng với nhau, trong các cột khác trong toàn bộ dữ liệu



Hình 5.10: Hiển thị giá trị lỗi của các cột và mối liên quan theo biểu đồ dendrogram

❖ Kết quả của giai đoạn tiền xử lý dữ liệu

Tập dữ liệu D_{ogr} được đề cập trong nghiên cứu [66] gồm 1072 bệnh nhân mắc cúm mùa hoặc cúm H1N1 và 413 bệnh nhân mắc CoViD-19. Tập dữ liệu chứa 50 thuộc tính, được coi là 50 đặc trưng, và một biến chẩn đoán được đánh dấu được sử dụng như biến phân loại. Áp dụng thuật toán 5.2 cho tập dữ liệu D_{ogr} đã tự động loại bỏ 12 thuộc tính (đánh dấu**) chi tiết về tập dữ liệu D_{ogr} cùng với các tỷ lệ lỗi được mô tả Bảng 5.4.

Bảng 5.4: Thuộc tính của bộ dữ liệu D_{org}

STT	Thuộc tính	Type	Số lượng giá trị của từng thuộc tính	Số mẫu dữ liệu	Tỷ lệ giá trị bị thiếu (%)
1	Diagnosis	int64	2	1485	0
2	ActivatedPartialThromboplastinTime**	float64	9	9	99.39
3	Age	float64	09	1457	1.89
4	Basophil**	float64	1	1	99.93
5	DaysInIncubation	float64	14	38	97.44
6	DaysToDeath**	float64	3	3	99.80
7	DurationOfIllness	float64	30	88	94.07
8	Eosinophils**	float64	3	8	99.46
9	Fibrinogene**	float64	9	9	99.39
10	Hematocrit**	float64	7	7	99.53
11	Hemoglobin	float64	18	24	98.38
12	Monocytes**	float64	1	1	99.93
13	Neutrophil	float64	91	103	93.06
14	NumberOfAffectedLobes	float64	5	24	98.38
15	NumberOfFamilyMembersInfected	float64	8	54	96.36
16	Platelets	float64	44	50	96.63
17	Procalcitonin	float64	23	33	97.78
18	RedBloodCells**	float64	4	4	99.73

19	SerumLevelsOfWhiteBloodCells	float64	27	151	89.83
20	Temperature	float64	35	629	57.64
21	Thrombocytes**	float64	1	1	99.93
22	TimeBetweenAdmissionAndDiagnosis	float64	15	47	96.84
23	Urea	float64	13	19	98.72
24	VapingStatus*	float64	0	0	100
25	Asymptomatic	object	2	21	98.59
26	BabyDeath	object	3	40	97.31
27	Cancer**	object	1	2	99.87
28	CommunityTransmission	object	7	249	83.23
29	Coughing	object	3	862	41.95
30	CReactiveProteinLevels	object	22	139	90.64
31	CReactiveProteinLevelsCategorical	object	4	158	89.36
32	CTScanResults	object	6	161	89.16
33	Diarrhea	object	2	450	69.7
34	Fatigue	object	2	531	64.24
35	Fever	object	3	926	37.64
36	GroundGlassOpacity	object	2	95	93.6
37	InitialPCRDagnosis	object	4	104	93
38	Lymphocytes	object	28	158	89.36
39	LymphocytesCategorical	object	6	195	86.87
40	NauseaVomitting	object	2	422	71.58
41	NeutrophilCategorical	object	4	148	90.03
42	Pregnant	object	3	65	95.62

43	PrematureDelivery	object	3	15	98.99
44	Region	object	63	318	78.59
45	RiskFactors	object	65	226	84.78
46	SerumLevelsOfWhiteBloodCellsCategorical	object	8	191	87.14
47	Sex	object	2	1409	5.12
48	ShortnessOfBreath	object	3	75	94.95
49	SmokingStatus**	object	3	10	99.33
50	SoreThroat	object	2	670	54.88
51	XrayResults	object	2	47	96.84

❖ Áp dụng thuật toán 5.2 cho quá trình trích xuất thông tin

Sau khi áp dụng thuật toán 5.1, Luận án tiếp tục sử dụng thuật toán 5.2 để chọn lựa thuộc tính có độ lợi thông tin cao để xây dựng bộ phân loại. Kết quả của quá trình này chúng tôi đã rút trích được 24 thuộc tính quan trọng và tập thuộc tính mới này được gọi là cho tập dữ liệu SmallData. Để đánh giá mức độ đúng đắn và hiệu quả của kết quả dữ liệu thu được, Luận án đã so sánh kết quả thu được ở giai đoạn này với kết quả của Wei Tse Li. Thông tin chi tiết về kết quả được trình bày ở Bảng 5.5

Để đánh giá độ chính xác và hiệu quả của kết quả được thu được từ mô hình, luận án so sánh kết quả ở giai đoạn này với kết quả của Wei Tse Li [77]. Chi tiết của so sánh được trình bày trong Bảng 5.5

Bảng 5.5: Kết quả so sánh giữa việc lựa chọn đặc trưng của mô hình của luận án và mô hình của Wei Tse Li [77]

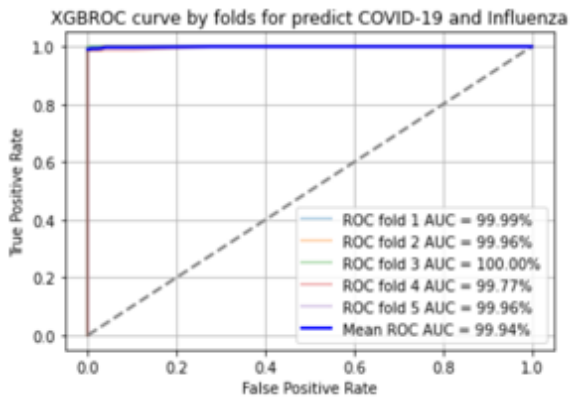
Tác giả	Phương pháp chọn lựa thuộc tính	Các đặc trưng quan trọng trong chọn lựa từ mô hình phân loại
Wei Tse Li [77]	Sử dụng kết hợp: Lựa chọn thủ công dựa vào thống kê và kết hợp	Diagnosis, Neutrophils, NeutrophilsCategorical, SerumLevelsOfWhiteBloodCells,

	mạng neural SOM để chọn thuộc tính.	SerumLevelsOfWhiteBloodCellsCategorical, Lymphocytes, LymphocytesCategorical, CTScanResults, RiskFactors, Diarrhea, Fever, Coughing, ShortnessOfBreath, SoreThroat, NauseaVomitting, Temperature, Fatigue, XrayResults
Đề xuất của luận án	Sử dụng các thuật toán 5.1, thuật toán 5.2 cho quá trình trích xuất thông tin	Diagnosis, Neutrophils, NeutrophilsCategorical, SerumLevelsOfWhiteBloodCells, SerumLevelsOfWhiteBloodCellsCategorical, Lymphocytes, LymphocytesCategorical, CTScanResults, RiskFactors, Diarrhea, Fever, Coughing, ShortnessOfBreath, SoreThroat, NauseaVomitting, Temperature, Fatigue, CReactiveProteinLevels, GroundGlassOpacity, InitialPCRDagnosis, NumberOfFamilyMembersInfected, Region

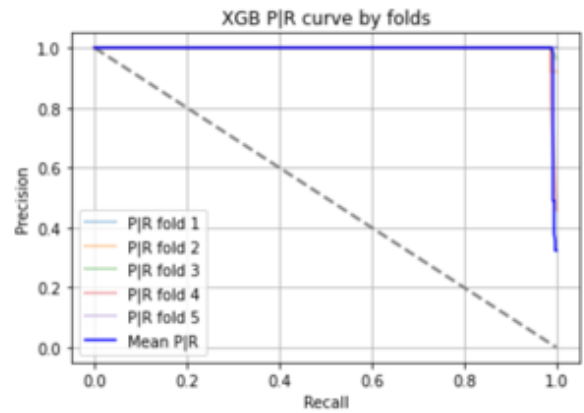
Như được thể hiện trong Bảng 5.5, mô hình của luận án đã có 24 thuộc tính được lựa chọn, và kết quả mà luận án thu được phản ánh gần như toàn bộ tập thuộc tính được đề xuất bởi Wei Tse Li . Tuy nhiên, có một thuộc tính bổ sung - kết quả X-quang. Một ưu điểm của mô hình của luận án là nó chọn hơn năm thuộc tính so với mô hình của Li, trong đó thuộc tính *GroundGlassOpacity* trong đó thuộc tính 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%. Điều này đặc biệt quan trọng cho chẩn đoán và điều trị bệnh nhân nặng. Do đó, 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 hiệu năng tiềm ẩn của dữ liệu. Kết quả đã rõ ràng cho thấy rằng việc lựa chọn đặc trưng cho mô hình phân loại của luận án là phù hợp, chọn ra các đặc trưng quan trọng cho mô hình dự đoán.

❖ Áp dụng thuật toán 5.3. xây dựng mô hình phân loại

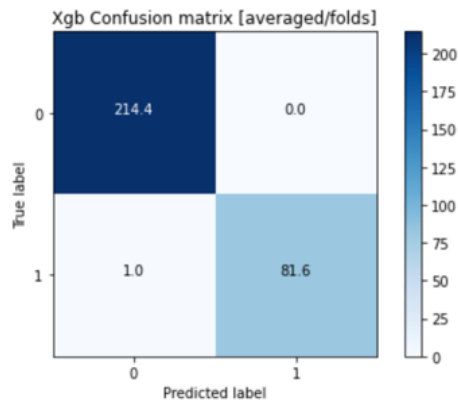
Với bộ dữ liệu DataSmall thu được sau khi áp dụng thuật toán 5.2, Luận án áp dụng thuật toán 5.3 để xây dựng bộ phân loại. Hiệu năng của mô hình phân loại được xây dựng được đánh giá hiệu năng bằng các độ đo đường cong ROC, và Precision-Recall Curve đoán và Confusion Matrix để đánh giá mô hình huấn luyện dữ liệu, luận án đã kiểm tra quá trình phân loại năm lần cho thuật toán XGBoost. Kết quả trung bình cho mỗi bước của cross-validation được thể hiện trong các hình 5.11 đến hình 5.14, cùng với kết quả đánh giá hiệu suất của phương pháp cross-validation năm lần trên tập SmallData.



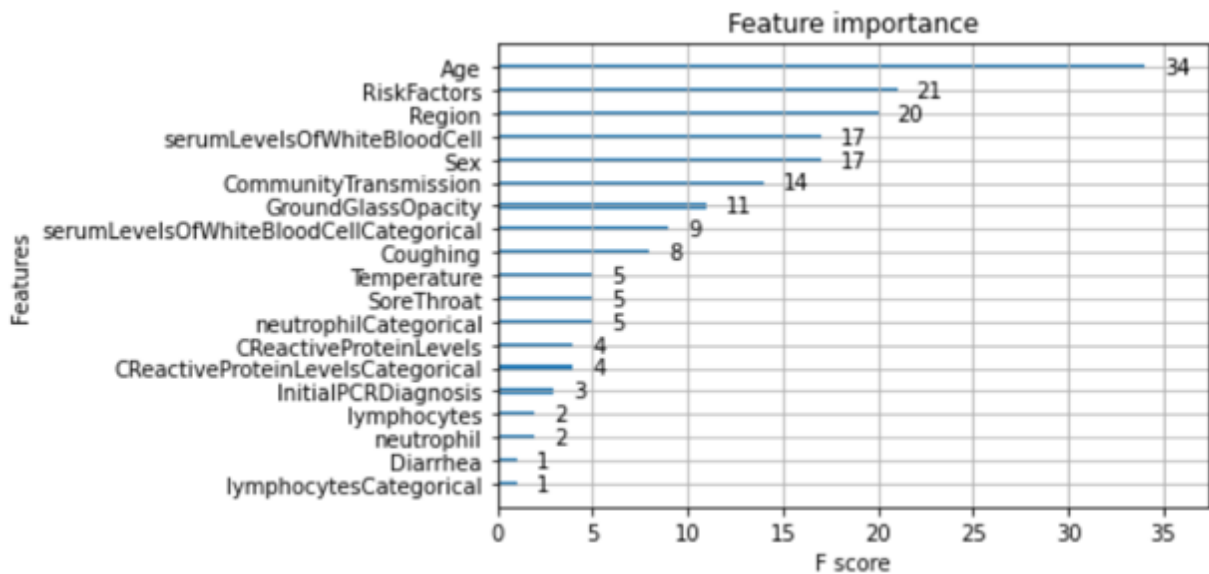
Hình 5.11: ROC curve cho mô hình dự đoán



Hình 5.12: Precision recall curve cho mô hình dự đoán



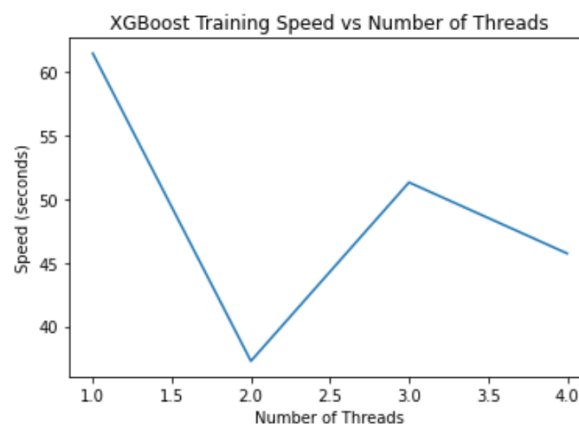
Hình 5.13: Confusion matrix cho mô hình dự đoán.



Hình 5.14: Thuộc tính quan trọng của mô hình dự đoán

❖ Thời gian thực hiện của mô hình đề xuất

Bên cạnh việc tính toán hiệu năng của mô hình phân loại được xây dựng, luận án còn đo đạt thời gian thực nghiệm. XGBoost hỗ trợ cơ chế xử lý song song nên việc tính toán thời gian được khảo sát theo luồng (threads) trong quá trình huấn luyện. Kết quả thời gian thực nghiệm được mô tả như hình 5.15



Hình 5.15: Thời gian thực nghiệm xây dựng mô hình phân loại

5.3.5. So sánh hiệu năng của mô hình đề xuất với mô hình phân loại khác trên cùng bộ dữ liệu và nhiệm vụ dự đoán

Tập các đặc trưng quan trọng thu được từ mô hình phân loại cũng được so sánh với tập đặc trưng của mô hình được xây dựng bởi Li. Trong nghiên cứu của Li đã sử dụng kỹ thuật mã hóa

one-hot để mã hóa tập dữ liệu định danh với tập đặc trưng đầu vào có 20. Thuật toán phân loại XGBoost đã giảm số lượng đặc trưng quan trọng từ 24 xuống 14. Thuộc tính *GroundGlassOpacity* vẫn được giữ lại bởi mô hình trong 14 đặc trưng quan trọng được sử dụng để dự đoán CoViD-19 hoặc cúm. Để kiểm tra tính chính xác của mô hình của luận án, luận án đã so sánh hiệu suất của nó với mô hình của Li. Li đã sử dụng nhiều phương pháp học máy để xây dựng mô hình phân loại, chẳng hạn như hồi quy RIDGE, rừng ngẫu nhiên, hồi quy LASSO và XGBoost. Ba kỹ thuật đầu tiên đã được thử nghiệm, và sử dụng phương pháp đánh giá AUC, kết quả lần lượt là 96,6%, 95,3% và 96,3%. Ưu điểm của phương pháp của luận án là sử dụng các phương pháp đánh giá mô hình AUC, ROC, độ nhạy và đặc hiệu. Chi tiết kết quả đánh giá được trình bày trong 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	-	-	-

Được xây dựng từ tập dữ liệu tổng hợp các triệu chứng lâm sàng của bệnh CoViD-19 và cúm, luận án này đã xây dựng một bộ phân loại và dự đoán cho CoViD-19 và cúm. Việc kết hợp hai phương pháp học máy LightGBM và XGBoost trong mô hình đề xuất đã mang lại hiệu quả tốt hơn trích xuất dữ liệu thô. Mô hình hoạt động tốt trong việc khám phá các biến quan trọng cho mô hình dự đoán từ kết quả thử nghiệm lâm sàng. Điều này tăng độ hiệu quả của dữ liệu đầu vào cho quá trình đánh giá và giảm số lượng mẫu bị lỗi. Kết quả cho thấy phương pháp đề xuất có thể kiểm soát và xử lý tự động các biến phân loại từ một tập dữ liệu thô, qua

đó tối đa hóa giá trị của tập biến số đầu vào cho mô hình phân loại. Kết quả của mô hình ổn định và có thể phục vụ làm cơ sở triển khai trên các tập dữ liệu lâm sàng lớn hơn tại các bệnh viện, từ đó làm cho quá trình chẩn đoán và phát hiện bệnh, cùng với điều trị, hiệu quả hơn.

5.4. 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 chẩn đoán bệnh và phân biệt bệnh. Đối với mô hình thứ nhất luận án đã đề xuất phương pháp tích hợp các thuật toán tự động loại bỏ cột thiếu dữ liệu, kỹ thuật KNNimputer và MOTE để xử lý dữ liệu trống và giảm mất cân bằng dữ liệu. Khi so sánh kết quả của mô hình thứ nhất với các nghiên cứu cùng mục tiêu và bộ dữ liệu, hiệu năng của mô hình đề xuất cho kết quả tốt hơn. Mô hình đề xuất thứ hai cũng sử dụng dữ liệu cận lâm sàng làm dữ liệu huấn luyện để xây dựng bộ phân định bệnh CoViD-19 và bệnh Cúm. Nghiên cứu này đã chứng minh việc sử dụng học máy thông qua mô hình kết hợp từ việc phối hợp các phương pháp xử lý dữ liệu tự động bằng kỹ thuật LightGBM cho việc rút gọn chiều và mô hình phân loại bằng XGBoost. Phương pháp đề xuất này giúp mô hình phân loại bệnh này đã phát huy hiệu quả khi phát hiện thêm các biến quan trọng cho mô hình dự đoán từ kết quả xét nghiệm lâm sàng. Điều này làm tăng số liệu đầu vào trong quá trình đánh giá và giảm đi số lượng mẫu lỗi. Về độ chính xác cũng như các hiệu năng khác đã được kiểm chứng với các công trình đã công bố có kết quả tốt hơn về hiệu năng cũng như ý nghĩa thực tiễn.

Luận án đã trình bày một phương án học máy có thể hỗ trợ cho việc phát hiện các ca bệnh có các triệu chứng lâm sàng gần giống như cúm. Kết quả của mô hình đề xuất chỉ thử nghiệm trên bộ dữ liệu được công đồng nghiên cứu cung cấp với các triệu chứng lâm sàng của bệnh thuộc chủng loại virus corona. Ngoài ra, phương pháp này có thể áp dụng cho các loại bệnh khác nhau. Việc áp dụng học máy vào chẩn đoán bệnh dựa trên dữ liệu lâm sàng hay cận lâm sàng đòi hỏi sự cẩn thận và kiểm soát chặt chẽ. Để đảm bảo tính chính xác và đáng tin cậy, mô hình cần được bổ sung dữ liệu huấn luyện đủ lớn, đồng thời dữ liệu này cần được xác minh và kiểm tra bởi các nghiên cứu lâm sàng và chuyên gia y tế. Kết quả của mô hình luôn ổn định và làm cơ sở để có thể triển khai trên tập dữ liệu lớn hơn từ nguồn dữ liệu lâm sàng thực tế tại các bệnh viện. Phương pháp được đề xuất từ các mô hình đề xuất của luận án đóng góp cho việc sàng lọc, phân loại và dự đoán chính xác bệnh nhân mắc bệnh. Mô hình đề xuất giảm thiểu chi

phí phát sinh trong quá trình khám chữa bệnh và giảm áp lực về thời gian trả kết quả xét nghiệm của bệnh nhân. Trong tương lai, mô hình được phát triển giúp cho việc chẩn đoán bệnh nhân trở nên đơn giản và chính xác 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 chẩn 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 hiện được công việc như sau:

- 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 các giải pháp đề xuất tập trung vào thực hiện các công việc như: 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 và cận lâm sàng trong về chẩn đoán bệnh.

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.

Các đóng góp trên đã giúp giải quyết các vấn đề trong lĩnh vực sinh học phân tử và chẩn đoán bệnh bằng sử dụng các phương pháp học máy và thuật toán khác nhau. Chúng cung cấp những phương pháp và kỹ thuật mới để xử lý dữ liệu và xây dựng các mô hình dự đoán hiệu quả trong các ứng dụng sinh học.

Các phương pháp thực nghiệm đã mang lại hiệu năng tốt hơn cho các mô hình học máy trong các bài toán về phân lớp gene, chẩn đoán bệnh. Tính hiệu quả của các mô hình phân lớp được chứng minh bằng kết quả thực nghiệm trên dữ liệu thực được lấy từ các kho dữ liệu Ngân hàng Gene quốc tế NCBI, từ các luận án công bố trên các tạp chí về tin sinh học. Để đánh giá tính tổng quát của các mô hình đề xuất luận án đã thực nghiệm trên cùng bộ dữ liệu với các nghiên cứu khác đã công bố và luận án tiến hành thu thập, xử lý dữ liệu, xây dựng, huấn luyện và đánh giá mô hình. Các mô hình đề xuất đã được đánh giá bằng cách so sánh kết quả độ chính xác phân lớp với các mô hình cơ bản khác. Ngoài ra, luận án cũng đo thêm thời gian thực hiện của các mô hình để phân tích và đánh giá hiệu năng giữa các mô hình. Kết quả tổng thể của các giải pháp tổng thể đề đạt hiệu năng cao hơn các nghiên cứu hiện tại trong cùng phạm vi dữ liệu và mục tiêu nghiên cứu. Trong tương lai các đề xuất này có thể triển khai trên nhiều bài toán khác nhau trong cùng lĩnh vực y sinh.

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]. **Duong Thi 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)

TÀI LIỆU THAM KHẢO

- [1] Mireya Martínez-García and Enrique Hernández-Lemu , “Data Integration Challenges for Machine Learning in Precision Medicine,” *Translational Medicine*, 2022.
- [2] Juan Jovel, Russell Greiner, “An Introduction to Machine Learning Approaches for Biomedical Research,” 2021.
- [3] Hoàng Trọng Phán, Trương Thị Bích Phượng, Giáo trình Di truyền học, vi sinh vật và ứng dụng, ĐH Huế, 2008.
- [4] Hoover, D.M. and J. Lubkowski, “DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis,” *Nucleic Acids Research*, 2002.
- [5] Annabel HAParret, HüseyinBesir, RobMeijers, “Critical reflections on synthetic gene design for recombinant protein expression,” *Elsevier*, tập 38, 2016.
- [6] Molly Hunter, Ping Yuan, Divya Vavilala, and Mark Fox , “Optimization of Protein Expression inMammalian Cells,” *John Wiley & Sons*, 2018.
- [7] Paulo Gaspar1, Jose’ Lui’s Oliveira , Jo”rg Frommlet, Manuel A.S. Santos and Gabriela Moura, “EuGene: Maximizing synthetic gene design for the heterologous expression,” *Bioinformatics applications note*, 2012.
- [8] Pere Puigbo, E.G., Antoni Romeu1 and Santiago Garcia-Vallve, “A web server for optimizing the codon usage of DNA sequences,” 2007.
- [9] Sharp, P.M, Tuohy, .TM, Mosurski, K.R, “Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed gene,” *Nucleic Acids Res*, 1987.
- [10] T. V. Lãng, Ứng dụng Tin học trong việc giải một số bài toán của Sinh học phân tử, Giáo Dục, 2008.
- [11] N. Đ. Thành, “Các kỹ thuật chỉ thị dna trong nghiên cứu và chọn lọc thực vật,” *Tạp chí sinh học*, pp. 265-294, 2014.
- [12] Wei Tse Li; Jiayan Ma; Neil Shende; Grant Castaneda; Jaideep Chakladar; Joseph C. Tsai; Lauren Apostol; Christine O. Honda; Jingyue Xu; Lindsay M. Wong; Tianyi Zhang; Abby Lee; Aditi Gnanasekar; Thomas K. Honda; Selena Z. Kuo; Michael Andrew Yu; Eric Y. Chang; Mahadevan; Rajasekaran; Weg M. Ongkeko, “Using machine learning of clinical data to diagnose CoViD-19: a systematic review and meta-analysis,” *BMC Medical Informatics and Decision Making*, tập 20, p. ID. 247, 2020.
- [13] Vanessa Damazio Teich, et all, “Epidemiologic and clinical features of patients with CoViD-19 in Brazil,” *einstein_journal*, 2020.

- [14] Wanshan Ning, Shijun Lei, Jingjing Yang, Yukun Cao, Peiran Jiang, Qianqian Yang, Jiao Zhang, Xiaobei Wang, Fenghua Chen, Zhi Geng, LiaLin Wang, Yu Xue and Zheng Wang, “Open resource of clinical data from patients with pneumonia for the prediction of CoViD-19 outcomes via deep learning,” *Nature biomedical engineering*, tập VOL 4, p. 1197–1207, DECEMBER 2020.
- [15] Dehua Wang, Yang Zhang và Yi Zhao, “LightGBM: An Effective miRNA Classification Method in Breast Cancer Patients,” trong *ICCB 2017: Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics*, Univ. of Nebraska at Omaha, USA, 2017.
- [16] Tianqi Chen; Tong He Michael Benesty; Vadim Khotilovich; Yuan Tang, “Xgboost: extreme gradient boosting,” *R package version 0.4-2*, tập 1, số 4, pp. 1-4, 2015.
- [17] J. a. J. L. Ye, “ Sparse methods for biomedical data.”, *ACM* , tập Sigkdd Explorations Newsletter 14, pp. 4-15, 2012.
- [18] H. a. X. L. Han, “The challenges of explainable AI in biomedical data science,” *BMC*, tập bioinformatics 22, pp. 1-3, 2021.
- [19] K. S. P. R. R. I. Z. a. B. Bentele, “Efficient translation initiation dictates codon usage at gene start,” *Molecular Systems Biology*, tập 9, p. 675, 2013.
- [20] T. J. B. T. L. S. J. W. T. a. T. J. White, “Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics.,” *PCR Protoc. a guide methods Appl*, tập 18 (1), p. 315–322, 1990.
- [21] S. a. I. S. Kaur, “ Artificial intelligence based clinical data management systems: a review,” *Informatics in Medicine Unlocked* 9, pp. 219-229, 2017.
- [22] Angela Serra†, Paola Galdi†, Roberto Tagliaferri, “ Machine learning for bioinformatics and neuroimaging,” *WIREs Data Mining Knowl Discov*, 2018.
- [23] H. A. a. G. B. Gaspar, “Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics." 20, no. 1 (2019): .,” *BMC bioinformatics*, Các tập %1 của %220, no. 1, pp. 1-11.
- [24] Malihe Ram, Ali Najafi, and Mohammad Taghi Shakeri, “Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest,” *Iranian Journal of Pathology*, 2017.
- [25] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, tập 29, pp. 1189-1232, 2001.
- [26] E. Weissler, T. Naumann, T. Andersson, R. Ranganath, O. Elemento, Y. Luo, D. Freitag, J. Benoit, M. Hughes, F. Khan và e. al., “The role of machine learning in clinical research: Transforming the future of evidence generation.,” *Trials*, 2021.

- [27] Buvailo, “A. Artificial Intelligence in Drug Discovery and Biotech: 2022 Recap and Key Trends,” *Available online*, Các tập %1 của %2<https://www.biopharmatrend.com/post/615-pharmaceutical-artificial-intelligence-key-developments-in-2022/> (truy cập ngày 31/1/ 2023, 2022).
- [28] A. Sharma, T. Virmani, V. Pathak, A. Sharma, K. Pathak, G. Kumar và D. Pathak, “Artificial Intelligence-Based Data-Driven Strategy to Accelerate Research, Development, and Clinical Trials of COVID Vaccine.,” *Biomed Res Int.* 2022, 2022.
- [29] S. Bagabir, N. Ibrahim và R. Ateeq, “CoViD-19 and Artificial Intelligence: Genome sequencing, drug development and vaccine discovery.,” *Infect. Public Health* 2022, 289–296.
- [30] Norah Alballa , Isra Al-Turaiki, “Machine learning approaches in CoViD-19 diagnosis, mortality, and severity risk prediction: A review,” *Informatics in Medicine Unlocked*, tập 24, 2021.
- [31] Patrick Schwab, August DuMont Schütte, Benedikt Dietz và Stefan Bauer, “Clinical Predictive Models for CoViD-19: Systematic Study,” *Journal of Medical Internet Research*, tập 22, số 10, p. ID. e21439, 2020.
- [32] Maryam AlJame, Imtiaz Ahmad , Ayyub Imtiaz, Ameer Mohammed, “Ensemble learning model for diagnosing CoViD-19 from routine blood tests.,” *Elsivier*, tập Informatics in Medicine Unlocked, 2020.
- [33] Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al., “Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19,” *JAMA Internal Medicine*, 2020.
- [34] Levy TJ, Richardson S, Coppa K, Barnaby DP, McGinn T, Becker LB, Davidson KW, Cohen SL, Hirsch JS, Zanos T., “Development and validation of a survival calculator for hospitalized patients with covid-19,” *medRxiv*, 2020.
- [35] Han Y, Zhang H, Mu S, Wei W, Jin C, Xue Y, Tong C, Zha Y, Song Z, Gu G. , “Lactate dehydrogenase, a risk factor of severe covid-19 patients,” *medRxiv*, 2020.
- [36] M. K. P. a. J. B. Van der Laan, “A new partitioning around medoids algorithm,” *Journal of Statistical Computation and Simulation* , Các tập %1 của %273, no. 8, pp. 575-584., 2003.
- [37] M. G. M. C. a. M. R.-A. Rodríguez-Casado, “ A priori groups based on Bhattacharyya distance and partitioning around medoids algorithm (PAM) with applications to metagenomics.,” *IOSR Journal of Mathematics*, pp. 24-32, (2017).

- [38] R. T. a. J. H. Ng, “ CLARANS: A method for clustering objects for spatial data mining.,” *IEEE transactions on knowledge and data engineering* , Các tập %1 của %214, no. 5, pp. 1003-1016, 2002.
- [39] Tianqi Chen và Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” trong *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, United States, 2016.
- [40] Tianqi Chen and Carlos Guestrin , “XGBoost: A Scalable Tree Boosting System,” *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2016.
- [41] G. e. a. Ke, “LightGBM: A highly efficient gradient boosting decision tree.,” *Adv. Neural Inf. Process. Syst. 2017-Decem*, p. 3147–3155, 2017.
- [42] Liudmila Prokhorenkova, et all, “CatBoost: unbiased boosting with categorical features,” *32nd Conference on Neural Information Processing Systems, NeurIPS*, 2018.
- [43] M. &. S. M. Meharunnisa, “CatBoost Encoded Tree-Based Model for the Identification of Microbes at Genes Level in 16S rRNA Sequence.,” *Communication and Intelligent Systems*, pp. 1137-1156, 2022.
- [44] P. B. Robson, “MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors.,” *Briefings in Bioinformatics*, p. 1–10, 2022.
- [45] P. Rousseeuw, “ Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Computational and Applied Mathematics*, p. 20: 53–65, 1987.
- [46] L. Breiman, “Random Forests,” *Statistics Department University of California Berkeley*, tập CA 94720, 2001.
- [47] Dương Thị Kim Chi, Trần Văn Lăng, “Mô hình dự đoán gen tương quan với hệ thống vật chủ dùng trong Tái tổ hợp,” *Kỹ yếu Hội nghị quốc gia lần thứ X Nghiên cứu cơ bản và ứng dụng CNTT*, pp. 407-415, 2017.
- [48] H.-H. Huang, “An ensemble distance measure of K-mer and Natural Vector for the phylogenetic analysis of multiple-segmented viruses,” *Journal of Theoretical Biology*, p. 136–144, 2016.
- [49] K. &. D. A. K. Acharya, “ Traditional and Ethno-medicinal knowledge of mushrooms in west Bengal.,” *India. Asian Journal of Pharmaceutical and Clinical Research*, pp. 35-41., 2014.

- [50] A. & P. S. Venkatachalapathi, “Exploration of wild medicinal mushroom species in Walayar valley, the Southern Western Ghats of Coimbatore District Tamil Nadu.,” *Mycosphere*, tập 7(2), pp. 118-130, 2016.
- [51] D. C. N. A. L. P. M. K. B. & D. M. D. Mossebo, “). *Termitomyces striatus* f. *pileatus* f. nov. and f. *brunneus* f. nov. From Cameroon with a key to central African species,” *Mycotaxon*, tập 107(1), pp. .315- 329., 2009.
- [52] R. A. B. S. C. J. S. F. Roe AD, “Multilocus species identification and fungal DNA barcoding: insights from blue stain fungal symbionts of the mountain pine beetle,” *Molecular Ecology Resources*, 2010.
- [53] K. S. P. J. H. N. R. O. O. Somervuo P, “ Unbiased probabilistic taxonomic classification for DNA barcoding.,” *Bioinformatics*, p. 32(19):2920–7, 2016.
- [54] N. R. A. K. T. L. T. A. B. M. B. S. B. T. B.-P. J. C. T. e. a. Kõljalg U, “Towards a unified paradigm for sequence-based identification of fung,” *Mol Ecol*, pp. 5271–7., 22(21), 2013.
- [55] W. S. R. T. H. J. H. M. H. E. L. R. O. B. P. D. R. C. e. a. Schloss PD, “Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities,” *Appl Environ Microbiol.*, 2009.
- [56] R. S. B. J. Z. M. A. J. Delgado-Serrano L, “Mycofier: a new machine learning-based classifier for fungal ITS sequences,” *BMC Res Notes.*, p. 402., 2016.
- [57] W. Q. G. P. C. M. P.-A. A. K. C. C. J. M. D. T.-D. N. .. Deshpande V, “Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences,” *Mycologia*, tập 108(1), số 1, 2016.
- [58] R. C. Edgar, “SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences.,” *biRxiv*, tập 074161., 2016.
- [59] T. K. S. S. G. R. T. a. A. R. R. Prabina Kumar Meher, “funbarRF: DNA barcode-based fungal species prediction using multiclass Random Forest supervised learning mode,” *BMC Genetics*, 2019.
- [60] R. R. A. & M. D. C. Das, “CNN_FunBar: Advanced Learning Technique for Fungi ITS Region Classification.,” *Genes*, tập 14(3), p. 634., 2023.
- [61] 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 môi,” *Kỹ yếu Hội nghị quốc gia lần thứ X1 Nghiên cứu cơ bản và ứng dụng CNTT*, pp. 846-853, 2018.
- [62] D. S. D. D. S. S. a. A. C. d. C. Robson P. Bonidia, “MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors,” *Briefings in Bioinformatics.*, p. 1–10, 2022.

- [63] A. F. S. B. S. H. P. a. H. F. Klein, “Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets.,” *Neural Netw*, pp. 106, 294–302, 2016.
- [64] G. V. A. G. V. M. B. T. O. G. M. B. A. M. J. N. G. L. P. P. R. W. V. D. Fabian Pedregosa, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research* 12, pp. 2825-2830, 2011.
- [65] Pablo Sieber, Domenica Flury, Sabine Güsewell, Werner C. Albrich, Katia Boggian,, Céline Gardiol,Matthias Schlegel1, Robert Sieber, Pietro Vernazza1 and Philipp Kohler, “Characteristics of patients with Coronavirus Disease 2019 (CoViD-19) and seasonal influenza at time of hospital admission: a single center comparative study,” *BMC Infectious Diseases*, 2021.
- [66] 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,” *STAIS-2022*, pp. 245-253, 2022.
- [67] Davide Brinati, Andrea Campagner, Davide Ferrari, Massimo Locatelli, Giuseppe Banfi and Federico Cabitza, “Detection of CoViD-19Infection from Routine Blood Exams with Machine Learning: A Feasibility Study,” *Medical Systems*, tập 44, số no. 8,, 2020.
- [68] Banerjee A, Ray S, Vorselaars B, Kitson J, Mamalakis M, Weeks S, Mackenzie LS., “Use of machine learning and artificial intelligence to predict sars-cov-2 infection from full blood counts in a population.,” *Int Immunopharm*, 2020.
- [69] Bao FS, He Y, Liu J, Chen Y, Li Q, Zhang CR, Han L, Zhu B, Ge Y, Chen S, et al., “Triaging moderate covid-19 and other viral pneumonias from routine blood tests,” *arXiv*, 2020.
- [70] Wei Tse Li, Jiayan Ma, Neil Shende, Grant Castaneda, Jaideep Chakladar, Joseph C. Tsai, Lauren Apostol, Christine O. Honda, Jingyue Xu, Lindsay M. Wong, Tianyi Zhang, Abby Lee, Aditi Gnanasekar, Thomas K. Honda, Selena Z. Kuo, Michael Andrew Yu,..., “Using machine learning of clinical data to diagnose CoViD-19: a systematic review and meta-analysis,” *BMC Medical Informatics and Decision Making*, tập 20, 2020.
- [71] Ben Hu, Hua Guo, Peng Zhou and Zheng-Li Shi, “, "Characteristics of SARS-CoV-2 and CoViD-19," , vol. p. , 2021.,” *Nature Reviews Microbiology*, Các tập %1 của %219,, p. 141–154, 2021.
- [72] Forrest Sheng Bao; Youbiao He; Jie Liu; Yuanfang Chen; Qian Li; Christina R. Zhang; Lei Han; Baoli Zhu; Yaorong Ge; Shi Chen; Ming Xu; Liu Ouyang, “Triaging moderate CoViD-19 and other viral pneumonias from routine blood tests,” *arXiv*, tập 2005.06546, p. <https://doi.org/10.48550/arXiv.2005.06546>, 2020.
- [73] Abhirup Banerjee, Surajit Ray, Bart Vorselaars, Joanne Kitson, Michail Mamalakis, Simonne Weeks, Mark Baker and Louise S. Mackenzie, “Use of Machine Learning and

- Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population,” *International immunopharmacology*, tập 86, 2020.
- [74] Jacob Cohen, Patricia Cohen, Stephen G. West and Leona S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd, Biên tập viên, New York, 2002.
 - [75] ShahlaFaisal and Gerhard Tutz, “Nearest neighbor imputation for categorical data by weighting of attributes,” *Information Sciences*, tập 592, pp. 306-319, 2022.
 - [76] Nitesh V. Chawla, et all, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, tập 16, p. 321–357, 2002.
 - [77] Tame Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago và Oteng Tabona, “A survey on missing data in machine Learning,” *Journal of Big Data* , tập 8, p. ID. 140, 2021.
 - [78] A. Bilogur, “Missingno: a missing data visualization suite,” *Journal of Open Source Software*, tập 3, 2018.
 - [79] Dehua Wang, Yang Zhang and Yi Zhao,, “LightGBM: An Effective miRNA Classifica,” *ICCB 2017: Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, Univ. of Nebraska at Omaha*, 2017.
 - [80] Xianlong Zhou, Zhichao Wang, Shaoping Li, Tanghai Liu, Xiaolin Wang, Jian Xia and Yan Zhao, “Machine Learning-Based Decision Model to Distinguish Between CoViD-19 and Influenza: A Retrospective, Two-Centered, Diagnostic Study,” *Risk Manag Healthc Policy*, tập 14, p. 595–604, 2021.
 - [81] Aurelle Tchagna Kouanou ,Thomas Mih Attia,Cyrille Feudjio, Anges Fleurio Djeumo, Adèle Ngo Mouelas,Mendel Patrice Nzogang, Christian Tchito Tchapga, and Daniel Tchiotso, “An Overview of Supervised Machine Learning Methods and Data Analysis for CoViD-19 Detection,” *Healthcare Engineering*, số Hindawi, 2021.
 - [82] Davide Brinati; Andrea Campagner; Davide Ferrari; Massimo Locatelli; Giuseppe Banfi; Federico Cabitza, “Detection of CoViD-19Infection from Routine Blood Exams with Machine Learning: A Feasibility Study,” *Journal of Medical Systems*, tập 44, số 8, p. ID.135, 2020.
 - [83] Krishnaraj Chadaga, Chinmay Chakraborty, Srikanth Prabhu, Shashikiran Umakanth, “Clinical and Laboratory Approach to Diagnose CoViD-19 Using Machine Learning,” *Interdisciplinary Sciences: Computational Life Sciences*, tập 14, p. 452–470, 2022.
 - [84] A. D. S. B. D. a. S. B. Patrick Schwab, “Clinical Predictive Models for CoViD-19: Systematic Study," *Journal of*, vol. 22, no. 10, p. ID. e21439, 2020.,” *Medical Internet Research*, tập 22, p. 10, 2020.

- [85] Wei Tse Li, et al, “Using machine learning of clinical data to diagnose CoViD-19: a systematic review and meta-analysis,” *BMC Medical Informatics and Decision Making*, tập 20, 2020.
- [86] Wei Tse Li, Jiayan Ma, Neil Shende, Grant Castaneda, Jaideep Chakladar, Joseph C. Tsai, Lauren Apostol, Christine O. Honda, Jingyue Xu, Lindsay M. Wong, Tianyi Zhang, Abby Lee, Aditi Gnanasekar, Thomas K. Honda, Selena Z. Kuo, Michael Andrew Yu, Eric Y. Chang, Mahadevan, Rajasekaran và Weg M. Ongkeko, “Using machine learning of clinical data to diagnose CoViD-19: a systematic review and meta-analysis,” *BMC Medical Informatics and Decision Making*, tập 20, p. ID. 247, 2020.