MÔ HÌNH CHẨN ĐOÁN BỆNH COVID-19 TỪ DỮ LIỆU LÂM SÀNG BẰNG PHƯƠNG PHÁP XGBOOST

Dương Thị Kim Chi[1,2]

(1) Trường Đại học Lạc Hồng; (2) Trường Đại học Thủ Dầu Một

Ngày nhận bài 03/4/2023; Ngày gửi phản biện 5/4/2023; Chấp nhận đăng 4/5/2023

Liên hệ email: chidtk@tdmu.edu.vn

https://doi.org/10.37550/tdmu.VJS/2023.03.435

Tóm tắt

Dữ liệu lâm sàng là các kết quả từ xét nghiệm công thức máu, xét nghiệm nước tiểu, đây cũng là phương thức y tế được thực hiện rất phổ biến trong quá trình thăm khám, điều trị và theo dõi bệnh tật. Đối với bác sĩ trực tiếp điều trị, kết quả xét nghiệm cận lâm sàng được xem là cách thức hỗ trợ đắc lực, nhất là trong các bệnh cảnh triệu chứng cơ năng, triệu chứng của bệnh nhân không rõ ràng hoặc không đặc hiệu. Hiện nay bệnh COVID-19 cũng là một dạng bệnh không triệu chứng hoặc triệu chứng không rõ ràng dễ gây nhầm lẫm với các cúm hay số xuất huyết. Sử dụng phương pháp học máy hiện đại để hỗ trợ cho quá trình chẩn đoán sàn lọc bệnh truyền nhiễm từ các mẫu dữ liệu lâm sàng sẽ giúp việc xác định bệnh nhanh chóng, chính xác có thể áp dụng đồng thời cho số lượng mẫu lớn. Điều này đã làm cho quá trình sàng lọc bệnh diễn ra nhanh, chính xác và tiết kiệm kinh phí điều trị. Nghiên cứu này đề xuất mô hình tự động xử lý dữ liệu lâm sàn và kết hợp mô hình phân loại Gradient Boosting để dự đoán bệnh COVID-19, mô hình đề xuất có thể học trực tiếp từ dữ liệu thô là kết quả của xét nghiệm lâm sàn mà không cần phải xóa bỏ dữ liệu trống. Mô hình đề xuất từ nghiên cứu này bao gồm hai giai đoạn: giai đoạn đầu sẽ đánh giá, xử lý dữ liệu; giai đoạn hai sẽ xây dựng mô hình phân loại bệnh dựa trên phương pháp XGBoost (Extreme Gradient Boosting). Để xây dựng mô hình thành công, nghiên cứu được thực hiện dựa trên bộ dữ liệu từ bệnh viện Israelita Albert Einstein ở Brazil đây là bộ dữ liệu do Teich tổng hợp từ các các bệnh nhân nhập viện tháng 4 đến tháng 5 năm 2020 và được xuất bản công khai trên tạp chí einstein_journal. Các kết quả từ nghiên cứu này cho thấy việc kết hợp kỹ thuật xử lý dữ liệu tự động và mô hình XGBoost tạo ra bộ phân loại bệnh COVID-19 từ dữ liệu lâm sàng có kết quả tốt và hiệu suất thu được từ mô hình là vượt trội hơn so với các nghiên cứu cùng chủ đề trên cùng bô dữ liệu, với chính xác tổng thể đạt trên 0,998. Để khẳng định tính chính xác cũng như hiệu năng của mô hình đề xuất đã tiến hành so sánh với nghiên cứu của các tác giả khác cho cùng chức năng dự đoán, nhận thấy mô hình cho kết quả tốt hơn về độ chính xác độ nhạy Recall, Độ đặc hiệu (Specificity), F1 score, ROC, Các kết quả đều đạt ở mức

là 0,99. Trong tương lai, mô hình từ nghiên cứu này sẽ giúp cho việc chẩn đoán bệnh của bệnh nhân trở đơn giản và chính xác. Đồng thời nó sẽ giúp hệ thống y tế tự động chẩn đoán bệnh mang lại nhiều cơ hội chữa bệnh kịp thời cho bệnh nhân và hỗ trợ ngăn chặn bùng phát dịch bệnh.

Từ khóa: *COVID-19, diagnostic model, machine learning, XGBoost*

Abstract

DIAGNOSTIC MODEL OF COVID-19 DISEASE FROM CLINICAL DATA BASED ON XGBOOST METHOD

Clinical data are results from blood count tests, urinalysis, which is also a medical procedure that is very commonly performed during examination, treatment and disease monitoring. For doctors directly treating, the results of subclinical tests are considered an effective way to support, especially in the case of functional symptoms, the patient's symptoms are unclear or non-specific. Currently, COVID-19 disease is also an asymptomatic disease or with unclear symptoms that can easily be confused with influenza or hemorrhagic numbers. Using modern machine learning methods to support the screening process of infectious diseases from clinical data samples will help to quickly and accurately identify diseases that can be applied simultaneously to a large number of samples. This has made the disease screening process fast, accurate and cost-effective. This study proposes an automatic model of clinical data processing and combines the Gradient Boosting classification model to predict COVID-19 disease, the proposed model can learn directly from the raw data as a result of the test. clinical trials without deleting blank data. The proposed model from this study includes two phases: the first phase will evaluate and process data; Phase two will build a disease classification model based on XGBoost (Extreme Gradient Boosting) method. To build a successful model, the study was carried out based on a dataset from the Israelita Albert Einstein hospital in Brazil, which is a dataset compiled by Teich from patients hospitalized April to May 2020 and published publicly in the journal einstein_journal. The results from this study show that combining the automated data processing technique and the XGBoost model to generate a COVID-19 disease classifier from clinical data has good results and performance obtained from the model. is superior to studies on the same topic on the same dataset, with overall accuracy above 0.998. To confirm the accuracy and performance of the proposed model, we compared it with other authors' studies for the same predictive function, and found that the model gave better results in terms of accuracy and sensitivity. Recall, Specificity, F1 score, ROC, Results were all at 0.99. In the future, the model from this study will help make the patient's diagnosis simple and accurate. At the same time, it will help the medical system to automatically diagnose diseases, bring more opportunities for timely treatment to patients and help prevent disease outbreaks.

1. Giới thiệu

Bệnh COVID-19 do virus SARS-COV-2 được phát hiện lần đầu vào tháng 12 năm 2019 tại Trung Quốc, đến nay đã lây lan đến 223 quốc gia và vùng lãnh thổ, làm tê liệt nền kinh tế toàn cầu. Hiện nay, Virus này tiến hóa thành nhiều biến thể với thời gian rất ngắn nên bệnh nhân mặc dù đã được tiêm vacxin phòng bệnh, nhưng vẫn có nhiều trường hợp nhiễm bệnh. Khi nhiễm bệnh, bệnh nhân thường có hoặc không có biểu hiện của bệnh Do đó, việc xác định chính xác bệnh nhân COVID-19 sẽ rất khó khăn. Ngoài ra, đối với những bệnh đang điều trị tại bệnh viện có thể lây nhiễm thêm bệnh COVID-19 điều này làm giảm hiệu quả của phát đồ điều trị. Ngoài ra việc lây nhiễm chéo giữa các nhóm bệnh khác nhau, hoặc cộng gộp nhiều nhóm bệnh sẽ làm ảnh hưởng không tốt đến sức khỏe của bệnh nhân, và lây lan bệnh cộng. Hiện nay tại hầu hết các quốc gia trên thế giới để chẩn đoán bệnh COVID-19 thường phải xét nghiệm bằng PCR (Polymerase Chain Reaction) để khẳng định mắc bệnh (Ben Hu và cs., 2021). Việc xét nghiệm bằng phương pháp này tốn nhiều nguồn sinh phẩm và mất nhiều thời gian mới xác định được bệnh. Đồng thời, chúng ta chỉ sử dụng phương pháp này khi và chỉ khi nghi ngờ bệnh nhân nhiễm bệnh. Do đó, yêu cầu cần thiết hiện nay phải có phương pháp chẩn đoán tự động từ các mẫu kết quả xét nghiệm về máu hay nước tiểu của bệnh nhân. Nhiều nghiên cứu về dự đoán bệnh COVID-19 dựa trên dữ liệu lâm sàng của bệnh nhân đã mang lại nhiều kết quả khả quan. Việc tổng hợp các nghiên cứu cùng lĩnh vực cũng như hiệu năng của các công bố được trình bày ở bảng 1.

Bảng 1. *Tổng hợp các nghiên cứu cùng mục tiêu dự đoán bệnh nhân COVID-19 bằng phương pháp học máy*

Ref.	Nguồn dữ liệu	Kích cỡ DL (COVID-19)	Số thuộc tính chọn	Phương pháp học máy	Độ chính xác
[2]	Albert Einstein Hospital, Brazil	5564/**(553)**	41features	Decision Tree (DT); Extremely Randomized Trees (ET);K-nearest neighbors (KNN); Logistic Regression (LR); Naıve Bayes (NB); **Random Forest (RF);**	0.98
[3]	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**); Naıve Bayes (NB); Random Forest (RF);	0.8
[4]	Albert Einstein Hospital, Brazil	5564/**(553)**	19 Blood features	**XGBoost.**	0,994
[5]	Albert Einstein Hospital, Brazil	5564/**(553)**	14 Blood features	**RF,** LR, GLMNET, ANN.	0,86
[6]	Hospitals in Wuhan, China	294 (**208**)	15 Blood features are available from the clinical data.	RF, **SVM**, ANN.	0,84
[7]	Albert Einstein Hospital, Brazil	5564/**(553)**	14 Blood features	**RF,** LR, XGBoost, KNN, SVM.	0,92

Dựa trên cơ sở dữ liệu dữ liệu của bệnh viện Israelita Albert Einstein ở Brazil (Kaggle, 2020) gồm 5644 ca nhập viện có triệu chứng giống bệnh Covid-19 từ tháng 4 đến tháng 5 năm 2020. Thực hiện xét nghiệm PCR đã xác định 553 trường hợp mắc bệnh COVID-19. Theo đó, tập dữ liệu chứa đựng 41 đặc điểm từ các thuộc tính của dữ liệu lâm sàng về các xét nghiệm mẫu máu, nước tiểu. Trong quy trình tính toán được đề xuất từ nghiên cứu này, bài viết thực hiện dựa trên 4 giai đoạn chính: (i) Khảo sát và xử lý các thuộc tính chứa dữ liệu bị thiếu ; (ii) phân tích tương quan của các thuộc tính và loại bỏ các dữ liệu nhiễu. (iii) Sử dụng kỹ thuật KNNImputer và SMOTE để xử lý dữ liệu trống và giảm mất cân bằng dữ liệu. (iv) Xây dựng bộ phân loại bệnh dựa trên thuật toán XGBoost. Các phần còn lại bài báo này được sắp xếp như sau: Phần II mô tả dữ liệu và phương pháp được sử dụng trong nghiên cứu này. Trong phần III, chúng tôi trình bày kết quả và phần IV trực quan các thuộc tính quan trọng của mô hình dự đoán. Cuối cùng là kết luận.

2. Phương pháp

2.1. Dữ liệu

2.1.1. Khảo sát dữ liệu

Dữ liệu của bệnh viện Israelita Albert Einstein ở Brazil từ tháng 4 đến tháng 5 năm 2020, gồm 5644 ca nhập viện trong đó có 553 trường hợp mắc bệnh COVID-19 (Kaggle, 2020; Vanessa và cs., 2020) bao gồm 110 thuộc tính mô tả về các ca bệnh. Theo các nghiên cứu lâm sàng về bệnh COVID-19 (Ben Hu và cs., 2021; Abhirup và cs., 2020; Forrest Sheng Bao và cs., 2020) các thuộc tính chứa dữ liệu lâm sàng về mẫu máu và nước tiểu có thể khẳng định khả năng mắc bệnh COVID-19. Dựa trên công bố (Ben Hu và cs., 2021; Davide Brinati và cs., 2020; Krishnaraj Chadaga và cs., 2022; Kaggle, 2020) của bài báo đã trích xuất 41 thuộc tính lâm sàn và chi tiết số lượng mẫu cho mô hình dự đoán bệnh COVID-19, thông tin chi tiết về các đặc tính lâm sàng được mô tả ở bảng 2.

Bảng 2. *Thông tin chi tiết các thuộc tính lâm sàng chuẩn đoán bệnh COVID-19 từ bệnh viện Israelita Albert Einstein*

STT	*Đặc tính lâm sàng*	*Số lượng mẫu*	*Kiểu dữ liệu*	*Ý nghĩa*
1	***SARS-Cov-2 exam result***	***644***	***object***	***Kết quả chuẩn đoán bệnh***
2	***Patient age quantile***	***644***	***int64***	***Tuổi bệnh nhân***
3	***Creatinine***	***424***	***float64***	***Chỉ số đào thải creatin phosphat ở cơ***
4	***Direct Bilirubin***	***182***	***float64***	***Xét nghiệm Bilirubin gián tiếp***
5	***Indirect Bilirubin***	***182***	***float64***	***Xét nghiệm Bilirubin trực tiếp***
6	***Lipase dosage***	***8***	***float64***	***Xét nghiệm hoạt độ lipase trong máu***
7	***Proteina C reativa mg/dL***	***506***	***float64***	***protein C***
8	***Serum Glucose***	***208***	***float64***	***Xét nghiệm glucose huyết tương***

9	***Urea***	***397***	***float64***	***Acid Uric trong máu***
10	***ctO2 (arterial blood gas analysis)***	***27***	***float64***	***Khí máu động mạch***
11	***Fio2 (venous blood gas analysis)***	***1***	***float64***	***Khí máu động mạch***
12	***Hb saturation (arterial blood gases)***	***27***	***float64***	***Khí máu động mạch***
13	***Hb saturation (venous blood gas analysis)***	***136***	***float64***	***Khí máu động mạch***
14	***HCO3 (arterial blood gas analysis)***	***27***	***float64***	***Khí máu động mạch***
15	***HCO3 (venous blood gas analysis)***	***136***	***float64***	***Khí máu động mạch***
16	***pCO2 (arterial blood gas analysis)***	***27***	***float64***	***Khí máu động mạch***
17	***pCO2 (venous blood gas analysis)***	***136***	***float64***	***Khí máu động mạch***
18	***pH (arterial blood gas analysis)***	***27***	***float64***	***Khí máu động mạch***
19	***pH (venous blood gas analysis)***	***136***	***float64***	***Khí máu động mạch***
20	***pO2 (arterial blood gas analysis)***	***27***	***float64***	***Khí máu động mạch***
21	***pO2 (venous blood gas analysis)***	***136***	***float64***	***Khí máu động mạch***
22	***Total CO2 (arterial blood gas analysis)***	***27***	***float64***	***Khí máu động mạch***
23	***Total CO2 (venous blood gas analysis)***	***136***	***float64***	***Khí máu động mạch***
24	***Eosinophils***	***602***	***float64***	***Xét nghiệm công thức máu***
25	***Hematocrit***	***603***	***float64***	***Xét nghiệm công thức máu***
26	***Hemoglobin***	***603***	***float64***	***Xét nghiệm công thức máu***
27	***Leukocytes***	***602***	***float64***	***Xét nghiệm công thức máu***
28	***Lymphocytes***	***602***	***float64***	***Xét nghiệm công thức máu***
29	***Mean corpuscular hemoglobin (MCH)***	***602***	***float64***	***Xét nghiệm công thức máu***
30	***Mean corpuscular hemoglobin concentration***	***602***	***float64***	***Xét nghiệm công thức máu***
31	***Mean corpuscular volume (MCV)***	***602***	***float64***	***Xét nghiệm công thức máu***
32	***Mean platelet volume***	***599***	***float64***	***Xét nghiệm công thức máu***
33	***Metamyelocytes***	***97***	***float64***	***Xét nghiệm công thức máu***
34	***Monocytes***	***601***	***float64***	***Xét nghiệm công thức máu***
35	***Myeloblasts***	***97***	***float64***	***Xét nghiệm công thức máu***
36	***Myelocytes***	***97***	***float64***	***Xét nghiệm công thức máu***
37	***Neutrophils***	***513***	***float64***	***Xét nghiệm công thức máu***
38	***Platelets***	***602***	***float64***	***Xét nghiệm công thức máu***
39	***Red blood cell distribution width (RDW)***	***602***	***float64***	***Xét nghiệm công thức máu***
40	***Red blood Cells***	***602***	***float64***	***Xét nghiệm công thức máu***
41	***Segmented***	***97***	***float64***	***Xét nghiệm công thức máu***

Các mẫu trong 41 tham số sử dụng cho mô hình chẩn đoán bệnh Covid-19 bao gồm*:* *Công thức mẫu máu* là 17 thuộc tính; *Khí máu động mạch* gồm 14 thuộc tính những tham số như: pH máu, áp suất khí carbonic trong máu động mạch (PaCO2), nồng độ bicarbonat trong huyết tương HCO3-, độ bão hòa oxy trong máu động mạch (SaO2) và áp suất riêng phần của oxy hòa tan trong máu động mạch (PaO-2)... *Các thuộc tính còn lại* là 10 bao

gồm: kết quả chẩn đoán, độ tuổi và các xét nghiệm về: Bilirubin toàn phần, gián tiếp và trực tiếp; Glucose huyết thanh; Liều lượng lipaza; urê; D-Dimer; Lactic Dehydrogenase; C-Protein phản ứng (CRP); Creatinin; và Thời gian thromboplastin từng phần (PTT) và Thời gian prothrombin Hoạt động từ đồ thị đông máu...

2.1.2. Xử lý dữ liệu lỗi, các thuộc tính có mức độ liên quan cao trong tập dữ liệu

Để có được dữ liệu tốt nhất cho mô hình phân loại, bài viết đã áp dụng phương pháp loại bỏ thuộc tính có tỷ lệ lỗi (null) hơn 99,9% và loại bỏ 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 (Baovà cs., 2020). Những thuộc tính có độ lỗi cao sẽ ảnh hưởng đến kết quả dự đoán. Để bỏ tất cả các thuộc tính của dữ liệu đầu vào có tỷ lệ lỗi lớn hơn 99,9 %.

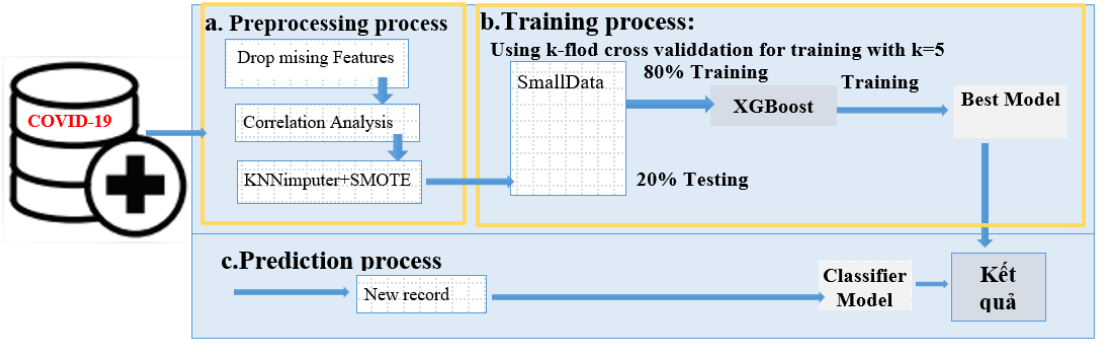
Mối tương quan giữa các thuộc tính là một yếu tố quan trọng trong quá trình đánh giá và chuẩn đoán bệnh cũng như có thể ảnh hưởng trực tiếp đến kết quả của mô hình dự đoán (Vanessavà cs., 2020). Các thuộc tính có mối tương quan cao sẽ không nhiều thông tin hữu ích (hoặc chỉ rất ít), nhưng chúng sẽ làm tăng độ phức tạp của thuật toán (Jacob Cohen và cs., 2002). Để loại bỏ các thuộc tính này, bài báo 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 sẽ 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 |±1|. Ngưỡng được chọn cho việc loại bỏ các thuộc tính tương quan là 0.95 và các tương quan lớn hơn ngưỡng sẽ bị loại bỏ.

2.1.3. Xử lý mất cân bằng dữ liệu của dữ liệu huấn luyện

Như đã phân tích nêu trên dữ liệu y tế dùng trong mô hình dự đoán có mức độ mất cân bằng giữa hai lớp là rất lớn nên hiệu quả của mô hình dự đoán là kém hiệu quả. Chúng tôi đã sử dụng kỹ thuật xử lý dữ liệu hiện đại KNNImputer (ShahlaFaisal và cs., 2022) để bổ sung thêm dữ liệu trống. Và sử dụng kỹ thuật Synthetic Minority Oversampling Technique (SMOTE) (Nitesh và cs., 2002) và làm đầy dữ liệu. Mô hình đề xuất được trình bày như hình 2.

2.2. Mô hình dự đoán bệnh nhân mắc bệnh COVID-19

XGBoost được phát triển dựa trên mô hình gốc Gradient Boosting Machine (GMB) của Friedman XGBoost là phiên bản cải tiến vượt trội của Gradient Boosting với nhiều ưu điểm (Jacob Cohen và cs., 2002) trong đó: nó có thể xử lý tính toán song song giữa các bộ dữ liệu khác nhau nên tốc độ xử lý có thể tăng gấp 10 lần so với GBM. Ngoài ra, nó giảm đáng kể hiện tượng Overfitting bằng cơ chế Regularization và bao gồm cơ chế tự động xử lý missing value bên trong nó. Chính vì những ưu điểm đó mà XGBoost được bài báo áp dụng để xây dựng mô hình dự đoán bệnh Covid-19. Hình 1 là mô hình phân loại tổng quát mà bài viết đề xuất cho giải pháp phân loại.



Hình 1. *Mô hình tổng quát dự đoán bệnh COVID-19*

Trong hình 1, bài viết đã mô tả nội dung thực hiện của đề xuất mô hình dự đoán, chi tiết của các bước như sau ba giai đoạn:

(a) Preprocessing process: Đây là quá trình quan trọng để chuẩn hóa dữ liệu bao gồm; (i)Xử lý trước các thuộc tính của số liệu như: khảo sát các giá trị dữ liệu còn thiếu, (ii) Phân tích tương quan của các thuộc tính và lọc các dữ liệu nhiễu. Sử dụng kỹ thuật KNNImputer và SMOTE để xử lý dữ liệu trống và giảm mất cân bằng dữ liệu**.**

(b) Training process: Kết quả của giai đoạn 1 tập dữ liệu sau khi đã chuẩn hóa được gọi là SmallData và dựa trên bộ dữ liệu này để xây dựng bộ phân loại bệnh. Áp dụng thuật toán XGBoost với các tham số được tối ưu, đã được đánh giá về độ chính xác. Tỷ lệ chia cho quá trình huấn luyện là 80:20, trong đó 80% dữ liệu của SmallData dùng cho việc huấn luyện, 20% dùng cho việc testing.Trong quá trình huấn luyện, bài viết 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 (Tame Emmanuel và cs., 2021). Quá trình này được tiếp diễn tự động và lặp lại năm lần nhằm mục tiêu tối ưu hóa các dữ liệu.

(c) Prediction process Các mẫu thử nghiệm mới có cùng thuộc tính như tập dữ liệu đã rút gọn SmallData sẽ được dùng để dự đoán bởi mô hình tốt nhất đã đề xuất.

2.3. Đánh giá hiệu năng mô hình

Việc áp dụng nhiều mô hình học máy khác nhau cho cùng một bộ dữ liệu nhằm tìm ra các giải pháp tối ưu cho quyết định phân loại và sàng lọc bệnh nhân trong quá trình chẩn đoán bệnh COVID-19. Trong y học lâm sàng, các chẩn đoán "Thật" và "Giả", phát biểu "Mô hình phân loại sai, mô hình chẩn đoán lầm, loại trừ nhầm, một số khái niệm khi áp dụng cho mục tiêu chẩn đoán: TP có thể dịch là "chẩn đoán đúng" true positive (TP), TN là "loại trừ đúng" true negative (TN), FP là "chẩn đoán sai" false positive (FP), FN là "bỏ sót, loại trừ nhầm" and false negative (FN), chúng được làm điều kiện dừng cho quá trình huấn luyện dữ liệu ban đầu. Để đánh giá hiệu suất của mô hình đề xuất, bài viết đã sử dụng các phương pháp đo đạc độ chính xác của mô hình học máy (Kaggle, 2020; Nitesh và cs., 2002) bao gồm:

Accuracy cho biết tỷ lệ các trường hợp được dự đoán đúng (TP và TN) trong toàn bộ dự đoán của tập dữ liệu, đươc tính theo công thức

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

Sensitivity (độ nhạy) - còn được gọi là Recall (PR), hay hit rate (tỉ lệ trúng đích), Độ nhạy True Positive rate (TPR) : tỉ lệ phân loại Positive đúng trên tổng số các trường hợp Positive và Recall

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

Precision: Do bộ dữ liệu đang sử dụng có số lượng mẫu bệnh nhân không nhiễm bệnh COVID-19 nhiều hơn bệnh nhân COVID-19, dẫn đến sự mất tại cân bằng (imbalance) trong tập dữ liệu đầu vào cho mô hình dự đoán. Nên chúng tôi sử dụng Precision để xác định tỉ lệ thực sự positive trên tổng số các trường hợp được mô hình dán nhãn "Positive". Precision là một thuật ngữ đo lường tính "xác định", hay khả năng phân loại Positive chính xác của mô hình.

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

F1 score: Được định nghĩa như trung bình điều hòa (harmonic mean) giữa Precision và Recall

$$F1 = 2\ \mathrm{x}\ \frac{Precision\ \mathrm{x}\ Recall}{Precision +\ Recall}$$

ROC (Receiver Operating Characteristics) được dùng để tính toán hiệu suất phân loại của mô hình. đường cong 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, sau đó tính chỉ số AUC cho đường cong này. Chỉ số AUC chính là con số thể hiện hiệu suất phân loại của mô hình, AUC (Area Under the Curve).

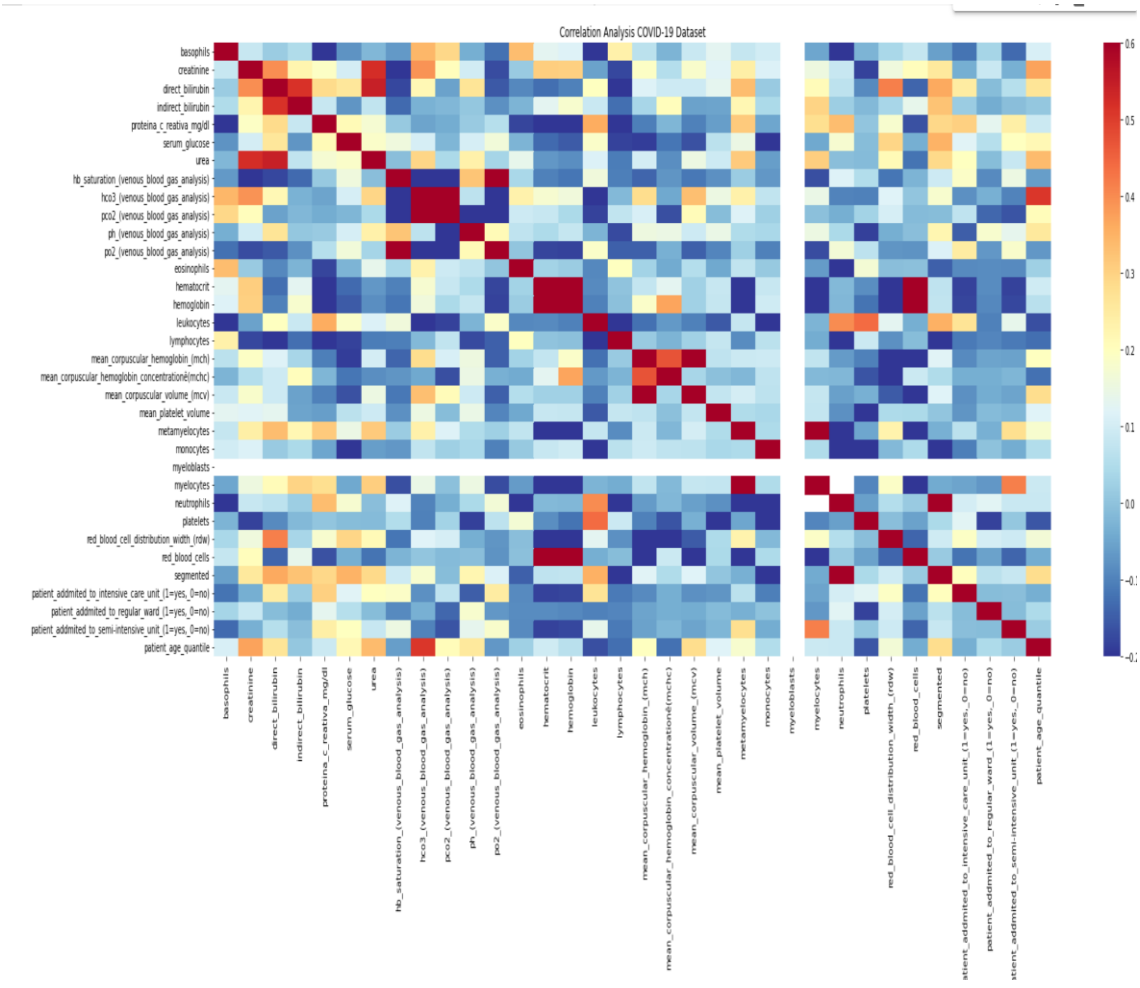
3. Kết quả thực nghiệm

Chúng tôi đã ứng dụng mô hình thực nghiệm bằng phần mềm Python và kết hợp sử dụng các gói phần mềm Sklearn, 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 sẽ được trình bày và được đánh giá so sánh với các nghiên cứu liên quan.

3.1. Xử lý dữ liệu lỗi

Qua quá trình áp dụng giải pháp **(a)** như đề xuất như ở hình 1 đã loại bỏ được 12 thuộc tính có tỷ lệ lỗi lên đến 99% : *'ph (arterial blood gas analysis)', 'po2 (arteria blood gas analysis)', 'pco2_(arteria blood gas analysis)', 'fio2 (venous blood gas analysis)', 'd-dimer', 'total co2 (arterial blood gas_analysis)', 'hco3 (arterial_blood gas analysis)', 'hb_saturation (arterial_blood gases)', 'cto2_(arterial blood gas_analysis)', 'lipase dosage', 'partial thromboplastin timeê(ptt)', 'prothrombin time (pt), activity'*.

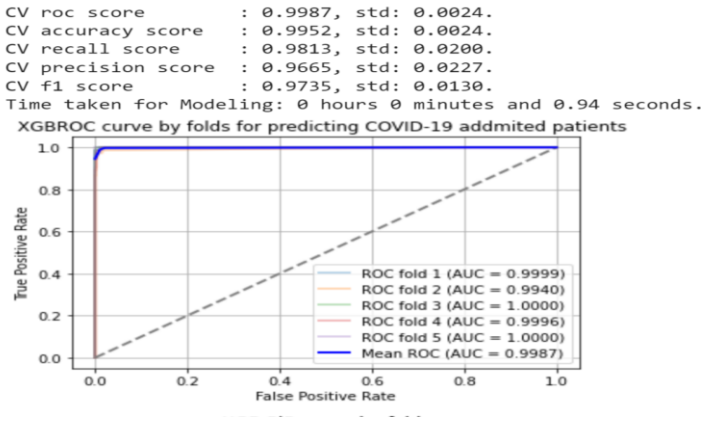
Ngoài ra chúng tôi còn kiểm tra mức độ tương quan của toàn bộ thuộc tính của tập dữ liệu và lựa chọn các thuộc tính có độ tương quan cao trên 95% để xóa, kết quả là không có 1 thuộc tính tương quan đạt mức độ liên quan là *total_co2_(venous blood gas analysis)*, nên không có dữ liệu cần loại bỏ cùng 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 2.



Hình 2. *Mối tương quan giữa các thuộc tính khảo sát*

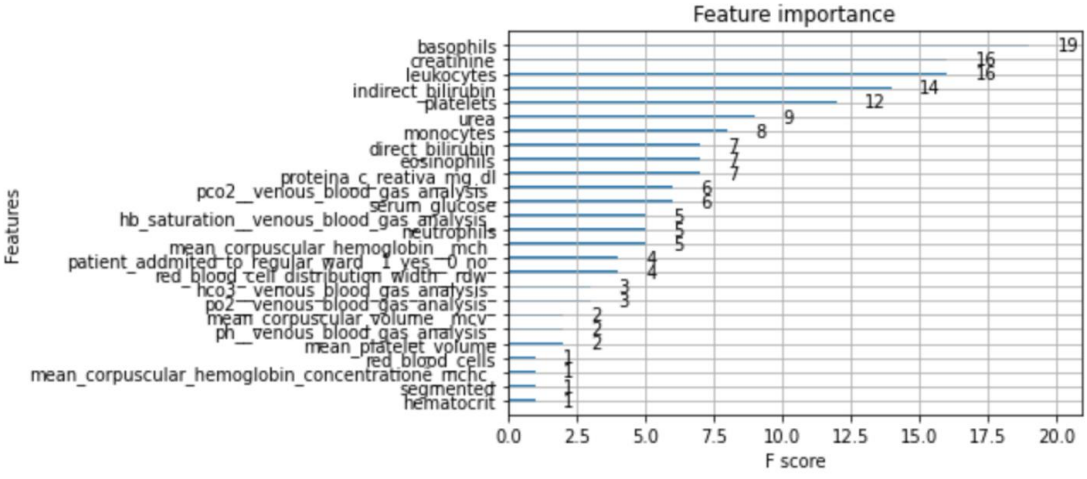
3.2. Kết quả mô hình phân loại từ dữ liệu lâm sàng

Áp dụng giải pháp **(b)** như mô tả ở hình 1: Tập dữ liệu sau khi loại bỏ dữ liệu lỗi được gọi là SmallData 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. Khi xây dựng mô hình dự đoán, chúng tôi sử dụng kỹ thuật kiểm tra chéo 5-folds với quy tắc: (i) lần đầu cho ngẫu nhiên số lần lập 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. Thông qua đó, chúng tôi sử dụng nhiều thuật toán khác để xây dựng bộ phân lớp dựa trên Datasmall như: XGBoost (Tianqi Chen và cs., 2016).



Hình 3. *Kết quả hiệu năng của mô hình đề xuất.*

Kết quả áp dụng thuật toán XGBoost với kỹ thuật kiểm tra chéo 5-folds để đảm bảo tính đúng đắn của mô hình thực nghiệm. Ngoài ra Mức độ quan trọng của từng thuộc tính cũng cho biết các thông điệp quý từ mô hình dự đoán. Đối với mô hình đề xuất kết quả dự đoán được trình bày chi tiết như hình



Hình 4. *Mức độ quan trọng chi tiết của các thuộc tính từ của mô hình đề xuất*

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

Các kết quả chi tiết về hiệu năng của mô hình được xây dựng bằng XGBoost được mô tả bởi các hình 5, 6,7. 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 Valter Augusto de Freitas Barbosa (2021). Các độ đo hiệu năng của mô hình đề xuất được liệt kê chi tiết ở bảng 3.

Bảng 3. *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	**XGBoost**	**0.998**	**0.966**	**0.981**	**0.973**
Valter Augusto de Freitas Barbosa (2021)	Random Forest	0.984	0.923	0.936	_

Trong mô tả ở bảng 3 cho thấy kết quả đề xuất đạt hiệu năng cao hơn hẳn về các chỉ số: AUC, Precision (Specificity), Recall (Sensitivity), F1 score. Như vậy cùng tập dữ liệu về bệnh COVID-19 bài viết đã đề xuất quy trình xây dựng mô hình dự đoán bệnh nhân mắc Covid-19 có độ chính xác cao. Trong tương lai có thể áp dụng quy trình này để chuẩn đoán các bệnh nguy hiểm khác.

4. Kết luận

Từ bộ dữ tổng hợp về chẩn đoán bệnh, bài viết đã xây dựng bộ phân loại và dự đoán bệnh COVID-19 từ xét nghiệm lâm sàng. Nghiên cứu này đã chứng minh việc sử dụng máy học thông qua mô hình kết hợp từ phương pháp xử lý dữ liệu tự động bằng kỹ thuật phân loại XGBoost. Thêm vào đó, XGBoost xây dựng mô hình phân loại tối ưu và mang lại hiệu quả đang mong đợi hơn so với nhiều nghiên cứu cùng loại. Mô hình từ bài viết đề xuất đã phát huy tác dụng 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. 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.

Kết quả thực nghiệm từ bài báo đã chứng minh tính hiệu quả của phương pháp này. Trong tương lai, phương pháp được đề xuất từ nghiên cứu này sẽ đóng góp cho việc sàn lọc, phân loại và dự đoán chính xác bệnh nhân mắc bệnh. Mô hình sẽ 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 COVID-19. Trong bối cảnh dịch bệnh COVID-19 đang trở nên phức tạp trên khắp thế giới thì biện pháp mở ra giải pháp tối ưu, an toàn và chính xác trong quá trình phân loại để điều trị bệnh nhân.

TÀI LIỆU THAM KHẢO

[1] Bilogur (2018). Missingno: a missing data visualization suite. *Journal of Open Source Software*, vol. 3.

[2] Abhirup Banerjee, Surajit Ray, Bart Vorselaars, Joanne Kitson, Michail Mamalakis, Simonne Weeks, Mark Baker and Louise S. Mackenzie (2020). Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. *International immunopharmacology*, vol. 86.

[3] Banerjee A, Ray S, Vorselaars B, Kitson J, Mamalakis M, Weeks S, Mackenzie LS. (2020). Use of machine learning and artificial intelligence to predict sars-cov-2 infection from full blood counts in a population. *Int Immunopharm*.

[4] Bao FS, He Y, Liu J, Chen Y, Li Q, Zhang CR, Han L, Zhu B, Ge Y, Chen S, et al. (2020). Triaging moderate covid-19 and other viral pneumonias from routine blood tests. *arXiv*.

[5] Ben Hu, Hua Guo, Peng Zhou and Zheng-Li Shi, (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, Vols. 19, p. 141-154.

[6] Davide Brinati, Andrea Campagner, Davide Ferrari, Massimo Locatelli, Giuseppe Banfi and Federico Cabitza (2020). Detection of COVID-19Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *Medical Systems*, *44*(8).

[7] Dehua Wang, Yang Zhang and Yi Zhao (2017). LightGBM: An Effective miRNA Classifica. *ICCBB 2017: Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics*. Univ of Nebraska at Omaha.

[8] Forrest Sheng Bao, Youbiao He, Jie Liu, Yuanfang Chen, Qian Li, Christina R. Zhang, Lei Han, Baoli Zhu, Yaorong Ge, Shi Chen, Ming Xu and Liu Ouyang (2020). Triaging moderate COVID-19 and other viral pneumonias from routine blood tests.

[9] Jacob Cohen, Patricia Cohen, Stephen G. West and Leona S. Aiken (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd, Ed., New York.

[10] K. (2020). Diagnosis of COVID-19 and its clinical spectrum. Kaggle., p. URL: www.kaggle.com/einsteindata4u/covid19.

[11] Kaggle (2020). Diagnosis of COVID-19 and its clinical spectrum. *Computer software manual*. Retrieved April 07.

[12] Krishnaraj Chadaga, Chinmay Chakraborty, Srikanth Prabhu, Shashikiran Umakanth (2022). Clinical and Laboratory Approach to Diagnose COVID 19 Using Machine Learning. *Interdisciplinary Sciences: Computational Life Sciences*, vol. 14, p. 452-470.

[13] Liudmila Prokhorenkova, et all (2018). *CatBoost: unbiased boosting with categorical features*. 32nd Conference on Neural Information Processing Systems, NeurIPS.

[14] Maryam AlJame, Imtiaz Ahmad , Ayyub Imtiaz, Ameer Mohammed (2020). *Ensemble learning model for diagnosing COVID-19 from routine blood tests*. Elsivier, vol. Informatics in Medicine Unlocked.

[15] Nitesh V. Chawla, et all. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, vol. 16, p. 321-357.

[16] ShahlaFaisal and Gerhard Tutz (2022). Nearest neighbor imputation for categorical data by weighting of attributes. *Information Sciences*, vol. 592, pp. 306-319.

[17] Tame Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago and Oteng Tabona (2021). A survey on missing data in machine Learning. *Journal of Big Data*, vol. 8, p. ID. 140.

[18] Tianqi Chen and Carlos Guestrin (2016). 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*.

[19] Valter Augusto de Freitas Barbosaa, Juliana Carneiro Gomesb, Maıra Araujo de Santana Clarisse Lins de at all (2021). Covid-19 rapid test by combining a Random Forest-based web system. *JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS*.

[20] Vanessa Damazio Teich, et all. (2020). Epidemiologic and clinical features of patients with COVID-19 in Brazil. *einstein_journal*.