NGHIÊN CỨU TỔNG QUAN VỀ DỰ BÁO ĐỢT CẤP BỆNH PHỔI TẮC NGHẼN MẠN TÍNH

Bùi Mỹ Hạnh[1,2,✉], Vương Thị Ngân[2], Hoàng Thị Hồng Xuyến[1]

[1]Trường Đại học Y Hà Nội
[2]Bệnh viện Đại học Y Hà Nội

Ứng dụng máy học dự báo đợt cấp bệnh phổi tắc nghẽn mạn tính (Chronic Obstructive Pulmonary Disease - COPD) là một xu thế tất yếu, có khả năng cách mạng hóa việc khám, điều trị, quản lý bệnh bằng cách cho phép phát hiện sớm, can thiệp cá nhân hóa, tối ưu hóa nguồn lực và trao quyền cho người bệnh. Nghiên cứu nhằm mục tiêu tổng quan các mô hình sẵn có về dự báo đợt cấp COPD từ cơ sở dữ liệu Pubmed theo hướng dẫn PRISMA. Kết quả đã xác định được 9/928 bài báo đáp ứng đầy đủ các tiêu chí lựa chọn, bao gồm: 7 quan sát hồi cứu đa trung tâm, 1 quan sát tiến cứu đơn trung tâm và 1 thử nghiệm lâm sàng đơn trung tâm. 117 yếu tố nguy cơ được đưa vào các mô hình dự báo, trong đó tuổi và giới xuất hiện phổ biến nhất (9/9 lần). Các mô hình có diện tích dưới đường cong (AUC) dao động từ 0,681 đến trên 0,9 với 3 mô hình có hiệu suất cao nhất lần lượt là Random Forest (> 0,9), Support Vector Machine (0,9) và Extreme Gradient Boosting (0,86) cần được ứng dụng, tiếp tục xây dựng và phát triển trên bộ dữ liệu của người Việt Nam.

Từ khóa: Đợt cấp bệnh phổi tắc nghẽn mạn tính, mô hình, dự báo, máy học.

I. ĐẶT VẤN ĐỀ

Bệnh phổi tắc nghẽn mạn tính (COPD) là một trong những nguyên nhân hàng đầu gây bệnh tật và tử vong trên toàn thế giới.[1] Đợt cấp COPD là biểu hiện xấu đi cấp tính của các triệu chứng hô hấp dẫn đến phải điều trị bổ sung, gây ảnh hưởng bất lợi đến tình trạng sức khỏe, tỷ lệ nhập viện và tiến triển bệnh.[1,2] Vì vậy, việc giảm nguy cơ bùng phát đợt cấp trong tương lai là mục tiêu chính của quản lý COPD. Các mô hình dự báo mang lại tiềm năng can thiệp sớm và quản lý bệnh chủ động dựa trên từng cá thể. Bằng cách xác định những cá nhân có nguy cơ tiến triển nặng trước khi chúng xảy ra, bác sĩ lâm sàng có thể điều chỉnh chế độ điều trị và thực hiện các biện pháp phòng ngừa như hỗ trợ cai thuốc lá và đưa ra các chương trình phục hồi chức năng phổi dựa trên đặc điểm của từng người bệnh.[3] Nguồn lực y tế hiện nay còn hạn chế và việc ưu tiên những người bệnh có nguy cơ đợt cấp cao hơn có thể giúp tối ưu hóa việc phân bổ thuốc men, máy thở, giường bệnh và nhân lực chăm sóc sức khỏe.[3] Do đó, việc xác định những người bệnh có nguy cơ cao bằng cách sử dụng các yếu tố có thể đo lường được, có tương quan với đợt cấp đóng vai trò quan trọng để giảm gánh nặng bệnh tật, ngăn ngừa tử vong sớm, giảm chi phí chăm sóc sức khỏe cao và nâng cao chất lượng cuộc sống.

Một số mô hình dự báo đợt cấp COPD đã được công bố kết hợp dữ liệu từ tiền sử bệnh, đặc điểm lâm sàng và kết quả xét nghiệm.[4-12] Ở Việt Nam, các nghiên cứu về đợt cấp COPD mới chỉ dừng lại ở việc tìm ra các yếu tố nguy cơ liên quan, mà chưa thiết lập một mô hình dự báo đợt bùng phát cụ thể bằng thuật toán trí tuệ nhân tạo cũng như chưa rõ mô hình nào chính xác nhất và có thể áp dụng rộng rãi.[13,14] Chính

Tác giả liên hệ: Bùi Mỹ Hạnh
Trường Đại học Y Hà Nội
Email: buimyhanh@hmu.edu.vn
Ngày nhận: 23/09/2024
Ngày được chấp nhận: 29/10/2024

vì vậy, chúng tôi tiến hành nghiên cứu với mục tiêu: ***Phân tích một số mô hình sử dụng trí tuệ nhân tạo dự báo đợt cấp bệnh phổi tắc nghẽn mạn tính.***

II. NỘI DUNG TỔNG QUAN

1. Đối tượng và phương pháp

Thiết kế tổng quan

Nghiên cứu tổng quan được thực hiện và báo cáo theo hướng dẫn Báo cáo tổng quan hệ thống và phân tích gộp (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA).

Cơ sở dữ liệu

Nghiên cứu sử dụng cơ sở dữ liệu Pubmed, được phát triển bởi Thư viện Y khoa Quốc gia Hoa Kỳ, Viện Y tế Quốc gia Hoa Kỳ (US National Library of Medicine - National Institutes of Health). Các bài báo được lựa chọn có thời gian xuất bản từ tháng 1 năm 2017 đến tháng 5 năm 2024 trên các tạp chí quốc tế có bình duyệt và được tiến hành trên bất kỳ quốc gia nào.

Chiến lược tìm kiếm

Chúng tôi sẽ tiến hành tìm kiếm toàn diện để xác định tất cả các ấn phẩm liên quan. Chúng tôi tập trung từ khóa vào ba phần chính bao gồm:

(1) Đợt cấp bệnh phổi tắc nghẽn mạn tính: Chronic Obstructive Pulmonary Disease Exacerbation, COPD Exacerbation, Chronic Obstructive Pulmonary Disease Flare-up, COPD Flare-up, Exacerbation of Chronic Obstructive Pulmonary Disease, Exacerbation of COPD, AECOPD;

(2) mô hình: model, tool, software, app, social media, cloud, web, web-database, web based, computer, mobile device, mobile data, data storage, online, network;

(3) dự báo: prediction, forecast, prognostication, estimate. Toán tử "OR" để liên kết tất cả các thuật ngữ và từ đồng nghĩa thành các nhóm cụ thể liên quan đến các từ khóa chính và toán tử "AND" để liên kết tất cả các nhóm thành chuỗi tìm kiếm cuối cùng.

Lựa chọn nghiên cứu

Việc lựa chọn nghiên cứu được thực hiện theo sơ đồ PRISMA. Hai thành viên nhóm nghiên cứu rà soát và sàng lọc dựa trên các tiêu chí lựa chọn, bao gồm:

(1) mô tả về mô hình dự báo hoặc công cụ dự báo cụ thể;

(2) dự báo liên quan đến đợt cấp bệnh phổi tắc nghẽn mạn tính.

Nghiên cứu bị loại trừ nếu có một trong các đặc điểm sau:

(1) Không phải mô hình dự báo đợt cấp bệnh phổi tắc nghẽn mạn tính bằng trí tuệ nhân tạo;

(2) Được công bố bằng ngôn ngữ không phải là tiếng Anh;

(3) Không có toàn văn của nghiên cứu.

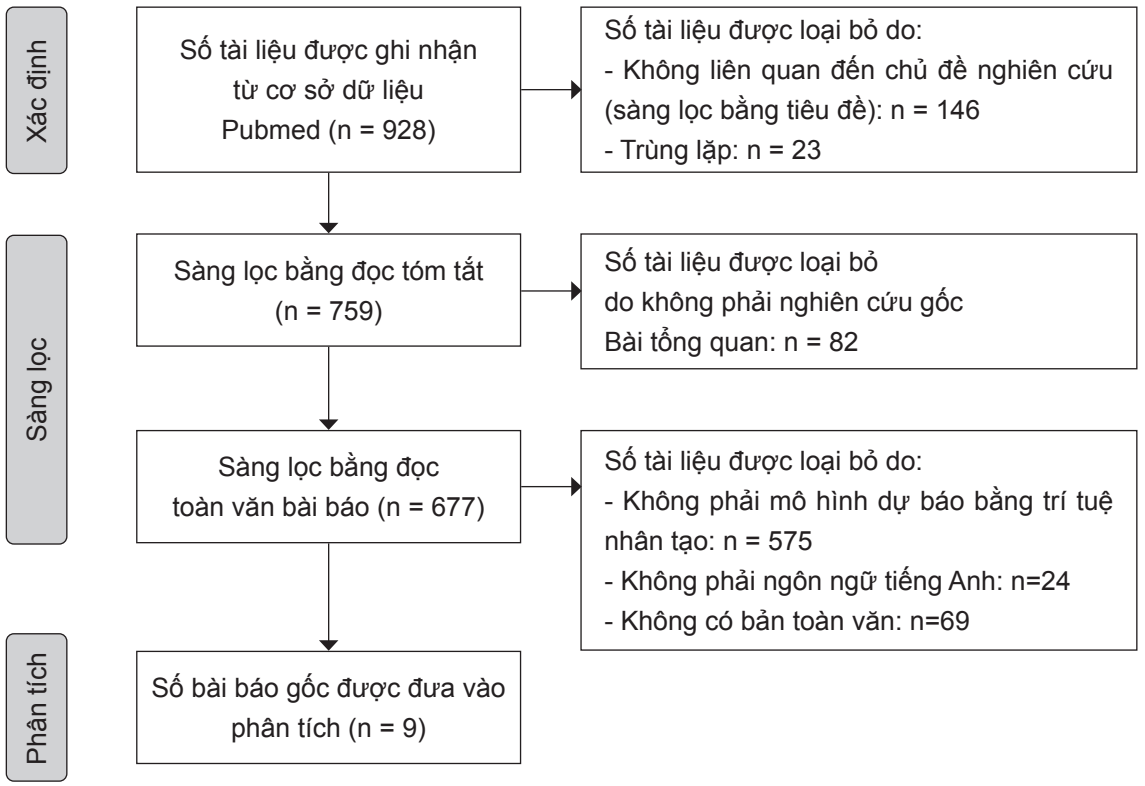
Trước tiên, các nghiên cứu được rà soát dựa trên tiêu đề và tóm tắt; nếu chưa đủ thông tin để ra quyết định lựa chọn hoặc loại trừ, các nghiên cứu được rà soát toàn văn. Để tránh sai số trong quá trình lựa chọn nghiên cứu, hai tác giả sẽ độc lập thực hiện toàn bộ quá trình rà soát và lựa chọn mô hình. Kết quả lựa chọn sẽ được so sánh, và các điểm không đồng nhất (nếu có) được giải quyết thông qua thảo luận để đi đến thống nhất cuối cùng.

Trích xuất dữ liệu

Tất cả dữ liệu được truy xuất từ bài báo toàn văn được nhập vào biểu mẫu thu thập thông tin. Bất kỳ sự khác biệt nào đều được giải quyết bằng cách thảo luận để đạt được sự đồng thuận.

III. KẾT QUẢ

1. Đặc điểm các nghiên cứu



Sơ đồ 1. Biểu đồ PRISMA về quy trình lựa chọn nghiên cứu

Quá trình sàng lọc tài liệu được trình bày ở hình 1. Dựa vào từ khóa được nêu, chúng tôi tìm được 928 tài liệu trên cơ sở dữ liệu Pubmed. Trong 928 tài liệu ban đầu, 146 tài liệu không liên quan đến chủ đề nghiên cứu (sàng lọc bằng tiêu đề), 23 tài liệu trùng lặp, 82 tài liệu không phải nghiên cứu gốc. Kết quả có 677 bài báo được sàng lọc đầy đủ, trong đó chúng tôi loại 575 bài báo không có mô hình dự báo cụ thể, 24 bài báo không viết bằng ngôn ngữ tiếng Anh, 69 bài báo không có bản toàn văn. Tổng cộng có 9 bài báo được đưa vào nghiên cứu này. Về thiết kế, 7/9 nghiên cứu là nghiên cứu quan sát hồi cứu, 1/9 nghiên cứu quan sát tiến cứu và 1/9 nghiên cứu thử nghiệm lâm sàng. Về địa điểm, 7/9 nghiên cứu đa trung tâm, 2/9 nghiên cứu đơn trung tâm. Tổng cỡ mẫu trong các nghiên cứu được thu nhận là 80.324 người bệnh, dao động từ 67 đến 43.576 người bệnh. 117 yếu tố nguy cơ khác nhau được đưa vào các mô hình dự đoán. Tuổi, giới là hai yếu tố dự đoán phổ biến nhất (9 lần). Các yếu tố dự đoán phổ biến tiếp theo là tình trạng hút thuốc (7 lần), điểm CAT (COPD Assessment Test), bệnh đái tháo đường (5 lần), BMI (Body mass index), bệnh tăng huyết áp và các dấu hiệu sinh tồn: mạch, nhiệt độ, huyết áp, nhịp thở, SpO_2 (4 lần), 40 yếu tố xuất hiện 2 hoặc 3 lần. Hơn một nửa số yếu tố dự đoán (65/117) chỉ được đưa vào một lần.

2. Mô hình Nomogram[12]

Mô hình dự báo dựa trên đặc điểm nhân khẩu học và các thông số xét nghiệm máu để dự đoán tần suất các đợt cấp COPD có AUC 0,681; độ nhạy 63,6% và độ đặc hiệu 65,0%.

Bảng 1. Các biến số trong mô hình Nomogram

Biến số đầu vào	- Thông tin chung: tuổi, giới, tình trạng hút thuốc, thời gian nằm viện - Bệnh đồng mắc: đái tháo đường, tăng huyết áp, suy tim, bệnh thận mãn tính, bệnh gan mạn tính. - Xét nghiệm: bạch cầu trung tính (NEU), bạch cầu ái toan (EOS), bạch cầu lympho (LYM), bạch cầu đơn nhân (MONO), tiểu cầu, glucose, glucose-to-lymphocyte ratio (GLR), bilirubin toàn phần (TBil), bilirubin gián tiếp (IBil), bilirubin trực tiếp (DBil), albumin, γ-glutamyl transpeptidase (GGT), apoprotein B, fibrinogen, D-dimer.
Biến số đầu ra	Tuổi, giới, bệnh gan, thời gian nằm viện, NEU, EOS, MONO, DBil, GGT, GLR.

3. Mô hình LASSO (Least Absolute Shrinkage and Selection Operator)[5]

Mô hình dự báo nhằm xác định người bệnh nhập viện vì đợt cấp COPD và nguy cơ tái nhập viện sau 30 ngày dựa trên tập dữ liệu lấy từ bệnh án điện tử. Mô hình LASSO có độ nhạy là 76,6%; độ đặc hiệu là 97,1% và chỉ số C-statistic là 0,975 ± 0,004.

Bảng 2. Các biến số trong mô hình LASSO

Biến số đầu vào	- Thông tin chung: tuổi, giới, thu nhập bình quân, số lần nhập viện năm trước, thời điểm nhập viện. - Bệnh đồng mắc: viêm phổi, cúm, suy tim, thiếu máu, rối loạn trầm cảm. - Dấu hiệu sinh tồn: nhịp thở, nhịp tim, huyết áp, nhiệt độ, SpO_2. - Xét nghiệm: hồng cầu, hemoglobin, bạch cầu, bạch cầu trung tính (NEU), bạch cầu ái toan (EOS), tiểu cầu, glucose, albumin, clorua, creatinin, kali, natri, bicarbonat, khoảng trống anion (AG), peptid natriuretic type B (BNP), INR, troponin, khí máu. - Thuốc: corticosteroid dạng hít (ICS), kháng cholinergic tác dụng ngắn (SAAC), kháng cholinergic tác dụng kéo dài (LAAC), chủ vận beta tác dụng ngắn (SABA), chủ vận beta tác dụng kéo dài (LABA), prednisone, methylprednisolone, kháng virus cúm, kháng sinh, lợi tiểu quai.
Biến số đầu ra	Prednisone, SAAC, LAAC, kháng sinh được chỉ định trong vòng 72 giờ đầu sau khi nhập viện.

4. Mô hình GLM (General Linear Model)[7]

Mô hình dự báo nguy cơ đợt cấp COPD mức độ trung bình và nặng gồm 11 yếu tố có AUC 0,71; giá trị tiên đoán dương (PPV) 48%; giá trị tiên đoán âm (NPV) 80%, đạt hiệu suất dự báo tốt hơn 10% so với mô hình rút gọn 3 yếu tố.

Bảng 3. Các biến số trong mô hình GLM

Biến số đầu vào	- Thông tin chung: tuổi, giới, BMI, chủng tộc, tình trạng hút thuốc, thời gian mắc COPD, số đợt cấp năm trước. - Bệnh đồng mắc: trào ngược dạ dày thực quản, đái tháo đường, tăng huyết áp, rối loạn lipid máu. - Triệu chứng: điểm CAT. - Xét nghiệm: NEU, EOS, FVC, FEV1, FEF25-75, PEF. - Thuốc: ICS, LAMA, LABA, Budesonide, Formoterol fumarate.
Biến số đầu ra	- Mô hình đầy đủ: giới, tình trạng hút thuốc, chủng tộc, số đợt cấp năm trước, điểm CAT, FEV1%, EOS, ICS, LAMA, LABA, Budesonide. - Mô hình rút gọn: số đợt cấp năm trước, EOS, ICS.

5. Mô hình RF (Random Forest)[4]

Nghiên cứu sử dụng dữ liệu người bệnh tự báo cáo cho ứng dụng sức khỏe số (myCOPD) có AUC là 0,727. Mô hình RF cung cấp các cảnh báo, đề xuất hành động dự phòng giảm thiểu nguy cơ xảy ra các đợt trầm trọng trong tương lai, nâng cao tính cá nhân hóa trong quản lý, chăm sóc COPD.

Bảng 4. Các biến số trong mô hình RF

Biến số đầu vào	Tuổi, giới, tình trạng hút thuốc, điểm CAT Điểm triệu chứng: (1) Các triệu chứng bình thường đối với người bệnh và dùng thuốc theo đơn, (2) Khó thở hơn bình thường nhưng không sốt/ thay đổi màu sắc/ lượng đờm, người bệnh có thể tự xịt thuốc giãn phế quản, (3) Khó thở hơn bình thường, ho ra đờm, thay đổi màu sắc đờm, cần tự dùng thuốc steroid và/hoặc kháng sinh (4) Khó thở nhiều hơn so với bình thường mặc dù đã được điều trị, hoặc bị đau ngực và/hoặc sốt cao cần nhập viện.
Biến số đầu ra	Tuổi, giới, tình trạng hút thuốc, điểm CAT, điểm triệu chứng.

6. Mô hình XGBoost (Extreme Gradient Boosting)[11]

Mô hình dự báo các đợt kịch phát mức độ nặng trong vòng một năm tới ở người bệnh COPD có AUC 0,866; độ chính xác 90,33%; độ nhạy 56,6%; độ đặc hiệu 91,17%; PPV 13,7%; NPV 98,83%.

Bảng 5. Các biến số trong mô hình XGBoost

Biến số đầu vào	117 yếu tố bao gồm: thông tin chung, bệnh đồng mắc, dấu hiệu sinh tồn, triệu chứng lâm sàng, cận lâm sàng, điều trị.
Biến số đầu ra	- Thông tin chung: tuổi, giới, chủng tộc, BMI, tình trạng hút thuốc, số năm mắc COPD. - Bệnh đồng mắc: hen suyễn, ung thư phổi, rối loạn lo âu/trầm cảm, viêm mũi dị ứng, bệnh chàm, suy tim sung huyết, tăng huyết áp, bệnh tim thiếu máu cục bộ, đái tháo đường, trào ngược dạ dày thực quản, viêm xoang, chứng ngưng thở lúc ngủ. - Thuốc: SAMA, SABA, SABA+SAMA, LAMA, LABA, LABA+LAMA, ICS+LABA, ICS+LABA+LAMA, Corticosteroid, Phosphodiesterase-4 inhibitor.

7. Mô hình dự báo dựa vào thông số lâm sàng và dấu ấn sinh học miễn dịch trong TOPDOCS (The Swiss Multicenter COPD Cohort Study)[6]

Nghiên cứu theo dõi các dấu ấn sinh học miễn dịch và đặc điểm lâm sàng trong vòng một năm ở 271 người bệnh COPD từ TOPDOCS dự báo tần suất đợt cấp COPD không thường xuyên (0 hoặc 1) hay thường xuyên (> 1) trong năm tiếp theo cho thấy hiệu suất mô hình sử dụng thông số lâm sàng và dấu ấn sinh học (AUC = 0,78) không vượt trội so với mô hình chỉ sử dụng thông số lâm sàng (AUC = 0,79).

Bảng 6. Các biến số trong mô hình từ TOPDOCS

Biến số đầu vào	- Thông tin chung: tuổi, giới, BMI, tình trạng hút thuốc. - Dấu hiệu sinh tồn: nhịp thở, nhịp tim, huyết áp, nhiệt độ, SpO_2. - Triệu chứng lâm sàng: điểm mMRC, điểm CAT. - Xét nghiệm: FEV1, WBC, PLT, NT- proBNP, CRP, Creatinin, Bilirubin. - Dấu ấn sinh học miễn dịch: IgA, IgM, IgE, IgG, IgG1, IgG2, IgG3, IgG4, Kiểu gen rs8099917GG, Kiểu gen rs8099917TT, Kiểu gen rs8099917TG. - Thuốc: LAMA, LABA, LAMA + LABA, LAMA + ICS, LABA + ICS, LAMA + LABA + ICS, SABA, SAMA, Steroid, Theophylline, Roflumilast.
Biến số đầu ra	- Mô hình đầy đủ (10 yếu tố: 8 thông số lâm sàng + 2 dấu ấn sinh học): tuổi, giới, BMI, nhịp tim, điểm CAT, điểm mMRC, FEV1, số lượng thuốc điều trị, PLT, IgG2. - Mô hình rút gọn: 8 thông số lâm sàng.

8. Mô hình dự báo nguy cơ nhập viện do đợt cấp COPD ở cơ sở chăm sóc ban đầu tại Thụy Điển[8]

Mô hình sử dụng cơ sở dữ liệu quốc gia bao gồm các thông tin về nhân khẩu học, chăm sóc sức khỏe và các yếu tố kinh tế xã hội dự đoán nguy cơ nhập viện do đợt cấp COPD trong vòng 10 ngày tới, có AUC là 0,86.

Bảng 7. Các biến số trong mô hình dự báo nguy cơ nhập viện do đợt cấp COPD ở cơ sở chăm sóc ban đầu tại Thụy Điển

Biến số đầu vào	- Thông tin chung: tuổi, giới, BMI, tình trạng hút thuốc, số đợt cấp trước đây, trình độ học vấn, thu nhập bình quân. - Bệnh đồng mắc: Chỉ số bệnh đi kèm Charlson (CCI): hen suyễn, viêm phổi, bệnh tim thiếu máu cục bộ, đái tháo đường, tăng lipid máu, rối loạn lo âu/trầm cảm, suy giảm nhận thức, trào ngược dạ dày thực quản, loãng xương, ung thư, bệnh thận, bệnh xơ nang. - Xét nghiệm: protein phản ứng C, huyết sắc tố, lipoprotein tỷ trọng cao (HDL), lipoprotein tỷ trọng thấp (LDL), LDL/HDL, HbA1C, WBC, Tổng lượng chất béo, TriGlyceride, FVC, FEV1, FEV1/FVC. - Thuốc: SABA, SAMA, LABA, LAMA, ICS, kháng leukontrien, roflumilast, steroid đường uống, kháng sinh.
Biến số đầu ra	Số đợt cấp trước đây, CCI, bệnh tim thiếu máu cục bộ, steroid đường uống.

9. Mô hình dự báo đợt cấp COPD sử dụng dữ liệu về lối sống[10]

Mô hình giám sát liên tục lối sống, triệu chứng của người bệnh cùng các yếu tố môi trường trong nhà tích hợp trên ứng dụng điện thoại thông minh, thiết bị đeo được và thiết bị cảm biến chất lượng không khí giúp phát hiện sớm đợt cấp COPD trong 7 ngày tới. Mô hình đạt độ chính xác 92,1%, độ nhạy 94% và độ đặc hiệu 90,4%; AUC > 0,9.

Bảng 8. Các biến số trong mô hình dự báo đợt cấp COPD sử dụng dữ liệu về lối sống

Biến số đầu vào	- Lối sống: số bước đi bộ, số bậc leo cầu thang, khoảng cách di chuyển hàng ngày, lượng calo tiêu thụ, chất lượng giấc ngủ. - Triệu chứng: điểm CAT, điểm mMRC. - Môi trường: nhiệt độ, độ ẩm, mức độ bụi mịn.
Biến số đầu ra	Số bước đi, số bậc leo cầu thang, khoảng cách di chuyển hàng ngày.

10. Mô hình SVM (Support vector machine)[9]

Mô hình SVM là một công cụ mạnh để xác định người bệnh đợt cấp COPD đạt độ nhạy 80%, độ đặc hiệu 83%, PPV 81%, NPV 85%, AUC 0,90.

Bảng 9. Các biến số trong mô hình SVM

Biến số đầu vào	- Thông tin chung: tuổi, giới, tình trạng hút thuốc, tình trạng uống rượu. - Dấu hiệu sinh tồn: nhiệt độ, nhịp tim, nhịp thở, huyết áp. - Bệnh đồng mắc: tăng huyết áp, đái tháo đường, rối loạn lipid máu. - Triệu chứng: ho, sốt, khó thở, khò khè, đờm, đau ngực. - Xét nghiệm: PaO2, PaCO2, FEV1/FVC. - Chất lượng cuộc sống: tư thế nằm, chất lượng giấc ngủ, khả năng hoạt động.
Biến số đầu ra	Ho, sốt, khó thở, khò khè, đờm, đau ngực, tư thế nằm, chất lượng giấc ngủ, khả năng hoạt động.

IV. BÀN LUẬN

Các mô hình học máy trên thế giới sử dụng bộ dữ liệu lớn và thuật toán phức tạp để dự báo các đợt trầm trọng dựa trên nhiều biến số lâm sàng, nhân khẩu học và môi trường, mang lại những lợi ích đáng chú ý trong việc cải thiện chất lượng chăm sóc và quản lý người bệnh COPD. Nghiên cứu này đã xác định và đánh giá 9 mô hình dự báo có độ chính xác cao trong đó có 2 mô hình được điều chỉnh từ mô hình gốc để khắc phục một số hạn chế. Tuy tổng hợp thành các mô hình như các kết quả từ bảng 2 đến bảng 11 nhưng chúng tôi nhận thấy sự không đồng nhất về số lượng và chủng loại yếu tố dự đoán, cỡ mẫu, khoảng thời gian, phương pháp thống kê và thước đo hiệu suất của các mô hình. Đối với nomogram có ưu điểm là cung cấp một công cụ trực quan có thể giải thích cho các bác sĩ lâm sàng, dễ sử dụng để dự đoán rủi ro tuy nhiên có thể thiếu độ chính xác với các mối quan hệ phức tạp, phi tuyến tính trong bệnh cảnh COPD. Mô hình LASSO có lợi thế giúp lựa chọn tính năng bằng cách thu nhỏ hệ số của các biến ít quan trọng hơn về 0, giảm độ phức tạp của mô hình và ngăn ngừa quá tải nhưng có thể không phù hợp khi mối quan hệ giữa các biến là phi tuyến tính. Mô hình GLM có ưu điểm linh hoạt, dễ hiểu, có thể mô hình hóa các loại kết quả khác nhau (nhị phân, liên tục) nhưng vẫn hạn chế trong việc xử lý các tương tác phi tuyến tính phức tạp. Mô hình Random Forest có thể xử lý các bộ dữ liệu lớn và các mối quan hệ phi tuyến tính phức tạp. Đây là một mô hình mạnh, phù hợp với nhiều biến nhưng sẽ phức tạp trong việc diễn giải và đòi hỏi kĩ thuật điều chỉnh (huấn luyện) để đạt được hiệu suất tối ưu. Kết quả nghiên cứu của cho thấy khi kết hợp dữ liệu về lối sống.[10] Mô hình dự đoán đợt cấp COPD dựa vào dữ liệu về lối sống sử dụng thuật toán RF huấn luyện ít biến số nhất (3 biến số), nhưng lại có hiệu suất dự báo mạnh mẽ nhất (độ chính xác 92,1%, độ nhạy 94% và độ đặc hiệu 90,4%; AUC > 0,9) cung cấp cách tiếp cận đơn giản và chính xác.[11] Tuy ít tiêu tốn tài nguyên xét nghiệm, nhưng việc sử dụng điện thoại thông minh, thiết bị đeo và cảm biến chất lượng không khí lại tạo ra các rào cản công nghệ, cùng với chi phí cao và khả năng tiếp cận hạn chế, đặc biệt là ở các khu vực thu nhập thấp. Đồng thời, sự phụ thuộc vào các biến số lối sống có thể bỏ qua một số dấu hiệu lâm sàng quan trọng trong diễn biến của bệnh. Mô hình sử dụng ít biến số tiếp theo là mô hình dự báo tích hợp vào cơ sở chăm sóc ban đầu tại Thụy Điển (4 biến số) sử dụng thuật toán XGBoost đạt AUC 0,86 có thể tạo điều kiện can thiệp sớm và quản lý bệnh tốt hơn.[8] Tuy nhiên, mô hình này cũng cần cải thiện hiệu suất dự báo dài hạn do thời gian theo dõi ngắn

(10 ngày). Mô hình SVM sử dụng 9 biến số đơn giản, dễ thu thập, dễ theo dõi về triệu chứng và chất lượng cuộc sống đem lại hiệu suất dự báo cao.[9] Với cỡ mẫu tương đối nhỏ (303 người bệnh), mức độ phản ánh tổng thể và khái quát hóa của tập dữ liệu huấn luyện còn hạn chế, có thể tạo ra kết quả dự báo không ổn định khi áp dụng thực tế.

Hiện nay, tại Việt Nam, các nghiên cứu về đợt cấp COPD mới chỉ dừng lại ở việc tìm ra các yếu tố nguy cơ liên quan, mà chưa thiết lập một mô hình tiên lượng đợt bùng phát cụ thể.[13,14] Căn cứ vào các kết quả nghiên cứu, chúng tôi nhận thấy rằng ba thuật toán: RF, SVM và XGBoost có tính khả thi cao khi xây dựng mô hình học máy sử dụng bộ dữ liệu đặc thù của người bệnh Việt Nam, tùy chỉnh theo các yếu tố môi trường địa phương, phơi nhiễm nghề nghiệp trong bối cảnh thực tế và các nguồn lực sẵn có. Việc ứng dụng các thuật toán này hứa hẹn nâng cao độ chính xác trong dự báo, giúp cung cấp các thông tin quan trọng hỗ trợ bác sĩ theo dõi và điều trị người bệnh đợt cấp bệnh phổi tắc nghẽn mạn tính, góp phần cải thiện chất lượng chăm sóc sức khỏe tại Việt Nam.

V. KẾT LUẬN – KHUYẾN NGHỊ

Các mô hình dự báo đợt cấp của bệnh phổi tắc nghẽn mạn tính có AUC dao động từ 0,681 đến trên 0,9. 3 mô hình có hiệu suất cao nhất lần lượt là Random Forest (> 0,9), Support Vector Machine (0,9) và Extreme Gradient Boosting (0,86).

Trong bối cảnh Việt Nam, đặc biệt với nguồn lực hạn chế, chúng tôi nhận thấy rằng có thể sử dụng ba thuật toán: RF, SVM và XGBoost trong xây dựng mô hình học máy đối với bộ dữ liệu đặc thù của người bệnh Việt Nam, tùy chỉnh theo các yếu tố về tiền sử bệnh, một số đặc điểm nhân khẩu học, các thông tin lâm sàng của đợt vào viện điều trị... Việc ứng dụng các thuật toán này hứa hẹn nâng cao độ chính xác trong dự báo, giúp cung cấp các thông tin quan trọng hỗ trợ bác sĩ theo dõi và điều trị người bệnh đợt cấp bệnh phổi tắc nghẽn mạn tính, góp phần cải thiện chất lượng chăm sóc sức khỏe tại Việt Nam.

TÀI LIỆU THAM KHẢO

1. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for Prevention, Diagnosis and Management of COPD: 2024 Report. In: GOLD, ed. Bethesda https://goldcopd.org/2024-gold-report.2024.

2. Bộ Y tế. Quyết định 2767/QĐ-BYT ngày 4/7/2023 về việc ban hành tài liệu chuyên môn "Hướng dẫn chẩn đoán và điều trị bệnh phổi tắc nghẽn mạn tính". In:2023.

3. Hurst J R, Skolnik N, Hansen G J, et al. Understanding the impact of chronic obstructive pulmonary disease exacerbations on patient health and quality of life. *European journal of internal medicine.* 2020;73:1-6.

4. Chmiel FP, Burns DK, Pickering JB, et al. Prediction of Chronic Obstructive Pulmonary Disease Exacerbation Events by Using Patient Self-reported Data in a Digital Health App: Statistical Evaluation and Machine Learning Approach. *JMIR Med Inform.* 2022; 10(3): e26499.

5. Fakhraei R, Matelski J, Gershon A, et al. Development of Multivariable Prediction Models for the Identification of Patients Admitted to Hospital with an Exacerbation of COPD and the Prediction of Risk of Readmission: A Retrospective Cohort Study using Electronic Medical Record Data. *COPD.* 2023; 20(1): 274-283.

6. Huebner ST, Henny S, Giezendanner S, et al. Prediction of Acute COPD Exacerbation in the Swiss Multicenter COPD Cohort Study (TOPDOCS) by Clinical Parameters, Medication Use, and Immunological Biomarkers. *Respiration.* 2022; 101(5): 441-454.

7. Singh D, Hurst J R, Martinez F J, et al. Predictive modeling of COPD exacerbation rates using baseline risk factors. *Therapeutic advances in respiratory disease.* 2022; 16: 17534666221107314.

8. Ställberg B, Lisspers K, Larsson K, et al. Predicting Hospitalization Due to COPD Exacerbations in Swedish Primary Care Patients Using Machine Learning - Based on the ARCTIC Study. *Int J Chron Obstruct Pulmon Dis.* 2021; 16: 677-688.

9. Wang C, Chen X, Du L, et al. Comparison of machine learning algorithms for the identification of acute exacerbations in chronic obstructive pulmonary disease. *Comput Methods Programs Biomed.* 2020; 188: 105267.

10. Wu CT, Li GH, Huang CT, et al. Acute Exacerbation of a Chronic Obstructive Pulmonary Disease Prediction System Using Wearable Device Data, Machine Learning, and Deep Learning: Development and Cohort Study. *JMIR Mhealth Uhealth.* 2021; 9(5): e22591.

11. Zeng S, Arjomandi M, Tong Y, et al. Developing a Machine Learning Model to Predict Severe Chronic Obstructive Pulmonary Disease Exacerbations: Retrospective Cohort Study. *J Med Internet Res.* 2022; 24(1): e28953.

12. Zhang Y, Zheng S P, Hou Y F, et al. A predictive model for frequent exacerbator phenotype of acute exacerbations of chronic obstructive pulmonary disease. *Journal of thoracic disease.* 2023; 15(12): 6502-6514.

13. Nguyễn Thị Ngọc, Vũ Văn Sơn, Bùi Mỹ Hạnh, và cs. Đặc điểm lâm sàng, cận lâm sàng, mức độ nặng bệnh nhân đợt cấp bệnh phổi tắc nghẽn mạn tính (COPD) tại bệnh viện Phổi Trung ương. *Tạp chí Y học Việt Nam.* 2020; 1(495): 4-9.

14. Phùng Thị Thanh, Chu Thị Hạnh, Trần Thị Nương, và cs. Một số yếu tố nguy cơ gây đợt cấp thường xuyên ở bệnh nhân có đợt cấp bệnh phổi tắc nghẽn mạn tính nhập viện tại Trung tâm Hô hấp, Bệnh viện Bạch Mai. *Tạp chí Y học Việt Nam.* 2022; 1(514): 100-104.

Summary

PREDICTION MODELS FOR CHRONIC OBSTRUCTIVE PULMONARY DISEASE EXACERBATION: A LITERATURE REVIEW

Machine learning techniques for predicting chronic obstructive pulmonary disease (COPD) exacerbation is the revolution in COPD management by allowing early detection, personalized intervention, resource optimization, and patient empowerment. The results identified 9/928 articles that fully met the selection criteria, including: 7 multicenters retrospective observations, 1 single-center prospective observation, and 1 single-center clinical trial. 117 risk factors were included in the prediction models, of which age and gender appeared most commonly (9/9 times). The models had the area under the curve ranging from 0.681 to over 0.9. The 3 highest performance models were Random Forest (> 0.9), Support Vector Machine (0.9), and Extreme Gradient Boosting (0.86), respectively, which need to be applied, further built, and developed on the Vietnamese dataset.

Keywords: Chronic Obstructive Pulmonary Disease Exacerbation, model, prediction, machine learning.