Tạp chí quốc tế về tin học dữ liệu và điện toán thông minh (IJDIIC)
Tập 2, Số 3, Tháng 9 năm 2023, trang 47~56
HAI: 10.59461/ijdiic.v2i3.73

Kỹ thuật học máy để dự đoán nguy cơ ung thư phổi bằng cách sử dụng bộ dữ liệu văn bản

Kumar Mohan1 , Bharaguram Thayyil2
1,2Khoa Công nghệ thông tin, Đại học Công nghệ và Khoa học ứng dụng-Shinas, Al Aqar, Oman

Thông tin bài viết

Lịch sử bài viết:

Nhận ngày 12 tháng 9 năm 2023
Đã sửa đổi ngày 20 tháng 9 năm 2023
Đã chấp nhận ngày 21 tháng 9 năm 2023

Từ khóa:

Học máy
Dự đoán
Ung thư phổi
Rừng ngẫu nhiên

TÓM TẮT

Các triệu chứng ban đầu của ung thư phổi, một mối đe dọa nghiêm trọng đối với sức khỏe con người, tương đương với các triệu chứng của cảm lạnh thông thường và viêm phế quản. Các chuyên gia lâm sàng có thể sử dụng các kỹ thuật học máy để tùy chỉnh các chiến lược sàng lọc và phòng ngừa theo nhu cầu riêng của từng bệnh nhân, có khả năng cứu sống và nâng cao chất lượng chăm sóc bệnh nhân. Các nhà nghiên cứu phải xác định các biến số lâm sàng và nhân khẩu học có liên quan từ hồ sơ bệnh nhân và xử lý trước và chuẩn bị tập dữ liệu để đào tạo mô hình học máy nhằm dự đoán chính xác sự phát triển của ung thư phổi. Mục tiêu của nghiên cứu là phát triển một mô hình học máy (ML) chính xác và dễ hiểu để dự đoán ung thư phổi giai đoạn đầu bằng cách sử dụng các biến số nhân khẩu học và lâm sàng, cũng như đóng góp vào lĩnh vực ứng dụng ML nghiên cứu y khoa đang phát triển có thể cải thiện kết quả chăm sóc sức khỏe. Để tạo ra mô hình dự đoán hiệu quả và chính xác nhất, các kỹ thuật học máy như Hồi quy logistic, Cây quyết định, Rừng ngẫu nhiên, Máy vectơ hỗ trợ, K-Nearest Neighbor (KNN) và Naive Bayes đã được sử dụng trong bài viết này.

Tác giả liên hệ:

Kumar Mohan

Khoa Công nghệ thông tin

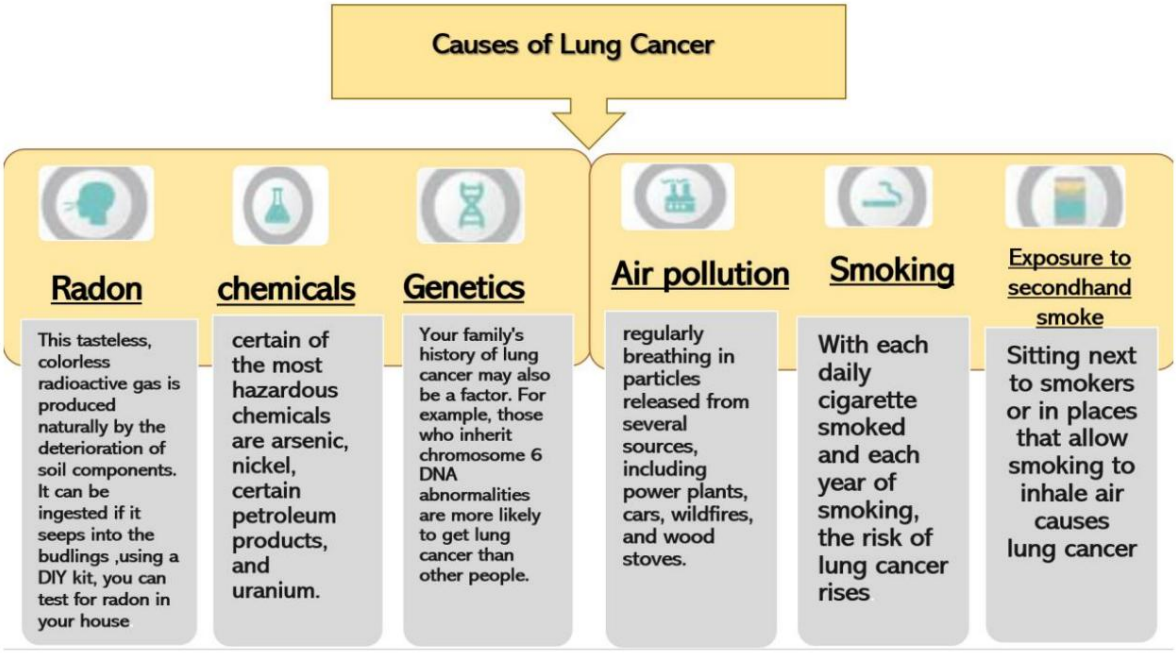
Đại học Công nghệ và Khoa học Ứng dụng-Shinas Al Aqar Oman

Email: kumar.mohan@shct.edu.om

1. GIỚI THIỆU

Trong thế giới ngày nay, ung thư phổi gây ra nhiều ca tử vong do các nguyên nhân liên quan đến ung thư hơn bất kỳ loại ung thư nào khác như ung thư vú, tuyến tiền liệt và ruột kết. Chẩn đoán và điều trị ung thư phổi sớm có sự cân nhắc thỏa đáng và tác động tích cực đến kết quả của bệnh nhân và tỷ lệ tử vong. Bệnh thường được phát hiện ở giai đoạn tiến triển trong nhiều trường hợp ung thư phổi vì các kỹ thuật sàng lọc hiện có không phải lúc nào cũng đáng tin cậy. Đối với chẩn đoán sớm ung thư phổi, các kỹ thuật học máy đưa ra một phương pháp khả thi [1]. Các mô hình học máy có thể dự đoán đáng tin cậy khả năng mắc ung thư phổi của bệnh nhân bằng cách xác định các mô hình và yếu tố rủi ro liên quan đến bệnh bằng cách đánh giá cơ sở dữ liệu lớn về hồ sơ bệnh nhân. Sử dụng dữ liệu hình ảnh y tế từ chụp CT và chụp X-quang, các mô hình học máy đã được tạo ra trong những năm gần đây để hỗ trợ việc xác định sớm ung thư phổi. Tuy nhiên, các kỹ thuật này có thể tốn kém và mất thời gian. Trong bài viết này, chúng tôi khám phá ứng dụng của các kỹ thuật học máy để ước tính nguy cơ ung thư phổi bằng cách sử dụng các tập dữ liệu văn bản. Mô hình này đặc biệt xem xét hồ sơ sức khỏe điện tử (EHR) bao gồm dữ liệu bệnh nhân như thông tin nhân khẩu học, ghi chú lâm sàng và tiền sử bệnh. Chúng ta cần dự báo đúng nguy cơ mắc ung thư phổi của bệnh nhân bằng cách trích xuất các yếu tố có liên quan từ các tập dữ liệu văn bản này và xây dựng

thuật toán học máy. Phần này cố gắng chứng minh đánh giá tổng thể trong các nghiên cứu đã phân tích dữ liệu biểu hiện gen trong ung thư. Điều này bao gồm nhiều phương pháp ma trận như (Ma trận nhầm lẫn, Độ chính xác, Độ chính xác, Thu hồi, Điểm F1) và các phương pháp thuật toán như (Hồi quy logistic, Cây quyết định, Rừng ngẫu nhiên, Máy vectơ hỗ trợ (SVM), K-Nearest Neighbor (KNN), Naive Bayes) để khám phá gen từ các mẫu và sử dụng các chữ ký biểu hiện này để phát triển mô hình dự đoán ung thư.



Hình 1. Sơ đồ nguyên nhân gây ung thư phổi.

Hình 1 cho thấy nguyên nhân gây ung thư phổi, thường được đưa vào cơ thể con người thông qua nhiều môi trường sống khác nhau như hút thuốc. Cả người hút thuốc và những người tiếp xúc với khói thuốc lá cũng vậy. Một số vết thương có thể xảy ra ở những người không hút thuốc, ví dụ, những người tiếp xúc với khí radon, tiếp xúc với các tác nhân gây ung thư và tiền sử gia đình bị suy giảm tế bào ở phổi.

2. TỔNG QUAN TÀI LIỆU

Đầu tiên, bài báo [4] trình bày các tính năng dựa trên học sâu và về mặt hiệu suất, mẫu đánh bại chiến lược học máy thông thường. Đối với mẫu học sâu, phương pháp thu được giá trị 0,5 cho chiến lược hồi quy và độ chính xác tối đa là 71,18 phần trăm cho phương pháp phân loại. Tuy nhiên, RMSE mẫu hồi quy cho các mô hình học máy thông thường vẫn không đổi ở mức lần lượt là 14,87 và 61,12 phần trăm. Trong bài báo [5], khoảng 34% khối u ác tính ở phổi được phát hiện ở các nốt phổi, là những tổn thương tròn hoặc không bằng nhau ở phổi. Do đó, chẩn đoán nốt phổi là rất quan trọng để phát hiện sớm ung thư phổi. Những phát hiện của thí nghiệm chứng minh rằng phương pháp dựa trên mạng nơ-ron tích chập để xác định và phát hiện ung thư phổi có mức độ chính xác cao hơn. Ngoài ra, bài báo [6] xem xét cơ sở dữ liệu các bài báo và bài đánh giá sách của Web of Science (WoS), nơi nó dẫn dắt các nhà nghiên cứu sử dụng phương pháp trắc lượng thư mục để tiến hành phân tích định lượng về sản phẩm khảo sát tại 24 quốc gia đứng đầu thế giới về nghiên cứu ung thư. Phân tích bao gồm 32.161 bài báo nghiên cứu về ung thư phổi từ 2085 tạp chí riêng biệt. Khoảng 5,6% trong tổng số các nghiên cứu về ung thư tập trung vào ung thư phổi vào năm 2013 và tăng khoảng 1,2% kể từ năm 2004. Thông tin cho thấy rằng, mặc dù có những hậu quả đáng kể về mặt lâm sàng, xã hội và tài chính liên quan đến ung thư phổi, nhưng lượng đầu ra nghiên cứu toàn cầu vẫn tụt hậu đáng kể so với các bệnh ác tính.

Trong bài báo [7], tiên lượng cho bệnh ung thư phổi không phải tế bào nhỏ tiến triển tại chỗ, không thể chữa khỏi đang dần mang lại kết quả tốt hơn bằng cách đưa hóa trị vào xạ trị triệt để và các phương pháp điều trị xạ trị mới. Với các loại thuốc hiện tại được dung nạp tốt hơn và dẫn đến cải thiện chất lượng cuộc

hóa trị liệu mang lại sự cải thiện khả năng sống sót vừa phải cho những cá nhân mắc ung thư phổi không phải tế bào nhỏ. Bài báo [8] trình bày kết quả tốt hơn về độ tuổi trung bình (khoảng tứ phân vị) của 17 322 bệnh nhân NSCLC, trong đó có 68 (61-74) và 13 361 (77,1%) trong số họ là người da trắng trong tiêu chí đánh giá. 10 273 khối u giai đoạn I (chiếm 59,3% trong số tất cả các loại ung thư) và 11 985 ung thư biểu mô tuyến (chiếm 69,2% trong số tất cả các loại khối u) chiếm phần lớn các khối u. Thời gian theo dõi là 24 (10–43) tháng là thời gian trung bình (khoảng tứ phân vị). Trong thời gian theo dõi, 3119 bệnh nhân đã tử vong vì ung thư phổi.

Bài viết [9] trình bày rằng ung thư phổi là nguyên nhân nghiêm trọng gây tử vong do ung thư trên toàn thế giới. Tỷ lệ tử vong do ung thư phổi giảm từ 20% đến 24% khi ung thư phổi được sàng lọc trên chụp cắt lớp vi tính (CT) ngực. Trong bài báo [10], Với việc sử dụng hình ảnh chụp CT, một phương pháp quét chi tiết được khai quật bằng một chiến lược kỹ thuật mới, thách thức trong việc xác định các loại ung thư phổi lành tính và ác tính có thể được giải quyết, hỗ trợ khả năng chẩn đoán ung thư phổi dễ dàng hơn của cộng đồng y tế.

Theo thử nghiệm hệ thống, tỷ lệ chính xác để dự đoán chẩn đoán ung thư phổi là lành tính hay ác tính là 83,33%. Bài báo [11] tập trung nhiều hơn vào tuổi, nguy cơ CT và PET và nguy cơ OS đã được sử dụng để dự đoán trong khi nguy cơ PET có tác động tiêu cực đến OS với HR là 0,67 (bao gồm tác động). Nguy cơ CT có tác động tích cực đến OS với HR là 1,35 (nguy cơ tăng cường) và bài báo cố gắng chứng minh

tuổi tác có tác động không đáng kể.

Trong bài báo [12], các tác giả xem xét mức độ dân số của tỷ lệ tử vong do NSCLC ở Hoa Kỳ giảm đáng kể trong giai đoạn 2013-2016, trong khi tỷ lệ sống sót sau khi chẩn đoán được cho là tăng đáng kể. Nghiên cứu này dự đoán tỷ lệ mắc bệnh liên quan đến những tiến bộ y tế, đáng chú ý là việc sử dụng và chấp thuận các loại thuốc nhắm mục tiêu. Bài báo [13] dự kiến một nỗ lực kỹ thuật nhằm tạo ra một hệ thống tự động để phát hiện ung thư phổi. Các phương pháp được sử dụng trong nghiên cứu này đã tạo ra độ chính xác của cơ sở dữ liệu bệnh viện là 92%. Công nghệ này cố gắng cải thiện độ chính xác và tốc độ của hệ thống phát hiện ung thư phổi.

Ngoài ra, nó còn hỗ trợ phát hiện ung thư sớm hơn.

Trong bài báo [14], những bệnh nhân ung thư phổi giai đoạn đầu không thể phẫu thuật hiện đang được xạ trị cắt bỏ định vị (SABR) là mức độ chăm sóc hợp lý. Phần lớn những bệnh nhân này là người cao tuổi và có đột biến trong EGFR (thụ thể yếu tố tăng trưởng biểu bì). Để xác định cách lão hóa và chuyển đổi EGFR ảnh hưởng đến kết quả điều trị và độc tính như thế nào, bài báo đã xem xét hồ sơ bệnh án của 71 bệnh nhân đã được xạ trị cắt bỏ định vị tại Bệnh viện Đa khoa Cựu chiến binh Đài Bắc trong giai đoạn 2015 và 2021 và mắc ung thư phổi không phải tế bào nhỏ (NSCLC) giai đoạn đầu không thể chữa khỏi. Xạ trị cắt bỏ định vị thân (SBRT) hoặc xạ trị cắt bỏ định vị (SABR) có thể kết thúc quá trình điều trị trong một đến hai tuần. Những phát hiện này cho thấy hiệu quả và tính an toàn của SABR ở những bệnh nhân ung thư phổi không phải tế bào nhỏ giai đoạn đầu không thể chữa khỏi không bị ảnh hưởng bởi tuổi tác hoặc tình trạng đột biến EGFR. 37 (52,1%) trong số những bệnh nhân này từ 80 tuổi trở lên, 50 (70,4%) mắc bệnh T1 và 21 (29,6%) mắc bệnh T2. Tình trạng đột biến EGFR của 33 (46,5%) bệnh nhân được biết đến và 16 (51,5%) trong số những bệnh nhân này có đột biến.

Trong bài báo [15], các tác giả đã trình bày nhiều lợi thế của việc sử dụng chẩn đoán ung thư sớm để dự đoán tỷ lệ sống sót. Sử dụng các phương pháp học sâu, một phương pháp tiếp cận chiến lược đã được phát triển và một số mô hình dự đoán khả năng sống sót cho bệnh nhân ung thư phổi và giải quyết vấn đề phân loại và hồi quy về khả năng sống sót của bệnh nhân ung thư đã được trình bày. Chúng tôi đã phân tích bài trình bày trên ba cấu trúc học sâu nổi tiếng nhất - Tổ chức não giả (ANN), Tổ chức não tích chập (CNN) và Tổ chức não không liên tục (RNN) trong khi xem xét hiệu suất của các mô hình học sâu so với mẫu AI thông thường. Trong cả phương pháp phân loại và hồi quy, các mô hình học sâu đều vượt trội hơn các mô hình học máy truyền thống về mặt hiệu suất. Các nhà nghiên cứu

đạt được tỷ lệ 0,5 cho phương pháp hồi quy và độ chính xác tối đa là 71,18 phần trăm trong khi phương pháp phân loại cho các mô hình học sâu. RMSE trong mô hình hồi quy cho các mô hình học máy truyền thống vẫn không đổi ở mức 14,87 phần trăm và 61,12 phần trăm.

Trong bài báo [16], các tác giả tạo điều kiện cho một cơ chế phát hiện sớm bằng cách nghiên cứu những năm gần đây. AI và máy học đã được nghiên cứu và sử dụng để phát hiện sớm căn bệnh này. Các phương pháp mới đã được phát triển bằng cách kết hợp phát hiện dữ liệu và xử lý hình ảnh y sinh. Trên một tập dữ liệu bệnh ung thư phổi, các thuật toán máy học đã được sử dụng để phân loại hình ảnh và tính toán độ chính xác, độ nhạy và các số liệu khác trong bài báo này. Giai đoạn đầu của ung thư phổi đã được phân tích bằng các thuật toán K-NN, Random Forest và SVM. Có thể đưa ra quyết định kết quả về tỷ lệ mắc ung thư phổi là tỷ lệ tử vong cao hơn các loại ung thư khác. Có thể điều trị bệnh kịp thời và cứu sống người bệnh trong trường hợp phát hiện sớm hơn. Chụp CT đã được áp dụng rộng rãi, đáng tin cậy hơn và chính xác hơn để phân biệt sự cải thiện bệnh ở phổi kể từ khi phát hiện ra bất kỳ sự cố tế bào nào không ngờ tới

trong các nốt phổi cũng có khả năng như vậy. Tuy nhiên, vì cường độ của hình ảnh CT thay đổi nên việc phát hiện chính xác vẫn là một thách thức đáng kể. Chẩn đoán hỗ trợ máy tính (CAD) đã nổi lên như một chiến lược hỗ trợ chính trong cuộc chiến chống lại vấn đề này. Ngoài ra, nghiên cứu đang được tiến hành trong lĩnh vực phát hiện ung thư phổi này để đạt được độ chính xác phát hiện 100%. Trong bài viết này, các cải tiến tiên tiến được sử dụng để dự đoán sự cố tế bào trong phổi bằng cách sử dụng hình ảnh từ bộ lọc CT được thảo luận và một cuộc kiểm tra được dẫn dắt họ để phân biệt phương pháp hàng đầu. 3 bộ phân loại (K-NN, Irregular woods và SVM) được sử dụng làm các biến thể kết quả và dựa trên kết quả và so sánh, các tác giả xác định rằng Random Forest, tiếp theo là SVM, với độ chính xác là 82,1 phần trăm, tạo ra kết quả hàng đầu trong số ba thuật toán.

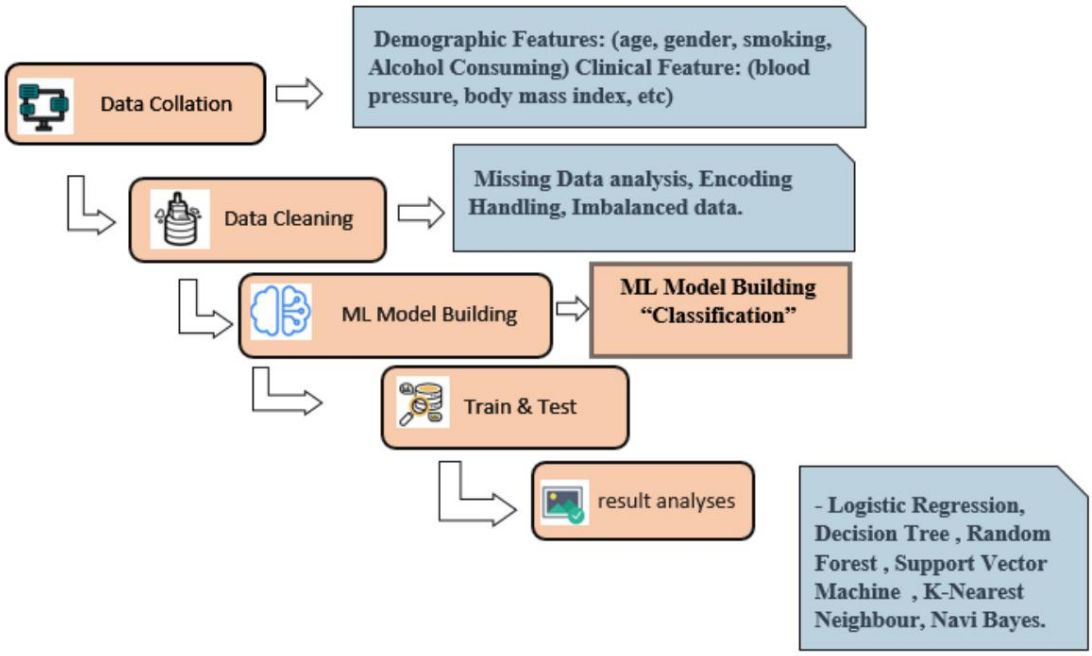
Trong bài báo [17], một cuộc khảo sát đã được thực hiện về nghiên cứu đưa ra phác thảo về các phương pháp tiếp cận dựa trên AI củng cố các phần thay đổi của sự cố tế bào trong kết luận và điều trị phổi, bao gồm vị trí sớm, phát hiện trợ lý, kỳ vọng dự báo và thực hành liệu pháp miễn dịch. Khoảng 2,20 triệu bệnh nhân mới được xác định mắc chứng cố tế bào ở phổi mỗi năm và 75% trong số họ tử vong trong khoảng thời gian năm năm phân tích. Sự phức tạp của các tế bào ung thư gây ra tình trạng kháng thuốc và tính không đồng nhất trong khối u (ITH) cao khiến việc điều trị ung thư trở nên khó khăn hơn. Sự tiến bộ liên tục của công nghệ nghiên cứu ung thư trong vài thập kỷ qua đã dẫn đến sự hình thành của nhiều cơ sở dữ liệu lâm sàng, hình ảnh y tế và bộ gen do nhiều dự án ung thư hợp tác lớn. Các cơ sở dữ liệu này giúp các nhà nghiên cứu dễ dàng hơn trong việc tìm hiểu toàn bộ các mô hình ung thư phổi, từ chẩn đoán đến điều trị và
phản ứng với hậu quả lâm sàng. Nghiên cứu hiện tại về phân tích -omics, bao gồm bộ gen, phiên mã, proteomics và chuyển hóa, đã cải thiện đáng kể các công cụ và năng lực nghiên cứu của chúng tôi.
Chụp cắt lớp vi tính (CT) liều thấp là kỹ thuật chính để theo dõi những người có nguy cơ mắc ung thư phổi. Mục tiêu của hệ thống chẩn đoán hỗ trợ máy tính (CAD) là tăng hiệu quả chẩn đoán bằng cách hỗ trợ bác sĩ giải thích dữ liệu hình ảnh y tế.

Bài báo [18] đã trình bày một mô hình nguyên mẫu để điều trị ung thư phổi có thể được tạo ra mà không gây nguy hiểm cho môi trường bằng cách sử dụng những phát triển mới nhất trong trí tuệ tính toán. Hệ thống tiết kiệm thời gian và tiền bạc vì nó có thể cắt giảm lượng tài nguyên bị lãng phí và lượng lao động cần thiết để thực hiện các tác vụ thủ công. Quá trình phát hiện từ tập dữ liệu ung thư phổi đã được tối ưu hóa bằng cách sử dụng mô hình học máy được xây dựng trên kiến trúc máy vectơ hỗ trợ (SVM). Bộ phân loại SVM được sử dụng để phân loại bệnh nhân ung thư phổi theo các triệu chứng của họ và Python được sử dụng để nâng cao việc triển khai mô hình. Chúng tôi đã đánh giá hiệu quả của mô hình SVM của mình bằng nhiều tiêu chí khác nhau để xây dựng kiến trúc dựa trên mô hình. Theo một số nghiên cứu, bệnh phổi chiếm hơn 13% trong số tất cả các phân tích về sự phát triển ác tính ở Hoa Kỳ vào năm 2015. Theo Hiệp hội Ung thư Hoa Kỳ, 27% trong số tất cả các trường hợp tử vong liên quan đến ung thư được cho là do ung thư phổi. Do đó, điều quan trọng là phải đánh giá và theo dõi đúng cách các nốt sần ở phổi trong giai đoạn phát triển của nó. Mục đích của khảo sát này là khám phá sự phát triển và tiến triển của bệnh ung thư bằng phương pháp ML và DL để dự đoán sự phát triển và tiến triển.

Tóm lại, bài tổng quan tài liệu cung cấp tổng quan toàn diện về ung thư phổi, nguyên nhân, chẩn đoán và các lựa chọn điều trị. Bài tổng quan nhấn mạnh tầm quan trọng của việc phòng ngừa và phát hiện sớm trong việc nâng cao kết quả cho bệnh nhân ung thư phổi. Hơn nữa, các nghiên cứu dự kiến sẽ xác định các lựa chọn điều trị mới và nâng cao tỷ lệ sống sót chung cho bệnh nhân ung thư phổi.

3. PHƯƠNG PHÁP

Bài viết này sử dụng sự kết hợp giữa thu thập dữ liệu, xử lý trước, phát triển mô hình, xác thực, tối ưu hóa và diễn giải như một phương pháp nghiên cứu cho các phương pháp học máy để dự đoán nguy cơ ung thư phổi bằng cách sử dụng các tập dữ liệu văn bản để tạo ra một mô hình dự đoán chính xác và đáng tin cậy cho vấn đề sức khỏe quan trọng này. Nghiên cứu này dựa trên phương pháp tiếp cận hỗn hợp (định lượng và định tính) và sử dụng dạng dữ liệu thứ cấp để có kết quả tốt hơn. Nghiên cứu thực nghiệm này chạy qua nhiều tiêu chí dựa trên sức khỏe khác nhau và xây dựng một phương pháp tiếp cận với mô hình thuật toán hỗn hợp.
Hệ thống đề xuất: Sau khi xử lý trước, dữ liệu hiện đã sẵn sàng để sử dụng trong quá trình tạo mô hình. Quy trình xây dựng mô hình yêu cầu sử dụng các thuật toán học máy và một tập dữ liệu được xử lý trước. Một số kỹ thuật được sử dụng bao gồm phân loại LR, DT và RF. Có một số yếu tố nhất định đã được đánh giá và thử nghiệm trong bài viết này. Sơ đồ mô tả hệ thống được thiết kế được hiển thị trong Hình 2. Các tiểu mục sau đây bao gồm tất cả các thành phần của sơ đồ khối một cách chi tiết [20]. Các kỹ thuật
của việc thu thập dữ liệu văn bản có thể được áp dụng là nghiên cứu khảo sát.



Hình 2. Sơ đồ sản xuất ung thư phổi.

Có một số lý do cần cân nhắc khi nghiên cứu khảo sát có thể được sử dụng trong nghiên cứu ung thư phổi:

Nhận biết các yếu tố rủi ro: Nghiên cứu khảo sát có thể được sử dụng để thu thập thông tin về thói quen, lối sống và các yếu tố môi trường có thể làm tăng nguy cơ ung thư phổi. Thông tin này có thể được các nhà nghiên cứu sử dụng để phát hiện các xu hướng và các yếu tố rủi ro liên quan đến bệnh tật.

Hiểu về trải nghiệm của bệnh nhân: Nghiên cứu khảo sát cũng có thể được sử dụng để hiểu về trải nghiệm của bệnh nhân sau khi nhận được chẩn đoán ung thư phổi. Các nhà nghiên cứu có thể tìm hiểu về các triệu chứng, trải nghiệm điều trị và mối quan tâm về chất lượng cuộc sống bằng cách phỏng vấn bệnh nhân.

Đánh giá các biện pháp can thiệp: Nghiên cứu khảo sát có thể được sử dụng để đánh giá hiệu quả của các chương trình phòng ngừa hoặc điều trị ung thư phổi. Ví dụ, các nhà nghiên cứu có thể sử dụng các cuộc khảo sát để thu thập thông tin về các chương trình sàng lọc ung thư phổi hoặc các sáng kiến giúp người hút thuốc bỏ thuốc.

Nhìn chung, nghiên cứu khảo sát là phương pháp rất hữu ích để thu thập thông tin về nhiều khía cạnh của ung thư phổi, có thể giúp hướng dẫn các kế hoạch phòng ngừa và điều trị. [21]

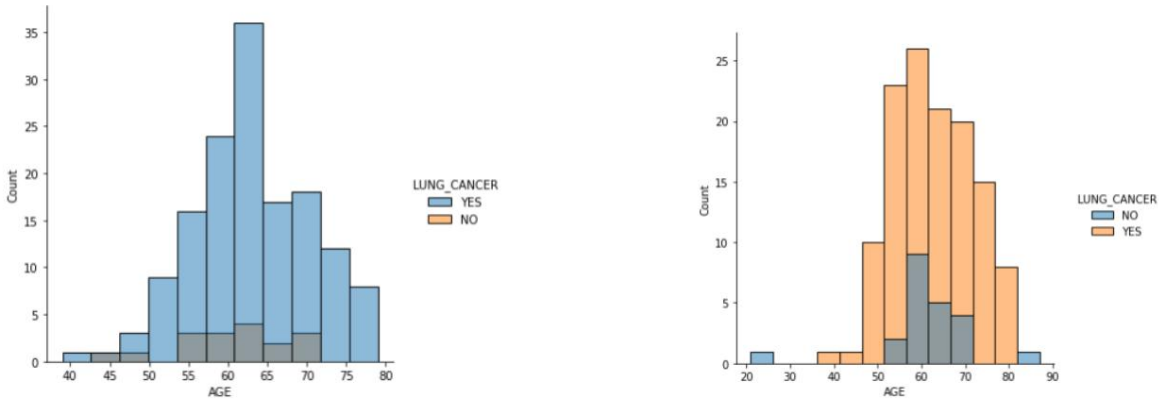
3.1. Tiền xử lý dữ liệu

Do các giá trị bị thiếu và/hoặc dữ liệu nhiễu, chất lượng của dữ liệu thô có thể tệ hơn chất lượng của dự báo cuối cùng. Để làm cho dữ liệu phù hợp hơn cho việc khai thác và phân tích, cần phải xử lý trước và giảm các giá trị dự phòng kỹ lưỡng, lựa chọn đặc điểm và phân loại dữ liệu được thực hiện. Cân bằng lớp bằng cách sử dụng phương pháp lấy mẫu lại là một thành phần khác của quá trình chuẩn bị dữ liệu. Chúng tôi đã sử dụng thông số kỹ thuật cấu trúc của SMOTE trong khuôn khổ được đề xuất để giảm bớt sự phân chia không đồng đều của những người tham gia trong nhóm ung thư phổi và không phải ung thư phổi. Cụ thể hơn, những người tham gia được phân bổ đều vì nhóm thiểu số trong tình trạng ung thư phổi là quá mức trong một số điều kiện nhất định. Ngoài ra, không có thông tin dữ liệu hoặc loại bỏ nào được sử dụng vì không có giá trị bị thiếu hoặc giá trị null.

Giới tính của mỗi người tham gia và nhóm tuổi mà họ thuộc về được thể hiện trong Hình 3 cùng với sự phân bố của học sinh trong mỗi lớp. Đối với đặc điểm lớp ung thư phổi, hình thứ hai

một phần cho thấy một phần đáng kể những người tham gia là những người trên 80 tuổi, trong khi độ tuổi phổ biến thứ hai là từ 30 đến 82. Ngoài ra, chúng ta có thể thấy từ hình này rằng ung thư phổi chủ yếu ảnh hưởng đến người cao tuổi. Ngược lại, Trong hình bên trái, một phần đáng kể những người tham gia là những người trên 75 tuổi, trong khi độ tuổi phổ biến thứ hai là từ 38 đến 79. Ngoài ra, chúng ta có thể thấy từ hình này rằng ung thư phổi chủ yếu ảnh hưởng đến người cao tuổi. Theo Hình 3, có khoảng 27% và 36% phụ nữ

hơn nam giới bị ung thư phổi. Điều đó cho thấy mặc dù ung thư phổi vẫn ảnh hưởng đến cả nam giới và nữ giới, nam giới có khả năng mắc bệnh cao hơn 9%.

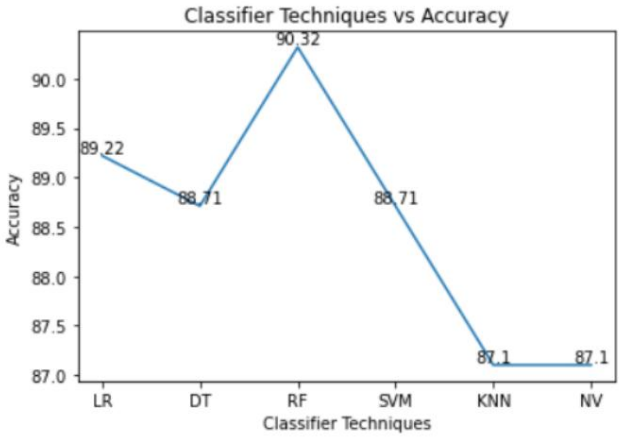


Hình 3. Phân bố người tham gia theo nhóm tuổi và giới tính trong tập dữ liệu cân bằng.

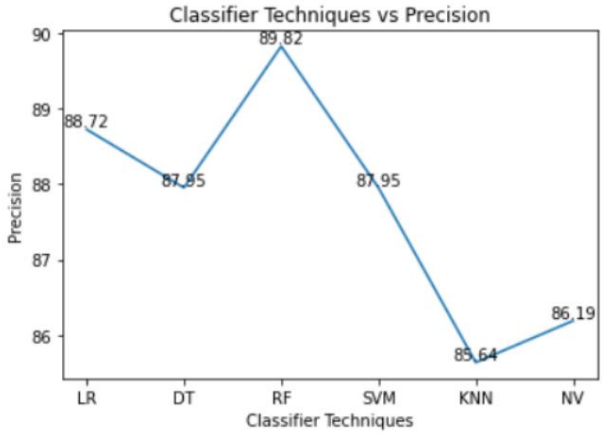
Phân tán tái diễn với các lớp vô hạn được hiển thị thông qua biểu đồ histogram. Đây là phác thảo của một vùng bao gồm các hình vuông có đáy ở các khoảng của các lớp tương phản và các vùng tỷ lệ thuận với tần suất của các lớp tương phản. Các hình vuông được kết nối khi đáy lấp đầy các vùng giữa các ranh giới lớp. Tỷ lệ của các hình vuông tạo nên các độ cao tương ứng với tần suất lớp tương đối và mật độ tái diễn cho các lớp khác nhau. Một số đặc điểm chính của biểu đồ histogram được minh họa trong Hình 4. Phân phối của tập dữ liệu được hiển thị trong biểu đồ histogram.

4. KẾT QUẢ VÀ PHÂN TÍCH

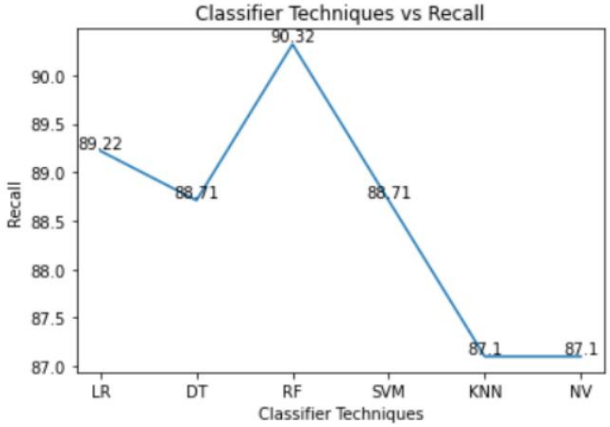
Khả năng của mô hình, dự đoán mô hình, phân tích và kết quả được đề cập trong phần này với thảo luận. Hình 4, 5, 6 và 7 minh họa hiệu suất của mô hình ML, dành riêng cho lớp ung thư phổi, về độ chính xác, khả năng thu hồi, độ chính xác và điểm F1.



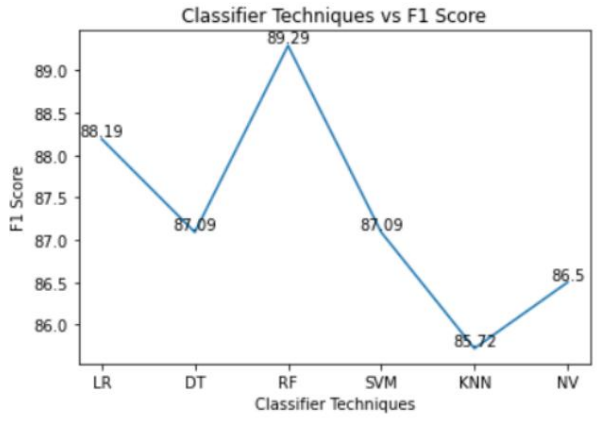
Hình 4. Các số liệu hiệu suất cho độ chính xác



Hình 5. Các số liệu hiệu suất cho độ chính xác



Hình 6. Các số liệu hiệu suất cho việc thu hồi



Hình 7. Chỉ số hiệu suất cho Điểm F1.

Bảng 1 và hình minh họa rõ ràng rằng mô hình RF là mô hình hiệu quả nhất trong số các mô hình đó. Tập trung vào các số liệu hiệu suất Độ chính xác, Độ chính xác, Độ thu hồi và Điểm F1, các mô hình RF có hiệu suất cao hơn, cho thấy rằng, với độ chính xác cao là 90,32%, Độ chính xác là 89,82%,

Thu hồi 90,32% và Điểm F1 89,29 có thể xác định thành công ung thư phổi từ các trường hợp không phải ung thư phổi.

Bảng 1. Hiệu suất của các mô hình ML

	Sự chính xác	Độ chính xác	Nhớ lại	Điểm F1
LỚP 1	89,22	88,72	89,22	88,19
Số 1	88,71	87,95	88,71	87.09
[illegible]	90,32	89,82	90,32	89,29
SVM	88,71	87,95	88,71	87.09
KNN	87,1	85,64	87,1	85,72
NV	87,1	86,91	87,1	86,5

Khi chúng tôi so sánh kết quả của mình với những kết quả khác, các bản sao tìm thấy trong các cuộc điều tra trước đây được đối chiếu trong Bảng 2 cho thấy chúng tôi đã có được kết quả tốt nhất trong Random Forest.

Bảng 2. Kết quả công việc đề xuất và kết thúc

ma trận	Bài viết thoát [16]	Rừng ngẫu nhiên được đề xuất
Sự chính xác	73,1%	90,32%
Độ chính xác	84%	89,82%
Nhớ lại	78,2%	90,32%
Điểm F1	81%	87,09%

5. KẾT LUẬN

Tóm lại, các yếu tố có thể được trích xuất từ cơ sở dữ liệu hồ sơ bệnh nhân khổng lồ bằng thuật toán máy học, sau đó có thể được sử dụng để dự báo khả năng phát triển ung thư phổi với độ chính xác cao.
Điều này có thể cải thiện kết quả điều trị cho bệnh nhân và tiết kiệm chi phí chăm sóc sức khỏe bằng cách cho phép chẩn đoán sớm hơn và điều trị bệnh tốt hơn. Nghiên cứu đánh giá mức độ các thuật toán khác nhau có thể dự đoán sự phát triển của ung thư phổi ở những cá nhân. Kết quả chứng minh rằng thuật toán rừng ngẫu nhiên đã vượt qua đối thủ cạnh tranh về hiệu suất tốt hơn.

TÀI LIỆU THAM KHẢO

[1] M. Guo et al., “Hóa trị liệu nhắm mục tiêu dựa trên vi hạt có nguồn gốc từ tế bào khối u tự thân ở bệnh nhân ung thư phổi có tràn dịch màng phổi ác tính,” Sci. Transl. Med., tập 11, số 474, tháng 1 năm 2019, doi: 10.1126/scitranslmed.aat5690.

[2] L. Zhang, Y. Hang, M. Liu, N. Li, và H. Cai, “Durvalumab tuyến đầu cộng với Platinum-Etoposide so với Platinum-Etoposide cho bệnh ung thư phổi tế bào nhỏ giai đoạn lan rộng: Phân tích hiệu quả chi phí,” Front. Oncol., tập 10, tháng 12 năm 2020, doi: 10.3389/fonc.2020.602185.

[3] S. C, H. SA và G. HL, “Các kỹ thuật loại bỏ hiện vật cho hình ảnh CT phổi trong phát hiện ung thư phổi,” Quốc tế J. Dữ liệu tin học Intell. Máy tính, tập. 1, không. 1, trang 21–29, tháng 9 năm 2022, doi: 10.59461/ijdiic.v1i1.14.

[4] E. Dritsas và M. Trigka, “Dự đoán rủi ro ung thư phổi bằng mô hình học máy”, Big Data Cogn. Comput., tập 6, số 4, trang 139, tháng 11 năm 2022, doi: 10.3390/bdcc6040139.

[5] Y. Zhang, B. Dai, M. Dong, H. Chen và M. Zhou, “Phương pháp phát hiện và nhận dạng ung thư phổi kết hợp mạng nơ-ron tích chập và các đặc điểm hình thái”, trong Hội nghị quốc tế lần thứ 5 về Công nghệ kỹ thuật máy tính và truyền thông (CCET) của IEEE năm 2022, IEEE, tháng 8 năm 2022, tr. 145–149. doi: 10.1109/CCET55412.2022.9906329.

[6] A. Aggarwal và cộng sự, “Tình hình nghiên cứu ung thư phổi: Phân tích toàn cầu”, J. Thorac. Oncol., tập 11, số 7, trang 1040–1050, tháng 7 năm 2016, doi: 10.1016/j.jtho.2016.03.010.

[7] SG Spiro và GA Silvestri, “Một trăm năm ung thư phổi,” Am. J. Respir. Crit. Care

Med., tập 172, số 5, trang 523–529, tháng 9 năm 2005, doi: 10.1164/rccm.200504-531OE.

[8] Y. She et al., “Phát triển và xác thực mô hình học sâu cho sự sống sót của bệnh ung thư phổi không phải tế bào nhỏ”, JAMA Netw. Open, tập 3, số 6, trang e205842, tháng 6 năm 2020, doi: 10.1001/jamanetworkopen.2020.5842.

[9] D. Deb, AC Moore và UB Roy, “Bối cảnh điều trị ung thư phổi toàn cầu năm 2021,” J. Thorac. Oncol., tập 17, số 7, trang 931–936, tháng 7 năm 2022, doi: 10.1016/j.jtho.2022.03.018.

[10] A. Hosny và cộng sự, “Học sâu để tiên lượng ung thư phổi: Nghiên cứu hình ảnh học đa nhóm hồi cứu”, PLOS Med., tập 15, số 11, trang e1002711, tháng 11 năm 2018, doi: 10.1371/journal.pmed.1002711.

[11] VK Raghu và cộng sự, “Xác thực mô hình dựa trên học sâu để dự đoán nguy cơ ung thư phổi bằng cách sử dụng X-quang ngực và dữ liệu hồ sơ y tế điện tử”, JAMA Netw. Open, tập 5, số 12, trang e2248793, tháng 12 năm 2022, doi: 10.1001/jamanetworkopen.2022.48793.

[12] K.-H. Yu và cộng sự, “Phương pháp học máy có thể tái tạo để phát hiện ung thư phổi bằng hình ảnh chụp cắt lớp vi tính: Phát triển và xác thực thuật toán”, J. Med. Internet Res., tập 22, số 8, trang e16709, tháng 8 năm 2020, doi: 10.2196/16709.

[13] P. Afshar et al., “$$\text {DRTOP}$$: radiomics dựa trên học sâu để dự đoán kết quả thời gian đến sự kiện ở ung thư phổi,” Sci. Rep., tập 10, số 1, trang 12366, tháng 7 năm 2020, doi: 10.1038/s41598-020-69106-8.

[14] Y. Wu et al., “Tuổi già và tình trạng đột biến <scp> EGFR </scp> ở những bệnh nhân ung thư phổi không phải tế bào nhỏ giai đoạn đầu không thể phẫu thuật đang được xạ trị cắt bỏ định vị: Kinh nghiệm của một viện duy nhất trên 71 bệnh nhân ở Đài Loan,” Thorac. Cancer, tập 14, số 7, trang 654–661, tháng 3 năm 2023, doi: 10.1111/1759-7714.14786.

[15] S. Doppalapudi, RG Qiu và Y. Badr, “Dự đoán và hiểu biết về thời gian sống sót của bệnh ung thư phổi: Các phương pháp học sâu”, Int. J. Med. Inform., tập 148, trang 104371, tháng 4 năm 2021, doi: 10.1016/j.ijmedinf.2020.104371.

[16] Rajesh N., A. Irudayasamy, MSK Mohideen và CP Ranjith, “Phân loại các hội chứng di truyền quan trọng liên quan đến bệnh tiểu đường bằng cách sử dụng phương pháp CapsNet dựa trên ANN,” Int. J. e-Collaboration, tập 18, số 3, trang 1–18, tháng 8 năm 2022, doi: 10.4018/IJeC.307133.

[17] Y. Li, X. Wu, P. Yang, G. Jiang, và Y. Luo, “Học máy để chẩn đoán, điều trị và tiên lượng ung thư phổi”, Genomics. Proteomics Bioinformatics, tập 20, số 5, trang 850–866, tháng 10 năm 2022, doi: 10.1016/j.gpb.2022.11.003.

[18] C. Anil Kumar và cộng sự, “Dự đoán ung thư phổi từ bộ dữ liệu văn bản bằng cách sử dụng máy học,” Biomed Res. Int., tập 2022, trang 1–10, tháng 7 năm 2022, doi: 10.1155/2022/6254177.

[19] E. Dritsas và M. Trigka, “Dự đoán rủi ro đột quỵ bằng các kỹ thuật học máy”, Sensors, tập 22, số 13, trang 4670, tháng 6 năm 2022, doi: 10.3390/s22134670.

[20] T. Tazin, MN Alam, NN Dola, MS Bari, S. Bourouis và M. Monirujjaman Khan, “Phát hiện và dự đoán bệnh đột quỵ bằng cách sử dụng các phương pháp học tập mạnh mẽ”, J. Healthc. Eng., tập. 2021, trang 1–12, tháng 11 năm 2021, doi: 10.1155/2021/7633381.

[21] G. Sailasya và GLA Kumari, “Phân tích hiệu suất dự đoán đột quỵ bằng thuật toán phân loại ML,” Int. J. Adv. Comput. Sci. Appl., tập 12, số 6, 2021, doi: 10.14569/IJACSA.2021.0120662.

TIỂU SỬ CÁC TÁC GIẢ

Kumar Mohan là Giảng viên giàu kinh nghiệm của Đại học Công nghệ và Khoa học ứng dụng – Shinas, Oman. Ông có kinh nghiệm giảng dạy và nghiên cứu hơn 17 năm. Ông đã hoàn thành chương trình Thạc sĩ Khoa học máy tính tại Đại học Sathyabama vào năm 2007 và hoàn thành chương trình Cử nhân Khoa học máy tính và kỹ thuật tại Đại học Periyar vào năm 2003. Ông đã nhận được 3 dự án được tài trợ từ URG và MOHERI Oman. Ông đã xuất bản nhiều bài báo và bài đánh giá cho nhiều Tạp chí uy tín. Bạn có thể liên hệ với ông qua email: kumar.mohan@shct.edu.om.

Bhraguram TM nhận bằng Tiến sĩ từ Đại học Kanpur, Ấn Độ và bằng Thạc sĩ Khoa học Máy tính và Kỹ thuật (CSE) từ Đại học Bharathidasan, Ấn Độ. Ông đã hoàn thành bằng Thạc sĩ Khoa học Máy tính và Kỹ thuật (CSE) từ trường Cao đẳng Kỹ thuật VMKV Tamilnadu, Ấn Độ và chuyên ngành An ninh mạng từ trường Cao đẳng Luật Chính phủ, Mumbai. Ông có bằng Thạc sĩ Kinh doanh chuyên ngành Kinh doanh Điện tử. Sự nghiệp lừng lẫy của ông kéo dài 15 năm học thuật và một năm kinh nghiệm làm việc tại công ty. Ông đã xuất bản hơn 19 bài báo bao gồm Scopus & Web of Science (WoS). Ông là thành viên của nhiều tổ chức chuyên môn nổi tiếng, cụ thể là IEEE, ACM, ISTE, IACSIT, CSTA, IAENGG, IAHFP và IARCP. Ông nắm giữ 3 bằng sáng chế đăng ký trong nhiều lĩnh vực khác nhau và mối quan tâm và lĩnh vực nghiên cứu của ông là Khai thác dữ liệu, Dịch vụ dựa trên đám mây, Pháp y kỹ thuật số, Metaverse và Chuỗi khối. Ông đã giành được nhiều giải thưởng và danh hiệu trong suốt sự nghiệp của mình và hiện đang làm giảng viên tại Đại học Công nghệ và Khoa học ứng dụng, Shinas, Vương quốc Hồi giáo Oman. Bạn có thể liên hệ với ông qua email: bhraguram.thayyil@shct.edu.om.