XÂY DỰNG BÀI TOÁN CHẨN ĐOÁN UNG THƯ CỔ TỬ CUNG SỬ DỤNG MÔ HÌNH LOGIT VỚI DEEP LEARNING

Lê Ngọc Hiếu[*], **Võ Phạm Huyền Khanh**[*], **Trần Công Hùng**[**]
[*] Trường Đại Học Mở Thành Phố Hồ Chí Minh
[**] Học Viện Công Nghệ Bưu Chính Viễn Thông Cơ Sở Thành Phố Hồ Chí Minh

Tóm tắt: Theo Global Cancer Observatory, ung thư cổ tử cung đã gây ra 604.127 trường hợp mắc mới và 341.831 trường hợp tử vong trên toàn cầu vào năm 2020 [1]. Bên cạnh đó, số lượng bệnh nhân không có dấu hiệu giảm trong những năm gần đây [1] [2]. Đa số người bệnh đến khám và điều trị khi bệnh đã ở giai đoạn muộn, lúc này người bệnh phải thực hiện cắt bỏ một phần hoặc thậm chí toàn bộ tử cung. Việc đưa ra những cảnh báo sớm về khả năng mắc ung thư cổ tử cung là việc làm thiết thực, giúp người bệnh có cái nhìn cụ thể về tình trạng sức khỏe của bản thân để có những hành động đúng đắn, hỗ trợ cải thiện sức khỏe hoặc điều trị sớm hơn. Nghiên cứu này là nghiên cứu tiếp tục dựa trên các nghiên cứu phát hiện ung thư cổ tử cung bằng cách sử dụng học máy chủ yếu là học sâu để giúp phát hiện ung thư cổ tử cung thông qua dữ liệu bệnh nhân đã được cung cấp. Với việc áp dụng phương pháp tiếp cận mới của một số công nghệ trí tuệ nhân tạo, ở đây là các kỹ thuật học sâu, phát hiện bệnh ung thư với dữ liệu được cung cấp của bệnh nhân. Dữ liệu đầu vào là các yếu tố nguy cơ của ung thư cổ tử cung và chúng tôi sử dụng mô hình được xây dựng bởi tập dữ liệu đào tạo lấy từ kho lưu trữ Học máy UCI [3]. Kết quả thực nghiệm với độ chính xác của mô hình là 94.18% so với kết quả chẩn đoán thực tế. Điều này có thể cho chúng ta thấy dấu hiệu tích cực về chẩn đoán ung thư cổ tử cung thông qua học máy, cho thấy trí tuệ nhân tạo có tiềm năng rất lớn để giúp bệnh nhân dự đoán và ngăn ngừa ung thư trước khi mắc phải.

Từ khóa: Chẩn đoán ung thư, ung thư cổ tử cung, chẩn đoán ung thư cổ tử cung, học máy, học sâu, trí tuệ nhân tạo.

I. GIỚI THIỆU

Ung thư [4] là một trong những căn bệnh gây tử vong hàng đầu trên toàn thế giới, với khoảng 10 triệu ca tử vong mỗi năm. Theo thống kê của GLOBOCAN [1], trên thế giới có khoảng 1/6 số ca tử vong do ung thư vào năm 2020. Hơn nữa tình trạng ung thư ở các nước có thu nhập thấp và trung bình rất nghiêm trọng. Cụ thể, có tới 70% số ca tử vong do ung thư xảy ra tại đây. Bệnh biểu hiện ở giai đoạn muộn và thiếu khả năng tiếp cận với chẩn đoán và điều trị là rất phổ biến. Kết quả từ báo cáo Global Survey 2019 [5] cho thấy điều trị toàn diện đạt mức hơn 90% các quốc gia có thu nhập cao nhưng lại dưới 15% các quốc gia có thu nhập thấp.

Ung thư cổ tử cung [6] phát triển trong cổ tử cung của phụ nữ (lối vào tử cung từ âm đạo). Đây là bệnh ung thư xếp thứ 4 về số ca mắc mới và thứ 6 về số ca tử vong ở phụ nữ trên toàn thế giới [1]. Tuy độ phổ biến không cao bằng ung thư vú và ung thư phổi, nhưng ung thư cổ tử cung rất nguy hiểm vì nó ảnh hưởng đến hệ thống sinh sản của phụ nữ. Phụ nữ ở độ tuổi hay mắc phải ung thư cổ tử cung là từ 30 trở lên, trung bình là 48-52 tuổi [6]. Mặc dù bệnh gây tổn thương lớn đến tử cung nhưng vì bệnh tiến triển âm thầm trong thời gian dài (từ 5 đến 20 năm) và các triệu chứng lại khá mờ nhạt, dễ gây nhầm lẫn với các bệnh phụ khoa khác nên rất khó để phát hiện khi bệnh ở giai đoạn đầu. Ở giai đoạn muộn, bệnh nhân phải tiến hành cắt bỏ đi một phần hay toàn bộ tử cung, điều này ảnh hưởng trực tiếp đến cơ quan sinh sản và thiên chức làm mẹ của người phụ nữ.

Đồng thời, trí tuệ nhân tạo [7] đã và đang đạt được rất nhiều sức hút và thành tựu trong nhiều lĩnh vực. Đặc biệt, trong lĩnh vực y tế, trí tuệ nhân tạo được áp dụng vào chẩn đoán bệnh và quản lý các loại vấn đề sức khỏe ngày càng nhiều. Những nghiên cứu ứng dụng trí tuệ nhân tạo, đặc biệt là các kỹ thuật học máy, học sâu trong chẩn đoán bệnh ung thư rất được ưa chuộng trong thời gian gần đây. Các nghiên cứu về chuẩn đoán ung thư cổ tử cung thông qua học máy, đa số tập trung nghiên cứu trên bộ dữ liệu về các yếu tố nguy cơ của bệnh ung thư cổ tử cung được tập hợp tại Bệnh viện Universitario de Caracas ở Caracas, ở Venezuela [8]. Khi nghiên cứu về tập dữ liệu này, chúng tôi nhận thấy bộ dữ liệu này thuộc loại không cân bằng và có độ lệch tương đối lớn. Đây là trường hợp thường thấy của những bài toán phân loại bệnh khi số lượng người được chẩn đoán là mắc bệnh khá ít ỏi so với số lượng người không mắc bệnh, đặc biệt là nhóm bệnh ung thư. Là một hướng nghiên cứu chuyên sâu của học máy, học sâu [9] tập trung giải quyết các vấn đề liên quan sử dụng mạng thần kinh nhân tạo để xây dựng mô hình dự đoán. Keras [10] là một thư viện được viết bởi ngôn ngữ Python hỗ trợ tốt về xây dựng các mô hình học sâu. Ta có thể kết hợp Keras với các thư viện học sâu trong việc xử lý tập dữ liệu không cân

Tác giả liên hệ: Trần Công Hùng,
Email: conghung@ptithcm.edu.vn
Đến tòa soạn: 25/5/2021, chỉnh sửa: 25/6/2021, chấp nhận đăng: 6/7/2021.

bằng [11] [12]. Dựa trên học sâu – mô hình logit được xây dựng từ thuật toán Logistic Regression – thuật toán phân loại nhị phân (binary classification) sẽ được sử dụng trong nghiên cứu này với mục đích là chẩn đoán một người có mắc ung thư cổ tử cung hay không.

Từ những lý do đó, bài báo này đề xuất, nghiên cứu chẩn đoán bệnh ung thư cổ tử cung thông qua các yếu tố nguy cơ của bệnh bằng mô hình logit, xây dựng mô hình chuẩn đoán bằng học sâu với Keras, từ đó giúp mọi người phát hiện sớm nguy cơ mắc hoặc mắc ung thư cổ tử cung ở giai đoạn đầu của bệnh để có phương pháp điều trị kịp thời.

Bài báo này được cấu trúc với 6 phần. Phần 1 giới thiệu thực trạng về ung thư cổ tử cung và phương pháp học máy được sử dụng trong nghiên cứu này. Phần 2 trình bày một số công trình liên quan mật thiết tới nghiên cứu này. Phần 3 là đề xuất phương pháp xử lý dữ liệu không cân bằng với mô hình Logit thông qua kỹ thuật của Keras. Phần 4 trình bày về tập dữ liệu và các phương pháp xử lý bộ dữ liệu trước khi xây dựng mô hình. Phần 5 trình bày các kết quả thực nghiệm. Phần 6 là kết luận chung của đề tài nghiên cứu.

II. CÁC CÔNG TRÌNH LIÊN QUAN

Trong giai đoạn gần đây, đã có rất nhiều nghiên cứu ứng dụng trí tuệ nhân tạo trong việc chẩn đoán bệnh đạt được độ chính xác và độ tin cậy cao. Học máy – một hướng chuyên sâu của trí tuệ nhân tạo và những thuật toán của nó đã được sử dụng rất nhiều trong việc xây dựng các mô hình phục vụ cho các bài toán nghiên cứu. Các thuật toán học máy được sử dụng trong các nghiên cứu gần đây là tiền đề và nền tảng tạo cảm hứng cho nghiên cứu này, là xây dựng mô hình nghiên cứu của bài báo này trong việc chẩn đoán ung thư cổ tử cung.

Muhammad Fazal Ijaz và cộng sự [13] vào năm 2020 đã đề xuất Mô hình dự báo ung thư cổ tử cung (Cervical Cancer Prediction Model - CCPM). Các tác giả đã sử dụng bộ dữ liệu yếu tố nguy cơ ung thư cổ tử cung từ Đại học California tại Irvine (UCI) bao gồm 36 thuộc tính (chứa bốn biến mục tiêu- Hinselmann, Schiller, Cytology, và Biopsy– và 32 yếu tố nguy cơ) của 858 bệnh nhân. Đầu tiên, các tác giả loại bỏ các giá trị ngoại lệ bằng cách sử dụng các phương pháp phát hiện ngoại lệ như Density-Based Spatial Clustering of Applications with Noise (DBSCAN) và Isolation Forest (iForest), tiếp theo là giải quyết vấn đề mất cân bằng dữ liệu thông qua Synthetic Minority Over-Sampling Technique (SMOTE) và SMOTE với liên kết Tomek (SMOTE Tomek). Cuối cùng, sử dụng Random Forest (RF) để phân lớp dữ liệu, chuẩn đoán bệnh. Do đó, CCPM được xây dựng dựa trên 4 trường hợp: (1) DBSCAN + SMOTETomek + RF, (2) DBSCAN + SMOTE + RF, (3) iForest + SMOTETomek + RF, và (4) iForest + SMOTE + RF. Sau khi quan sát, các tác giả nhận thấy rằng RF hoạt động tốt nhất so với các mô hình còn lại. Hơn nữa, CCPM được đề xuất cho thấy độ chính xác tốt hơn so với các phương pháp được đề xuất trước đây về dự đoán ung thư cổ tử cung trên cùng bộ dữ liệu.

Nghiên cứu của Ahmed Mostafa Khalil và cộng sự [14] vào năm 2020, về việc phát triển một hệ thống mới sử dụng hệ chuyên gia mờ để dự đoán ung thư phổi. Quá trình dự đoán sử dụng hệ thống chuyên gia mờ này trải qua bốn bước, với dữ liệu thu được từ Khoa Hô hấp Bệnh viện Ngực Nam Kinh ở Trung Quốc chứa các triệu chứng quan trọng nhất của ung thư phổi. Sau khi chạy kiểm tra cho thấy độ chính xác đạt 100%. Đây là một hướng nghiên cứu phát hiện ung thư phổi bằng hệ chuyên gia mờ và đạt được tỉ lệ chính xác tối đa.

Riham Alsmariy và cộng sự [15] vào năm 2020 đã đề xuất nghiên cứu sử dụng thuật toán học máy để tìm ra một mô hình có khả năng chẩn đoán ung thư với độ chính xác và độ nhạy cao. Bài báo đã sử dụng bộ dữ liệu các yếu tố ung thư cổ tử cung [8] làm đầu vào để xây dựng mô hình phân loại thông qua phương pháp bỏ phiếu kết hợp với ba bộ phân loại: Cây quyết định (Decision tree), Hồi quy logistic (Logistic regression) và Rừng ngẫu nhiên (Random forest). Thông qua phương pháp bỏ phiếu Synthetic Minority Oversampling Technique (SMOTE), vấn đề mất cân bằng tập dữ liệu được giải quyết, sau đó sử dụng phương pháp Principal Component Analysis (PCA) để tăng hiệu suất mô hình và cuối cùng là kỹ thuật 10-fold cross-validation để ngăn chặn các vấn đề liên quan overfitting. Sau khi kết hợp các phương pháp trên, đường đặc tính (ROC_AUC) của bốn mô hình dự đoán cho từng biến mục tiêu cho tỷ lệ cao hơn, độ chính xác, độ nhạy và tỷ lệ độ chính xác thực của dự báo được cải thiện từ 0,93% lên 5,13%, 39,26% lên 46,97% và 2% lên 29 %, tương ứng cho tất cả các biến mục tiêu.

Cũng sử dụng phương pháp bỏ phiếu SMOTE và kỹ thuật 10-fold cross-validation, vào năm 2019, Talha Mahboob Alam và cộng sự [16] đã đề xuất nghiên cứu về dự đoán ung thư cổ tử cung bằng cách sử dụng các thuật toán phân loại: Cây quyết định tăng cường (Boosted decision tree), Rừng quyết định (Decision forest), và Rừng nhiệt đới quyết định (Decision jungle) trên cùng một tập dữ liệu [8] nhưng đã được cân bằng. Quá trình tiền xử lý dữ liệu trải qua năm bước. Bộ dữ liệu mới bao gồm 32 thuộc tính và 1226 bệnh nhân với 563 bệnh nhân ung thư, 663 bệnh nhân không ung thư. Nghiên cứu sử dụng bốn phương pháp sàng lọc là bốn thuộc tính đích Hinselmann, Schiller, Cytology, và Biopsy để chẩn đoán ung thư. Sau khi chạy mô hình, kết quả dự đoán với thuật toán Boosted decision tree có độ chính xác lên đến 97,8% trên đường cong AUROC với phương pháp sàng lọc Hinselmann. Với bài báo này, các tác giả đã công bố một hướng xử lý cân bằng và sử dụng các đặc tính nổi bật của ung thư để xây dựng mô hình, bỏ qua được sai số của các yếu tố không liên quan do vậy độ chính xác khá cao.

Bài báo công bố của Abisoye Blessing và cộng sự [17] vào năm 2019, đã đề xuất một nghiên cứu dự đoán ung thư cổ tử cung bằng cách kết hợp Giải thuật di truyền (Genetic Algorithm - GA) và Support Vector Machine (SVM). Bộ dữ liệu ung thư cổ tử cung được xử lý để loại bỏ nhiễu, sử dụng kỹ thuật chuẩn hóa Min-Max để tránh các vấn đề về xử lý số. Sau đó, từ tập dữ liệu các yếu tố nguy cơ ung thư cổ tử cung, các trường dữ liệu phù hợp được chọn thông qua phương pháp GA. Cuối cùng, các tác giả sử dụng SVM để huấn luyện dữ liệu, sử dụng dữ liệu đã chuẩn bị để phân loại ung thư cổ tử cung. Kết quả độ chính xác khi chạy mô hình phân loại với thuộc tính Biopsy là 95%. Với kết quả

95% của nghiên cứu này, ta thấy sự kết hợp khá hài hòa giữa thuật toán di truyền (GA) để xử lý dữ liệu trước khi đưa vào xây dựng mô hình phân lớp chuẩn đoán bệnh với SVM.

Một nghiên cứu khác đã được công bố do tác giả B. Nithya1 và V. Ilango ở Ấn Độ vào năm 2019 [18]. Trong nghiên cứu này, các tác giả ứng dụng học máy bằng cách áp dụng các kỹ thuật học máy trong R để phân tích các yếu tố nguy cơ của ung thư cổ tử cung. Ngoài ra, bài viết này xây dựng một số mô hình phân loại sử dụng các thuật toán bao gồm: C5.0, Rừng Ngẫu Nhiên (Random Forest), rPast, K láng giềng gần nhất (k-Nearest Neighbor) và Máy vectơ hỗ trợ (Support Vector Machine). Kỹ thuật lựa chọn các đặc trưng được sử dụng trong quá trình tiền xử lý dữ liệu này là phương pháp wrapper để tìm kiếm các mô hình dữ liệu chính xác. Các phương pháp phân loại được xây dựng với k-fold cross-validation cùng 26 thuộc tính của tập dữ liệu sau khi làm sạch. Kết quả độ chính xác của mô hình phân loại thông qua hai thuật toán C5.0 và Rừng Ngẫu Nhiên với một số thuộc tính quan trọng đã chọn lọc được nâng cấp đáng kể từ 99% lên 100%. Trong nghiên cứu này, các tác giả đã thử nghiệm các mô hình học máy điển hình lần lượt với bộ dữ liệu ung thư cổ tử cung, và cách lựa chọn các thuộc tính phù hợp với từng mô hình. Mô hình được xây dựng từ thuật toán C5.0 và Random Forest là có kết quả với độ chính xác cao nhất.

Bài báo của Shanjida Khan Maliha và cộng sự [19] vào năm 2019 đã sử dụng các thuật toán Naive Bayes, K láng giềng gần nhất (k-Nearest Neighbor) và J48 kết hợp với phương pháp 10-fold cross-validation để dự đoán bệnh ung thư. Dữ liệu của bệnh nhân đang trải qua căn bệnh được sử dụng trong bài nghiên cứu đã được cung cấp bởi các bác sĩ tham gia đề tài này. Tập dữ liệu này chứa 61 thuộc tính về các triệu chứng và một số thuộc tính về kết quả xét nghiệm của bệnh ung thư, và 1 thuộc tính đại diện cho các loại bệnh ung thư của 1059 bệnh nhân. Công cụ Weka được sử dụng với mục đích đo lường độ chính xác của bộ dữ liệu bệnh ung thư, bao gồm 9 loại bệnh ung thư: ung thư não, ung thư máu, ung thư tuyến tuỵ, ung thư tuyến tiền liệt, ung thư buồng trứng, ung thư vú, ung thư thực quản, ung thư phổi và ung thư đại trực tràng. Kết quả sau khi chạy mô hình cho thấy độ chính xác của mô hình với các thuật toán Naive Bayes, k-NN và J48 lần lượt là 98,2%, 98,8% và 98,5%. Cả 3 thuật toán áp dụng đều có độ chính xác lên tới 98%, điều này cho thấy bộ dữ liệu do các bác sĩ tham gia xây dựng và cung cấp khá chuẩn xác và đáng tin cậy.

Sau khi nghiên cứu các một số công trình liên quan và điển hình là các công trình vừa khảo sát ở trên, chúng tôi nhận thấy việc sử dụng Keras với mô hình logit để xây dựng mô hình dự đoán ung thư cổ tử cung là khá tiềm năng và hợp lý. Bộ dữ liệu ung thư cổ tử cung [8] là không cân bằng và cần phải xử lý dữ liệu này trước khi xây dựng mô hình dự đoán bệnh. Việc xây dựng mô hình logit với Keras trên bộ dữ liệu không cân bằng này là rất khả thi và hiện tại cũng chưa có nhiều công bố nào áp dụng phương pháp này. Do đó, để nghiên cứu thêm về dự đoán bệnh ung thư cổ tử cung trên bộ dữ liệu UCI này, chúng tôi xin đề xuất, áp dụng mô hình logit với Keras vào nghiên cứu này.

III. DỮ LIỆU CỦA BÀI TOÁN

A. Bộ dữ liệu các yếu tố nguy cơ ung thư cổ tử cung

Bộ dữ liệu về các yếu tố nguy cơ ung thư cổ tử cung sử dụng trong nghiên cứu này được cung cấp công khai từ Trường đại học California, Irvine (UCI) Machine Learning Repository [3], đây là cổng thông tin cung cấp các bộ dữ liệu, lý thuyết miền và trình tạo dữ liệu được cộng đồng học máy sử dụng để thực nghiệm nghiên cứu phát triển các thuật toán học máy.

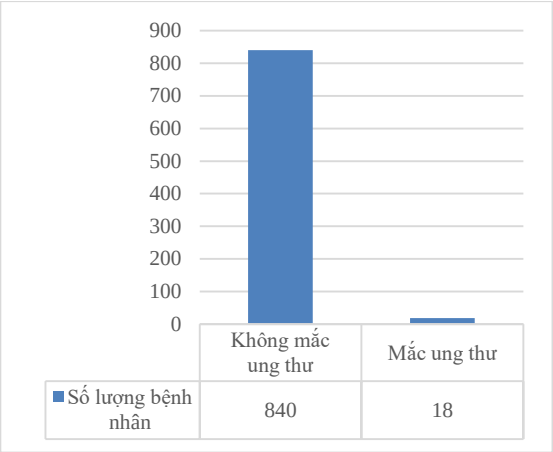
Bộ dữ liệu này được thu thập tại Bệnh viện Universitario de Caracas ở Caracas, Venezuela vào năm 2017. Nó tập trung vào việc chẩn đoán ung thư cổ tử cung cho 858 bệnh nhân. Vì một số bệnh nhân quyết định không trả lời một số câu hỏi do lo ngại về quyền riêng tư nên bộ dữ liệu có khá nhiều trường bị thiếu dữ liệu và không hoàn chỉnh. Mô tả và kiểu dữ liệu của 36 thuộc tính được trình bày trong bảng 1.

BẢNG 1. MÔ TẢ CÁC TRƯỜNG DỮ LIỆU TRONG BỘ DỮ LIỆU CÁC YẾU TỐ UNG THƯ CỔ TỬ CUNG

TT	Tên thuộc tính	Kiểu DL	SL quan sát NULL
1	Age	Int	0
2	Number of sexual partners	Int	26
3	First sexual intercourse	Int	7
4	Num of pregnancies	Int	56
5	Smokes	Bool	13
6	Smokes (years)	Float	13
7	Smokes (packs/year)	Int	13
8	Hormonal Contraceptives	Bool	108
9	Hormonal Contraceptives (years)	Float	108
10	IUD	Bool	117
11	IUD (years)	Float	117
12	STDs	Bool	105
13	STDs (number)	Int	105
14	STDs:condylomatosis	Bool	105
15	STDs:cervical condylomatosis	Bool	105
16	STDs:vaginal condylomatosis	Bool	105
17	STDs:vulvo-perineal condylomatosis	Bool	105
18	STDs:syphilis	Bool	105
19	STDs:pelvic inflammatory disease	Bool	105
20	STDs:genital herpes	Bool	105
21	STDs:molluscum contagiosum	Bool	105
22	STDs:AIDS	Bool	105
23	STDs:HIV	Bool	105
24	STDs:Hepatitis B	Bool	105
25	STDs:HPV	Bool	105
26	STDs: Number of diagnosis	Int	0
27	STDs: Time since first diagnosis	Int	787
28	STDs: Time since last diagnosis	Int	787
29	Dx:Cancer	Bool	0
30	Dx:CIN	Bool	0
31	Dx:HPV	Bool	0
32	Dx	Bool	0
33	Hinselmann	Bool	0
34	Schiller	Bool	0
35	Cytology	Bool	0
36	Biopsy	Bool	0

B. Các vấn đề của bộ dữ liệu

Bộ dữ liệu được sử dụng trong bài báo là một bộ dữ liệu nhỏ và không cân bằng. Thuộc tính ***Dx:Cancer*** được sử dụng để phân loại với số lượng người mắc bệnh rất nhỏ, số liệu được trình bày trong biểu đồ bên dưới.



Hình 1. Thống kê số lượng bệnh nhân mắc ung thư trong bộ dữ liệu nghiên cứu

Với mẫu có kích thước 858, chỉ có 18 người được chẩn đoán mắc ung thư, tương đương với 2.1%. Sự mất cân bằng này khá lớn, điều này sẽ ảnh hưởng đến chất lượng dự đoán của mô hình.

Thêm vào đó, Bảng 1 cũng đã chỉ ra số lượng giá trị bị thiếu của bộ dữ liệu là rất cao (có tới 26 trên 36 thuộc tính bị thiếu dữ liệu), với 2 thuộc tính bị thiếu tới 787 giá trị và 18 thuộc tính bị thiếu trên 100 giá trị.

IV. PHƯƠNG PHÁP NGHIÊN CỨU

Việc lựa chọn các phương pháp và thuật toán phù hợp cho tập dữ liệu là quan trọng và cần thiết để xây dựng một mô hình hiệu quả và chính xác. Phần này mô tả cơ sở lý thuyết liên quan tới các phương pháp phân loại và thuật toán được áp dụng để xây dựng mô hình dự đoán bệnh trong bài báo này.

A. Cơ sở lý thuyết

1) Mô hình Logit trong học sâu

Mô hình Logit được xây dựng từ thuật toán Hồi quy Logistic (Logistic Regression) thuộc nhóm Classification. Đây là một thuật toán phân loại nhị phân (binary classification) được sử dụng phổ biến. Cụ thể chúng tôi sẽ áp dụng mô hình này vào việc chẩn đoán bệnh nhân mắc và không mắc ung thư cổ tử cung.

Có nhiều thông số [20] ta có thể được sử dụng để đo lường hiệu suất của mô hình phân lớp nhị phân (binary classification); các trường khác nhau có các tùy chọn khác nhau cho các số liệu cụ thể do các mục tiêu khác nhau. Trong y học, độ nhạy (sensitivity) và độ đặc trưng (specificity) thường được sử dụng, trong khi độ chính xác truy xuất thông tin thường được thể hiện thông qua 2 chỉ số precision và recall. Một điểm khác biệt quan trọng là giữa các chỉ số không phụ thuộc vào tần suất xuất hiện của mỗi loại trong tổng thể (tỷ lệ prevalence) và các chỉ số phụ thuộc vào tỷ lệ này - cả hai nhóm đều hữu ích, nhưng chúng có các thuộc tính rất khác nhau.

Đưa ra phân loại của một tập dữ liệu cụ thể, có bốn kết hợp cơ bản giữa danh mục dữ liệu thực tế và danh mục được chỉ định: dương tính thực TP (True Positive – những lớp biểu hiện 1 là đúng), âm tính thực TN (True Negative – những lớp biểu hiện 0 là đúng), dương tính giả FP (False Positive – những lớp biểu hiện 1 là sai) và âm tính giả FN (False Negative – những lớp biểu hiện 0 là sai).

Assigned / Actual	Test outcome *positive*	Test outcome *negative*
Condition positive	True *positive*	False *negative*
Condition negative	False *positive*	True *negative*

Hình 2. Kết quả đánh giá của một mô hình phân lớp nhị phân

Chúng có thể được sắp xếp thành một bảng như hình 1, với các cột tương ứng với giá trị thực - điều kiện dương tính hoặc điều kiện âm tính - và các hàng tương ứng với giá trị phân loại - kết quả chạy mô hình dương tính hoặc kết quả chạy mô hình âm tính.

2) Dữ liệu không cân bằng

Sự mất cân bằng dữ liệu [21] là một trong những hiện tượng rất phổ biến của bài toán phân loại nhị phân. Nếu tỷ lệ dữ liệu giữa hai lớp là 50:50 thì dữ liệu đó được xem là cân bằng. Đến khi sự cách biệt giữa hai lớp là 60:40 thì dữ liệu đó xuất hiện hiện tượng mất cân bằng, nhưng những trường hợp này ảnh hưởng không đáng kể tới khả năng dự báo của mô hình. Tuy nhiên nếu lớn hơn tỷ lệ này thì độ chính xác của mô hình không còn đáng tin cậy nữa. Đặc biệt với các trường hợp tỷ lệ đạt mức 90:10 thì được xem là mất cân bằng nặng. Lúc này, độ chính xác (accuracy) có thể đạt được rất cao mà không cần tới mô hình. Vì thế độ chính xác không còn đủ tin cậy để đánh giá mô hình vào lúc này.

Ngoài ra, việc mất cân bằng dữ liệu nặng thường dẫn tới dự báo kém chính xác trên nhóm thiểu số. Do hầu hết kết quả dự báo thường nghiêng về nhóm đa số và rất ít trên nhóm thiểu số. Trong khi việc dự báo được chính xác một mẫu thuộc nhóm thiểu số quan trọng hơn nhiều so với dự báo mẫu thuộc nhóm đa số. Để cải thiện kết quả dự báo ta cần những điều chỉnh thích hợp để mô hình đạt được độ chính xác cao trên nhóm thiểu số. Lúc này để có thể đánh giá độ chính xác của mô hình một cách đúng đắn ta có thể cân nhắc tới một số chỉ số đánh giá thay thế khác như precision, recall, f1-score,... Hiện tượng overfitting [22] là một hiện tượng không mong muốn mà người xây dựng mô hình học máy thường gặp, đặc biệt với bộ dữ liệu không cân bằng. Đây là hiện tượng mô hình tìm được quá khớp (fit) với dữ liệu training. Việc quá khớp này có thể làm cho mô hình dự đoán nhầm nhiễu và chất lượng của mô hình không còn đạt kết quả tốt trên dữ liệu test nữa. Overfitting xảy ra khi mô hình quá phức tạp để mô phỏng dữ liệu training. Điều này đặc biệt xảy ra khi lượng dữ liệu training quá nhỏ trong khi độ phức tạp của mô hình quá cao. Hiện tượng này cũng giống như việc học tủ của con người, khi đó mô hình không học được gì từ dữ liệu training. Vì thế khi xây dựng mô hình ta sẽ hạn chế tối đa việc gặp overfitting bằng cách sử dụng những kỹ thuật khác nhau.

3) Keras tensorflow trong việc xử lý dữ liệu không cân bằng

Keras [10] là một thư viện được phát triển trên nền tảng ngôn ngữ Python. Keras có thể sử dụng chung với các thư viện nổi tiếng như Tensorflow, CNTK, Theano. Keras có cú pháp đơn giản hơn TensorFlow khá nhiều. Ta có thể kết hợp keras với các thư viện học sâu trong việc xử lý tập dữ liệu không cân bằng.

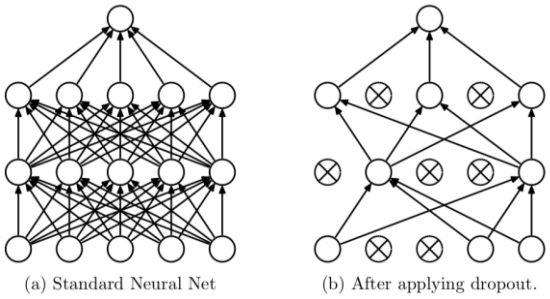
Trong TensorFlow và Keras, ta có thể làm việc với các tập dữ liệu không cân bằng theo nhiều cách:

- **Giảm kích thước mẫu** (Undersampling hoặc Random Undersampling): là phương pháp làm giảm số lượng quan sát của nhóm đa số để nó trở nên cân bằng với nhóm thiểu số. Đây là một cách làm cân bằng mẫu một cách nhanh chóng, dễ dàng thực hiện. Tuy nhiên phương pháp sẽ làm kích thước mẫu bị giảm đi đáng kể.
- **Tăng kích thước mẫu** (Oversampling hoặc Random Oversampling): phương pháp này đối lập với giảm kích thước mẫu. Nghĩa là ta sẽ làm tăng kích thước của mẫu thuộc nhóm thiểu số. Phương pháp này sẽ cân bằng lại dữ liệu và thường khả quan hơn phương pháp giảm kích thước mẫu.
- **Áp dụng trọng số lớp** (class weights): bằng cách làm cho các lớp có số lượng dữ liệu cao hơn ít quan trọng hơn trong quá trình tối ưu hóa mô hình, có thể đạt được cân bằng lớp ở mức tối ưu hóa.
- **Làm việc với F1 score** thay vì độ chính xác (precision) và thu hồi (recall): bằng cách sử dụng một số liệu cố gắng tìm sự cân bằng giữa mức độ liên quan của tất cả các kết quả và số lượng kết quả có liên quan được tìm thấy, ta có thể giảm tác động của cân bằng lớp đối với mô hình của mình mà không cần loại bỏ nó.

Trong bài báo này chúng tôi sẽ sử dụng kỹ thuật đánh trọng số lớp cho mô hình phân loại để cải thiện mô hình dự đoán bệnh.

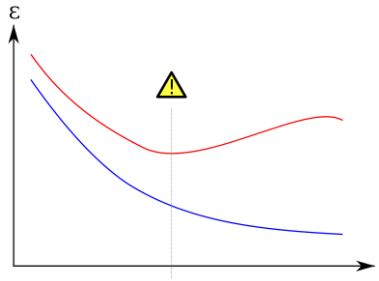
Regularization [22] nghĩa là thay đổi mô hình một chút để tránh overfitting trong khi vẫn giữ được tính tổng quát của nó (tính tổng quát là tính mô tả được nhiều dữ liệu, trong cả tập training và test). Một cách cụ thể hơn, ta sẽ tìm cách *di chuyển* nghiệm của bài toán tối ưu hàm mất mát tới một điểm gần nó. Hướng di chuyển sẽ là hướng làm cho mô hình *ít phức tạp hơn* mặc dù giá trị của hàm mất mát có tăng lên một chút. Điều này còn được hiểu là cách khắc phục độ phức tạp của một mô hình. Một số kỹ thuật của regularization bao gồm [23]:

- **L_1 regularization:** phạt các trọng số tương ứng với tổng các giá trị tuyệt đối của các trọng số. Trong các mô hình dựa trên sparse features, L_1 regularization giúp đẩy trọng số của các tính năng không liên quan hoặc hầu như không liên quan về chính xác 0, điều này sẽ xóa các tính năng đó khỏi mô hình.
- **L_2 regularization:** phạt các trọng số tương ứng với tổng bình phương của các trọng số. L_2 regularization giúp đẩy trọng số ngoại lệ (những giá trị dương tính cao hoặc âm tính thấp) về gần 0 nhưng không hoàn toàn về 0. (ngược lại với L_1 regularization) L_2 regularization luôn cải thiện tính tổng quát hóa trong các mô hình tuyến tính.
- **Dropout regularization:** Một hình thức regularization hữu ích trong việc đào tạo mạng nơ-ron. Dropout regularization hoạt động bằng cách loại bỏ ngẫu nhiên một số lượng cố định các nơ-ron được lựa chọn trong một lớp cho một bước gradient duy nhất. Càng nhiều nơ-ron được loại bỏ, regularization càng mạnh. Điều này tương tự như việc huấn luyện mạng để mô phỏng một nhóm lớn theo cấp số nhân của các mạng nhỏ hơn.



Hình 3. Mô hình mạng nơ-ron Drop. Bên trái: Một mạng nơ-ron tiêu chuẩn với 2 lớp hidden. Phải: Ví dụ về lưới mỏng được tạo ra bằng cách áp dụng tính năng dropout cho mạng ở bên trái. Các đơn vị bị gạch chéo đã bị loại bỏ [24]

- **Early stopping:** đây là phương pháp liên quan đến việc kết thúc việc đào tạo mô hình trước khi kết thúc đào tạo mất mát giảm. Trong Early stopping, ta kết thúc đào tạo mô hình khi tổn thất trên tập dữ liệu xác thực (validation dataset) bắt đầu tăng lên, tức là khi hiệu suất tổng quát hóa xấu đi.



Hình 4. Early Stopping với đường màu xanh là train error, đường màu đỏ là validation error. Trục x là số lượng vòng lặp, trục y là error. Mô hình được xác định tại vòng lặp mà validation error đạt giá trị nhỏ nhất [25]

Chúng tôi sẽ áp dụng hai phương pháp Dropout regularization và Early stopping để tránh gặp phải hiện tượng overfitting cho mô hình.

B. Phát biểu bài toán

Bộ dữ liệu về các yếu tố nguy cơ của bệnh ung thư cổ tử cung [8] là bộ dữ liệu thuộc loại không cân bằng (imbalanced data) và có độ lệch tương đối lớn. Với nghiên cứu về nhóm bệnh ung thư, thì bộ dữ liệu này là hợp lý và rất thường thấy.

Để mô hình xây dựng có độ chính xác và phù hợp tốt hơn, thì việc xử lý dữ liệu và tìm phương pháp phù hợp với bộ dữ liệu là cực kỳ quan trọng và cần thiết. Với mục tiêu giải quyết tình trạng không cân bằng trong bộ dữ liệu kết

hợp với học sâu, ta có rất nhiều phương pháp, cụ thể với bộ dữ liệu hiện tại thì nghiên cứu này xin đề xuất xử lý dữ liệu không cân bằng bằng kết hợp với mô hình logit (hay còn gọi là binary classification) mà bộ thư viện Keras hỗ trợ, kết hợp với thư viện của Tensorflow, sử dụng kỹ thuật để nâng cao độ chính xác của mô hình dự đoán là class weighting.

C. *Đề xuất phương pháp và thuật toán xử lý*

Quy trình hoạt động của phương pháp được đề xuất như sau:

1) Input: dữ liệu bệnh nhân

Đầu vào của mô hình là dữ liệu khám bệnh của một bệnh nhân bất kì. Dữ liệu này chạy qua mô hình dự báo để dự đoán ra bệnh nhân này có bệnh hay không. Mô hình này xây dựng từ bộ dữ liệu UCI [8] để dự đoán ung thư cổ tử cung. Các thông tin cần thiết của dữ liệu đầu vào chính là yếu tố nguy cơ gây ung thư cổ tử cung [6] như: tuổi tác, độ tuổi quan hệ tình dục đầu tiên, số lượng bạn tình, số lần mang thai, sử dụng thuốc lá, thời gian, số lượng thuốc lá đã sử dụng, sử dụng thuốc tránh thai, có mắc phải các bệnh lây lan qua đường tình dục,... Những thuộc tính này sẽ được mô tả một cách chi tiết và đầy đủ hơn vào phần 4 của bài báo.

2) Processing: xây dựng mô hình dự đoán bệnh

Quá trình xây dựng mô hình dự đoán bệnh ung thư cổ tử cung của bài báo được hiển thị theo sơ đồ dưới đây:

Hình 5. Các bước xây dựng mô hình dự đoán

Bước 1: Đầu tiên từ bộ dữ liệu các yếu tố nguy cơ của bệnh ung thư cổ tử cung [8], ta tiến hành quá trình xử lý dữ liệu như sau:

- Làm sạch dữ liệu: bộ dữ liệu có nhiều giá trị bị bỏ trống, trong trường hợp này ta sẽ tiến hành thống kê số lượng dữ liệu ấy sau đó chỉnh sửa và loại bỏ một số thuộc tính không quan trọng hoặc có giá trị null nhiều.
- Chia nhỏ dữ liệu: chia bộ dữ liệu thành các tập train, test, validation.
- Chuẩn hóa dữ liệu: sau khi chia nhỏ tập dữ liệu, tập train sẽ được chuẩn hóa bằng kỹ thuật StandardScaler. Kỹ thuật này sẽ đặt giá trị trung bình thành 0 và độ lệch chuẩn thành 1.

Bước 2: Sau khi hoàn thành quá trình xử lý dữ liệu, ta thu được bộ dữ liệu đầy đủ và chuẩn hóa. Với bộ dữ liệu mới này, ta sẽ tiến hành xây dựng mô hình dự đoán cho bài toán. Mô hình được xây dựng sau đó sẽ áp dụng phương pháp đánh trọng số lớp để gia tăng độ tin cậy của mô hình. Phương pháp này sẽ chuyển các trọng số Keras cho mỗi lớp thông qua một tham số với mục đích là giúp cho mô hình "chú ý nhiều hơn" đến các thành phần có tỷ lệ dương tính thấp.

Bước 3: dự đoán kết quả trên tập test và đánh giá độ chính xác của mô hình.

3) Output: kết luận là có bệnh hay không bệnh, với độ chính xác bao nhiêu phần trăm.

Sau khi có mô hình dự đoán bệnh đã xây dựng ở trên, ta có thể lấy thông tin bất kỳ của một bệnh nhân để làm đầu vào và đưa qua mô hình dự đoán bệnh. Sau khi nhập đầu đủ thông tin các thuộc tính cần, mô hình dự đoán sẽ xuất ra kết quả là bệnh nhân này có bệnh hay không và một số thống kê cơ bản. Đây là hướng phát triển tích hợp vào các hệ hỗ trợ khám chữa bệnh của bệnh viện & các bác sĩ.

D. *Xử lý dữ liệu*

Việc trích xuất thông tin có giá trị và kết quả phụ thuộc chủ yếu vào chất lượng của dữ liệu, do đó cần phải xử lý dữ liệu trước khi bắt đầu quá trình học máy để xây dựng mô hình. Vì lý do này, chúng tôi nhận thấy việc xử lý dữ liệu là quan trọng để cải thiện chất lượng cho mô hình. Trong nghiên cứu của chúng tôi, dữ liệu được xử lý theo những bước dưới đây.

1) Làm sạch dữ liệu

- Loại bỏ những thuộc tính mang số lượng giá trị mất mát cao bao gồm 2 thuộc tính: *STDs: Time since first diagnosis, STDs: Time since last diagnosis*.
- Loại bỏ những thuộc tính mang ý nghĩa chẩn đoán khác, bao gồm 7 thuộc tính: *Dx:CIN, Dx:HPV, Dx, Hinselmann, Schiller, Citology, Biopsy*.
- Xử lý các giá trị bị thiếu: Đối với những thuộc tính có kiểu luận lý (đúng, sai) sẽ được thay bằng giá trị "0" nếu không có dữ liệu. Đối với những thuộc tính có kiểu số thay bằng giá trị trung bình của thuộc tính đó.

Sau quá trình chọn lọc các thuộc tính bộ dữ liệu được sử dụng cho mô hình, thì bộ dữ liệu sau xử lý bao gồm 27 thuộc tính – với 26 thuộc tính là các yếu tố nguy cơ và 1 thuộc tính ***Dx:Cancer*** để phân loại.

2) Chia nhỏ dữ liệu

Bộ dữ liệu được chia ra thành 3 tập train, validation và test. Sử dụng tập validation là đặc biệt quan trọng với bộ dữ liệu không cân bằng vì vấn đề overfitting là mối quan tâm hàng đầu trong việc thiếu dữ liệu huấn luyện cho mô hình. Chúng tôi cũng đã sử dụng kỹ thuật *Early Stopping* của *Keras* để tránh việc huấn luyện quá nhiều cho mô hình để làm giảm hiện tượng overfitting.

Cụ thể với mẫu 858, bộ dữ liệu được chia ra làm 2 tập train và test – chiếm 20% của tập train tương ứng với 172 dòng dữ liệu. Tập train lớn này sau đó sẽ được chia ra làm 2 tập nữa là tập train chính thức và tập validation – cũng chiếm 20% của tập train này tương ứng với 138 dòng dữ liệu.

Từ 3 tập train, validation, test có tổng cộng 27 thuộc tính, được chia ra thành tập labels chứa lớp phân loại của mô hình chính là thuộc tính ***Dx:Cancer***, tập features chứa 26 thuộc tính về các yếu tố ung thư cổ tử cung. Kích thước cụ thể của các tập được miêu tả trong hình dưới đây:

```
Training labels shape: (548,)
Validation labels shape: (138,)
Test labels shape: (172,)
Training features shape: (548, 26)
Validation features shape: (138, 26)
Test features shape: (172, 26)
```

Hình 6. Kích thước của các tập train, validation, test

3) Chuẩn hóa dữ liệu

Chúng tôi đã sử dụng *StandardScaler* để chuẩn hóa các thuộc tính đầu vào. Điều này sẽ thiết lập giá trị trung bình thành 0 và độ lệnh chuẩn thành 1. Kỹ thuật này được Keras cung cấp và sử dụng dễ dàng.

V. KẾT QUẢ THỰC NGHIỆM

A. Mô hình cơ sở (Baseline model)

Baseline model [26], là mô hình đầu tiên của bài toán khi chưa áp dụng bất cứ kỹ thuật xử lý gì. Ta sẽ bắt đầu với việc xây dựng một baseline model, sau đó cố gắng đưa ra các giải pháp phức tạp hơn vào mô hình để đạt kết quả cao hơn so với baseline model. Nếu nhận được kết quả tốt hơn, nghĩa là mô hình của ta hoạt động hiệu quả hơn.

Tạo một baseline model bằng *Sequential* [27] (được cung cấp bởi Keras) với 3 lớp: dense, dropout (hidden layer), dense_1 (output). Lớp hidden được kết nối với input, lớp dropout được sử dụng để làm giảm overfitting cho mô hình và lớp output để trả về kết quả người bệnh có mắc ung thư. Tổng tham số của mô hình là 449, sử dụng epochs = 40 và batch_size = 50. Đồng thời chúng tôi cũng đã thiết lập *bias* để làm giảm sự mất mát của mô hình. Tất cả các kỹ thuật này đều được Keras hỗ trợ.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 16)                432
_________________________________________________________________
dropout (Dropout)            (None, 16)                0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 17
=================================================================
Total params: 449
Trainable params: 449
Non-trainable params: 0
```

Hình 7 Mô hình xây dựng với các lớp và tham số

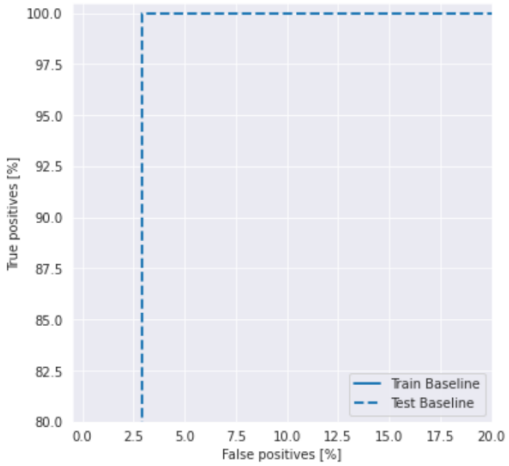
- AUC đề cập đến khu vực dưới đường cong (Area Under the Curve) của Receiver Operating Characteristic Curve (ROC-AUC) [28]. Chỉ số này bằng với xác suất mà bộ phân loại sẽ xếp hạng mẫu dương tính ngẫu nhiên cao hơn mẫu âm tính ngẫu nhiên.
- AUPRC đề cập đến khu vực dưới đường cong (Area Under the Curve) của Precision-Recall Curve [29]. Chỉ số này tính toán các cặp precision-recall cho các ngưỡng xác suất khác nhau.

Đánh giá mô hình trên tập dữ liệu test và hiển thị kết quả cho các thông số.

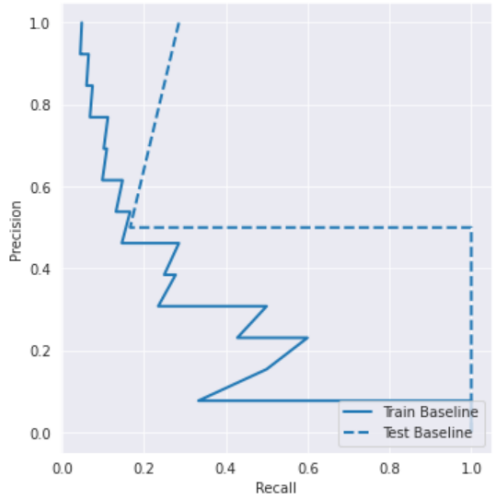
```
loss :  0.04956606775522232
tp :  0.0
fp :  0.0
tn :  170.0
fn :  2.0
accuracy :  0.9883720874786377
precision :  0.0
recall :  0.0
auc :  0.9823529720306396
prc :  0.6000000238418579
```

Hình 8 Kết quả xây dựng mô hình cơ sở (baseline model)

Với mẫu 172 người, mô hình dự đoán đúng 170 người không bị bệnh (Âm tính thực - True Negatives) và không dự đoán được 2 người mắc bệnh (Âm tính giả - False Negatives). Nếu mô hình đã dự đoán mọi thứ một cách hoàn hảo, đây sẽ là một ma trận đường chéo trong đó các giá trị nằm ngoài đường chéo chính, cho biết các dự đoán không chính xác sẽ bằng không. Trong trường hợp này, ma trận cho thấy rằng có tương đối ít dự đoán sai, có nghĩa là có tương đối ít trường hợp bị bệnh nhưng không được phát hiện. Tuy nhiên, ta có thể muốn có ít âm tính giả (False Negatives) hơn mặc dù phải tăng số lượng dương tính giả (False Positives). Việc đánh đổi này có thể phù hợp hơn vì có ít âm tính giả sẽ an toàn hơn là có nhiều dương tính giả.



Hình 9 Biểu đồ ROC của mô hình cơ sở (baseline)



Hình 10 Biểu đồ AUPRC của mô hình cơ sở (baseline)

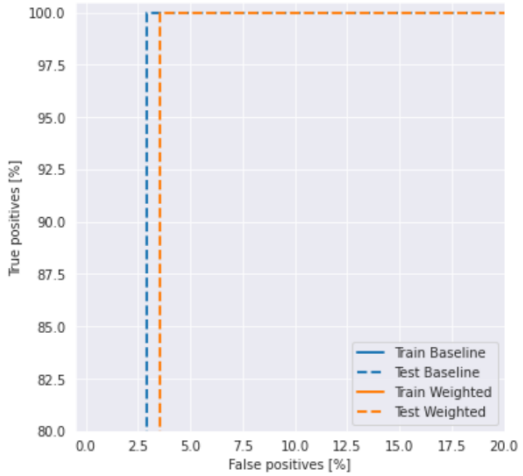
Trong hình 9 cho thấy tỷ lệ dương tính giả (FP – False Positive) thấp chỉ khoảng 2.5%. Biểu đồ này cho thấy tập test có sai số, nhưng tập train không có. Trong hình 10, cho thấy sự phân bố của *Precision* và *Recall* của mô hình cơ sở. Đối với tập train nếu precision giảm thì recall sẽ tăng, còn tập test gãy khúc rõ nét hơn do tập dữ liệu không cân bằng. Trong ví dụ này, âm tính giả (người bị bệnh không được phát hiện) có thể gây ra dự đoán sai dẫn đến nguy hiểm, trong khi dương tính giả (người không bị bệnh nhưng lại được chẩn đoán là mắc bệnh) có thể gây lo lắng cho người kiểm tra. Điều này cho thấy dữ liệu chưa tốt, và chúng ta cần phải sử dụng một số kỹ thuật để nâng cao độ phù hợp của mô hình.

B. Hiệu chỉnh mô hình với phương pháp đánh trọng số lớp (Class Weighting)

Nhằm nâng cao độ phù hợp của mô hình cơ sở, chúng tôi sử dụng kỹ thuật Class Weighting do Keras cung cấp. Mục đích là để xác định những người bị mắc bệnh, nhưng ta không có nhiều mẫu dương tính để làm việc, vì vậy ta sẽ muốn bộ phân loại có heavily weight đối với một số trường hợp có sẵn. Ta có thể thực hiện việc này bằng cách chuyển các trọng số Keras [30] cho mỗi lớp thông qua một tham số. Những điều này sẽ khiến mô hình "chú ý nhiều hơn" đến các trường hợp từ một lớp có ít thể hiện. Cụ thể ở đây là những trường hợp bị ung thư. Với kỹ thuật này, sau khi thiết lập, trọng số của mô hình với class 0 và 1 lần lượt là 0.51 và 23.83.

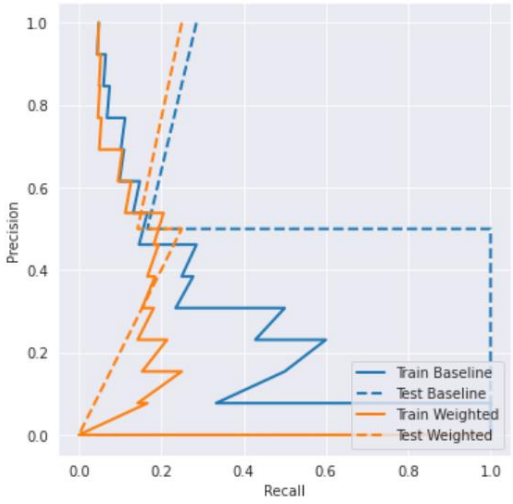
```
loss :  0.18796148896217346
tp :  2.0
fp :  10.0
tn :  160.0
fn :  0.0
accuracy :  0.9418604373931885
precision :  0.1666666716337204
recall :  1.0
auc :  0.9735293984413147
prc :  0.16788256168365479
```

Hình 11 Kết quả xây dựng mô hình sau khi sử dụng phương pháp đánh trọng số của Keras



Hình 12. Biểu đồ AUPRC của mô hình sau khi sử dụng phương pháp đánh trọng số

Cùng với mẫu 172, mô hình đã dự đoán được 160 người không bị bệnh (Âm tính thực), 10 người bị bệnh nhưng không chính xác (Dương tính giả) và đã dự đoán đúng 2 người mắc bệnh (Dương tính thật) với độ chính xác là 94.18%. Mặc dù mô hình dự đoán nhầm đến 10 người không bị bệnh nhưng đổi lại đã tìm ra được người mắc bệnh. Cụ thể ở trường hợp này là việc xác định việc bệnh nhân có mắc phải bệnh này hay không quan trọng hơn việc chuẩn đoán nhầm rất nhiều.



Hình 13 Biểu đồ AUPRC của mô hình sau khi sử dụng phương pháp đánh trọng số.

Sau khi sử dụng class weighting, mô hình thu được khá mượt mà và phù hợp hơn với dữ liệu. Hình 12, cho thấy tỷ lệ dương tính giả sau khi sử dụng đánh trọng số, của tập test tăng lên và lớn hơn 2.5% và xấp xỉ 3.0%. Trong hình 13, cho thấy rõ nét của sự phân bố Precision và Recall hợp lý hơn, tập test sau khi đánh trọng số xích lại gần tập train, cho thấy sự phù hợp mô hình tăng lên.

Bên cạnh *class weighting*, còn khá nhiều kỹ thuật xử lý do Keras hỗ trợ để xử lý nâng cao độ phù hợp của mô hình xây dựng từ dữ liệu không cân bằng. Trong nghiên cứu này dừng lại ở *class weighting*.

VI. KẾT LUẬN

Bài báo này là kết quả nghiên cứu về logit model, bộ dữ liệu không cân bằng và các kỹ thuật học sâu cung cấp bởi Kera. Chúng tôi đã đề xuất một hệ thống chẩn đoán bệnh thông qua xây dựng mô hình dự đoán bệnh bằng cách sử dụng bộ dữ liệu không cân bằng chứa các yếu tố nguy cơ của bệnh ung thư cổ tử cung. Trong số 172 hồ sơ dữ liệu testing, chương trình chạy ra kết quả chẩn đoán trùng khớp với hồ sơ 2 dương tính chính xác đạt tỷ lệ 94.18%. Những cảnh báo ban đầu này sẽ là cơ sở để người bệnh có cái nhìn khách quan hơn, cũng như tầm quan trọng về bệnh ung thư cổ tử cung của mình để có thể lập kế hoạch chăm sóc và điều trị phù hợp.

LỜI CẢM ƠN

Xin chân thành cảm ơn trường Đại học Mở Thành phố Hồ Chí Minh và Học viện Công nghệ Bưu chính Viễn thông đã tạo điều kiện cho chúng tôi thực hiện các nghiên cứu này. Bên cạnh đó, chúng tôi cũng xin chân thành cảm ơn quý thầy cô đã luôn ủng hộ, động viên và giúp đỡ chúng tôi trong suốt quá trình nghiên cứu.

TÀI LIỆU THAM KHẢO

[1] Globocan, "World Cancer," Cancer today, March 2021. [Online]. Available: https://gco.iarc.fr/today/fact-sheets-cancers.

[2] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M and et al, "Global Cancer Observatory: Cancer Today," Lyon: International Agency for Research on Cancer, 2020. [Online]. Available: https://gco.iarc.fr/today. [Accessed February 2021].

[3] "UCI Machine Learning Repository," [Online]. Available: https://archive.ics.uci.edu/ml/. [Accessed 09 April 2021].

[4] W. contributors, "Cancer," Wikipedia, The Free Encyclopedia., [Online]. Available: https://en.wikipedia.org/wiki/Cancer. [Accessed 13 February 2021].

[5] G. W. HealthOrganization, "Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2019 global survey," 2020.

[6] W. contributors, "Cervical cancer," Wikipedia, The Free Encyclopedia, [Online]. Available: https://en.wikipedia.org/wiki/Cervical_cancer. [Accessed 15 April 2021].

[7] B. Copeland, "Artificial intelligence," Encyclopedia Britannica, 11 August 2020. [Online]. Available: https://www.britannica.com/technology/artificial-intelligence. [Accessed 15 April 2021].

[8] "UCI Machine Learning Repository: Cervical cancer (Risk Factors)," 3 March 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. [Accessed 16 April 2021].

[9] M. HARGRAVE, "Deep Learning," Investopedia, 6 April 2021. [Online]. Available: https://www.investopedia.com/terms/d/deep-learning.asp.

[10] "Keras," [Online]. Available: https://keras.io/about/.

[11] "Classification on imbalanced data," Tensorflow, 7 April 2021. [Online]. Available: https://www.tensorflow.org/tutorials/structured_data/imbalanced_data.

[12] C. Versloot, "Working with Imbalanced Datasets with TensorFlow 2.0 and Keras," MachineCurve, 10 November 2020. [Online]. Available: https://www.machinecurve.com/index.php/2020/11/10/working-with-imbalanced-datasets-with-tensorflow-and-keras/.

[13] I. Muhammad, A. Muhammad and S. Youngdoo, "Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods," *MDPI Open Access Journals*, pp. 1-22, 2020.

[14] K. Ahmed, L. Sheng, L.Yong, L.Hong-Xia and M. Sheng-Guan, "A new expert system in prediction of lung cancer disease based on fuzzy soft sets," *Springer*, p. 14179–14207, 2020.

[15] A. Riham, H. Graham and A. Hoda, "Predicting Cervical cancer using machine learning methods," *(IJACSA) International Journal of Advanced Computer Science and Applications*, pp. 173-184, 2020.

[16] A. Talha, K. Muhammad, I. Muhammad, W. Abdul and M. Mubbashar, "Cervical cancer prediction through different screening methods using data mining," *International Journal of Advanced Computer Science and Applications (IJACSA)*, p. 2019, 388–396.

[17] B. Abisoye, A. Ekundayo, L. Kehinde and A. Opeyemi, "Prediction of cervical cancer occurrence using genetic algorithm and support vector machine," in *3rd International Engineering Conference*, 2019.

[18] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classifcation methods for cervical cancer prediction," *A Springer Nature Journals*, 2019.

[19] S. K. Maliha, R. R. Ema, S. K. Ghosh, H. Ahmed, M. R. J. Mollic and T. Islam, "Cancer Disease Prediction Using Naive Bayes,K-Nearest Neighbor and J48 algorithm," in *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019.

[20] W. contributors, "Binary classification," Wikipedia, The Free Encyclopedia, [Online]. Available: https://en.wikipedia.org/wiki/Binary_classification. [Accessed 9 May 2021].

[21] P. Đ. Khang, "Mất cân bằng dữ liệu (imbalanced dataset)," Khoa học dữ liệu, 17 February 2020. [Online]. Available: https://phamdinhkhanh.github.io/2020/02/17/ImbalancedData.html.

[22] V. H. Tiệp, "Overfitting," Machine Learning cơ bản, [Trực tuyến]. Available: https://machinelearningcoban.com/2017/03/04/overfitting/.

[23] Google Machine Learning, "Machine Learning Glossary," [Trực tuyến]. Available: https://developers.google.com/machine-learning/glossary/. [Đã truy cập 6 January 2021].

[24] Srivastava, Nitish and Hinton, Geoffrey and Krizhevsky, Alex and Sutskever, Ilya and Salakhutdinov, Ruslan, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[25] "Overfitting," Wikipedia, The Free Encyclopedia, [Online]. Available: https://en.wikipedia.org/wiki/Overfitting. [Accessed 25 May 2021].

[26] "Baseline Model," ScienceDirect, [Online]. Available: https://www.sciencedirect.com/topics/computer-science/baseline-model.

[27] Tensorflow, "tf.keras.Sequential," [Trực tuyến]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/Sequential. [Đã truy cập 14 May 2021].

[28] "Classification: ROC Curve and AUC," Machine Learning Crash Course, [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc. [Accessed 10 April 2020].

[29] Davis, Jesse and Goadrich, Mark, "The Relationship between Precision-Recall and ROC Curves," in *Association for Computing Machinery*, {Pittsburgh, Pennsylvania, USA, 2006.

[30] "Classification on imbalanced data," TensorFlow, 7 April 2021. [Trực tuyến]. Available: https://www.tensorflow.org/tutorials/structured_data/imbalanced_data#class_weights.

[31] F, Asadi et al, "Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer," *Journal of biomedical physics & engineering*, vol. 10, pp. 513-522, 2020.

[32] K. Ajay, S. Rama and T. Kumar, "Machine Learning Based Approaches for Cancer Prediction: A Survey," in *Proceedings of 2nd International Confer-ence on Advanced Computing and Software Engineering (ICACSE) 2019*, 2020.

[33] S. Rati, Y.Vikash, P. Parashu and P. Pankaj, "Machine Learning Techniques for Detecting and Predicting Breast Cancer," *International Journal of Innovative Technology and Ex-ploring Engineering (IJITEE)*, pp. 2658-2662, 2019.

[34] S. Jaswinder and S. Sandeep, "Prediction of Cervical Cancer Using Machine Learning Techniques," *International Journal of Applied Engineering Research ISSN*, vol. 14, pp. 2570-2577, 2019.

BUILDING CERVICAL CANCER DIAGNOSIS PROBLEM USING LOGIT MODEL WITH DEEP LEARNING

***Abstract*:** According to the Global Cancer Observatory, cervical cancer globally caused 604,127 new cases and 341,831 deaths in 2020 [1]. Besides, the number of cervical cancer patients with no sign has increased in recent years [1] [2]. The majority of patients come for medical examination and treatment when the cancer is in a late-stage, at which time patients must perform a partial or even total removal of their uterus. Providing early warning for the existence of cervical cancer is a practical job. This would help patients to have a specific view of their health status, and then they can take the right actions or take support in improving their health and getting earlier treatment. This study is continuously discovering cervical cancer using AI, especially in machine learning. We aim to apply a new approach of deep learning with a logit model to help to detect cervical cancer with the provided data from patients. We use Keras's deep learning and its techniques - class weighting for getting better results. The system is designed that the input data is risk factors of cervical cancer, and we use a model which is built with the training data set obtained from the UCI Machine Learning Repository [3]. The experimental result with model accuracy is 94,18% compared to the actual diagnostic result. This result shows us a positive prediction of cervical cancer with the logit model, proves that AI and Machine Learning are effectively helping patients predict and prevent cancer before getting it.

***Keywords*:** Cancer Diagnosis, Cervical Cancer, Cervical Cancer Diagnosis, Machine Learning, Deep learning, Artificial intelligence.

TIỂU SỬ CỦA TÁC GIẢ

Trần Công Hùng sinh năm 1962. Ông nhận bằng Kỹ sư điện tử viễn thông hạng Nhất của trường Đại học Công nghệ HOCHIMINH Việt Nam năm 1987. Ông nhận bằng Kỹ sư tin học và máy tính của trường Đại học Công nghệ HOCHIMINH Việt Nam, 1995. Ông nhận bằng thạc sĩ kỹ thuật viễn thông khóa sau đại học Trường Đại học Bách Khoa Hà Nội, 1998. Ông nhận bằng TS. tại Đại học Bách khoa Hà Nội, Việt Nam, 2004. Lĩnh vực nghiên cứu chính: các tham số và phương pháp đo hiệu suất B - ISDN, QoS trong mạng tốc độ cao, MPLS. Ông hiện là Phó Giáo sư Tiến sĩ thuộc Khoa Công nghệ Thông tin II, Học viện Công nghệ Bưu chính Viễn thông tại Hồ Chí Minh, Việt Nam.

Lê Ngọc Hiếu bắt đầu làm việc trong lĩnh vực CNTT với vai trò Kiến trúc sư Hệ thống CNTT từ năm 2010. Hiện tại đang là giảng viên CNTT cho trường Đại học Mở TP.HCM (Việt Nam). Nghiên cứu chính: điện toán đám mây, hiệu quả đám mây để có dịch vụ tốt hơn, ứng dụng của AI, Học máy và Ngôn ngữ học máy tính. Các ngành học phụ: giáo dục, giáo dục về CNTT, Giảng dạy ngôn ngữ (tiếng Trung và tiếng Anh), Kinh doanh và Kinh tế.
Thông tin thêm: *https://www.researchgate.net/profile/Hieu-Le-24*

Võ Phạm Huyền Khanh là sinh viên CNTT trường Đại học Mở TP.HCM. Bắt đầu nghiên cứu về dữ liệu và máy học từ năm 2020. Lĩnh vực nghiên cứu quan tâm đến dữ liệu y học và cách phát hiện bệnh bằng dữ liệu. Cô ấy sẽ tiếp tục nghiên cứu sâu hơn và tiếp tục học lên sau khi tốt nghiệp.