Chẩn đoán ung thư vú bằng học máy thông qua dữ liệu tế bào

Bùi Hải Đăng, Vũ Nguyên Đan, Phan Tuấn Hiệp
Viện Trí tuệ nhân tạo
Trường Đai học Công nghệ
Đại học Quốc gia Hà Nội

Ngày 5 tháng 5 năm 2025

Tóm tắt nội dung

Ung thư vú là một trong những nguyên nhân gây tử vong hàng đầu ở phụ nữ trên toàn thế giới, bao gồm cả Việt Nam. Do đó, việc áp dụng các phương pháp chẩn đoán chính xác và kịp thời là vô cùng cần thiết. Trong nghiên cứu này, chúng tôi đề xuất một mô hình học máy kết hợp Linear Discriminant Analysis (LDA) và Support Vector Machine (SVM) để phân loại các tế bào ung thư vú thành lành tính hoặc ác tính, sử dụng tập dữ liệu Breast Cancer Wisconsin (Diagnostic). Dữ liệu bao gồm các đặc trưng hình thái học từ hình ảnh sinh thiết tế bào. Quy trình nghiên cứu được thực hiện qua ba giai đoạn chính: tiền xử lý dữ liệu (chuẩn hóa, xử lý giá trị thiếu), giảm chiều dữ liệu bằng LDA để trích xuất các đặc trưng phân biệt tối ưu, và phân loại bằng SVM với kernel tuyến tính. Mô hình được đánh giá thông qua các chỉ số độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc hiệu (specificity), và giá trị F1-score, đồng thời so sánh với các phương pháp truyền thống như SVM đơn lẻ và Logistic Regression. Kết quả thực nghiệm cho thấy mô hình kết hợp LDA-SVM đạt độ chính xác rất cao: Accuracy đạt 98.83%, độ nhạy (Sensitivity) đạt 96.88%, độ đặc hiệu (Specificity) đạt 100% và F1-Score đạt 98.41%, vượt trội so với các phương pháp cơ sở, khẳng định tiềm năng ứng dụng trong hỗ trợ chẩn đoán lâm sàng. Nghiên cứu này không chỉ góp phần cải thiện hiệu quả chẩn đoán ung thư vú mà còn làm nổi bật vai trò của việc kết hợp các kỹ thuật giảm chiều dữ liệu và phân loại trong lĩnh vực y sinh.

1 Giới thiệu

Ung thư vú là sự tăng trưởng bất thường của các tế bào trong vú. U cục hoặc khối thường có thể sờ thấy và được gọi là u. U phát triển khi các tế bào trong vú phân chia không kiểm soát và sinh ra mô thừa. U vú có thể lành tính (không ung thư) hoặc ác tính (ung thư). Theo thống kê của Tổ chức Y tế Thế giới (WHO), vào năm 2022, có 2,3 triệu phụ nữ được chẩn đoán mắc ung thư vú và 670.000 ca tử vong trên toàn cầu. Trong đó Việt Nam là khoảng 24.600 ca mới mắc và hơn 10.000 ca tử vong. Chính vì lý do đó, chúng ta cần những phương pháp chẩn đoán sớm và chính xác. Chẩn đoán kịp thời không chỉ giúp nâng cao hiệu quả điều trị mà còn giảm tỷ lệ tử vong đáng kể. Trong bối cảnh đó, các kỹ thuật học máy đã trở thành công cụ đầy hứa hẹn trong việc phân tích dữ liệu y sinh, đặc biệt là dữ liệu hình thái học tế bào từ sinh thiết - yếu tố then chốt để phân biệt khối u lành tính và ác tính.

Tập dữ liệu Breast Cancer Wisconsin (Diagnostic) là một trong những cơ sở dữ liệu chuẩn mực, cung cấp các đặc trưng hình thái học chi tiết của nhân tế bào (như bán kính, độ nhẵn, độ lồi lõm, và độ đối xứng) được trích xuất từ hình ảnh vi mô. Dù nhiều nghiên cứu trước đây đã ứng dụng thành công các mô hình như Support Vector Machine (SVM), Random Forest, hay Logistic Regression trên tập dữ liệu này, nhưng việc kết hợp giữa kỹ thuật giảm chiều dữ liệu và phân loại vẫn còn hạn chế. Đặc biệt, dữ liệu y sinh thường chứa nhiều đặc trưng dư thừa hoặc không liên quan, làm giảm hiệu suất mô hình và tăng nguy cơ overfitting [1].

Trong nghiên cứu này, chúng tôi đề xuất một phương pháp kết hợp giữa Linear Discrimi-

nant Analysis (LDA) và Support Vector Machine (SVM) để tối ưu hóa quy trình chẩn đoán. LDA được sử dụng để giảm chiều dữ liệu, tập trung vào các đặc trưng phân biệt tối ưu giữa hai lớp lành tính và ác tính, trong khi SVM với kernel tuyến tính được tận dụng để phân loại dựa trên không gian đặc trưng mới này. Sự kết hợp này nhằm giải quyết hai thách thức chính: (1) Giảm nhiễu và tăng tính phân tách của dữ liệu trước khi phân loại, (2) Nâng cao độ chính xác và ổn định của mô hình so với các phương pháp đơn lẻ [2, 3].

Nghiên cứu được thực hiện theo ba giai đoạn chính: Tiền xử lý dữ liệu (chuẩn hóa, xử lý giá trị thiếu), giảm chiều dữ liệu bằng LDA, và huấn luyện mô hình SVM. Kết quả thực nghiệm trên tập dữ liệu Wisconsin cho thấy mô hình đề xuất đạt độ chính xác cao (98.83%), vượt trội so với SVM đơn lẻ và Logistic Regression. Đóng góp chính của nghiên cứu bao gồm: (1) Xác định tính khả thi của việc tích hợp LDA-SVM trong chẩn đoán ung thư vú, (2) Cung cấp một quy trình tối ưu từ tiền xử lý dữ liệu đến khâu phân loại.

2 Các nghiên cứu liên quan

Năm 2016, Asri et al. [4] đã so sánh các thuật toán học máy khác nhau, bao gồm SVM, cây quyết định (C4.5), Naïve Bayes (NB) và K-NN trên tập dữ liệu Breast Cancer Wisconsin (Original) để dự đoán và chẩn đoán nguy cơ ung thư vú. Kết quả thí nghiệm cho thấy SVM đạt độ chính xác cao nhất với tỷ lệ lỗi thấp.

Vào năm 2021, Bài nghiên cứu "Linear discriminant analysis and support vector machines for classifying breast cancer" [5] đã tập trung vào việc so sánh hai phương pháp LDA (Linear Discriminant Analysis) và SVM (Support Vector Machines) trong phân loại ung thư vú. Đã cho thấy SVM có thể được sử dụng để hỗ trợ chẩn đoán ung thư chính xác hơn.

Vào năm 2023, bài nghiên cứu "Diagnosis of Breast Cancer Using Random Forests" [6] tập trung vào việc sử dụng thuật toán Random Forest để phân loại ung thư vú với độ chính xác rất cao.

Khourdifi [7] đã áp dụng bốn kỹ thuật học máy, bao gồm SVM, Random Forest (RF), Naïve Bayes và K-nearest neighbor (K-NN), trên tập dữ liệu ung thư vú Wisconsin từ kho dữ liệu học máy UCI. Các tác giả sử dụng phần mềm Waikato Environment for Knowledge Analysis (WEKA) để mô phỏng thuật toán. Kết quả cho thấy SVM có hiệu suất tổng thể tốt nhất về hiệu quả và độ chính xác.

Amreenbatool và Yung-Cheolbyun [8] đã cải thiện phân loại ung thư vú bằng thuật toán học tập tổng hợp bỏ phiếu thích ứng. Thí nghiệm trên đã sử dụng bốn mô hình Extra Tree, Light Gradient Boosting Machine (LightGBM), Ridge Classifier (RC), Linear Discriminant Analysis (LDA) để từ đó đào tạo ra một mô hình Voting Classifier đạt độ chính xác cao (97.63%).

Ricciardi *et al.* [9] đã sử dụng kết hợp phân tích phân biệt tuyến tính (LDA) và phân tích thành phần chính (PCA) để phân loại bệnh động mạch vành. Trong đó, PCA được sử dụng để tạo đặc trưng mới, còn LDA được áp dụng để phân loại, giúp cải thiện khả năng chẩn đoán bệnh nhân.

Trong một nghiên cứu khác, Sivakami và Saraswathi [10] đã áp dụng mô hình lai giữa cây quyết định (DT) và máy vector hỗ trợ (SVM) để dự đoán ung thư vú. Cây quyết định được sử dụng để chọn đặc trưng, và phương pháp đề xuất đã cải thiện độ chính xác dự đoán lên 91%. Trong một nghiên cứu gần đây về ung thư vú, Wu và Hicks [11] đã khảo sát bốn thuật toán học máy: máy vector hỗ trợ (SVM), K-nearest neighbor (KNN), Naïve Bayes và cây quyết định để phân loại ung thư vú ba âm tính và không ba âm tính dựa trên dữ liệu biểu hiện gen. Kết quả cho thấy SVM có hiệu suất phân loại tốt hơn ba thuật toán còn lại.

Zheng *et al.* [12] đã nghiên cứu thuật toán K-means và Support Vector Machine (K-SVM) dựa trên xác thực chéo 10 lần, và phương pháp được đề xuất đã nâng cao độ chính xác dự đoán ung thư vú lên 97,38% khi thử nghiệm trên tập dữ liệu Wisconsin Diagnostic Breast Cancer (WDBC). Các tác giả đã đề xuất một sự kết hợp mới của các thuật toán học máy, cụ thể là sử dụng K-means để nhận diện riêng biệt các mẫu ẩn trong khối u ác tính và lành tính, sau đó dùng SVM để tạo ra bộ phân loại mới trong quá trình xác thực chéo 10 lần. Phương pháp mới của họ đạt độ chính xác 97,38%, cao hơn so với điểm số của sáu thuật toán khác.

Vào năm 2020, Boeri *et al.* [13] đã sử dụng Artificial Neural Network (ANN) và SVM để tiên lượng tái phát ung thư vú cũng như tử vong của bệnh nhân trong vòng 32 tháng sau phẫu thuật. SVM đạt hiệu suất tốt nhất, với độ chính xác

96,86%.

Tóm lại, dựa vào các nghiên cứu đã đề cập ở trên, nghiên cứu chúng tôi đề xuất đã kết hợp hành vi của các thuật toán học máy (Support Vector Machine - SVM) và các kỹ thuật trích xuất đặc trưng (Linear Discriminant Analysis - LDA) để áp dụng trên tập dữ liệu ung thư vú Wisconsin, hỗ trợ chẩn đoán chính xác ung thư vú. Mục tiêu của nghiên cứu là đạt được độ chính xác cao nhất với tỷ lệ lỗi thấp nhất trong phân tích dữ liệu. Để làm được điều đó, hiệu suất của mô hình sẽ được đánh giá dựa trên các tiêu chí như độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc hiệu (specificity), và giá trị F1-score.

3 Dữ liệu và Phương pháp nghiên cứu

3.1 Dữ liệu

Tập dữ liệu Breast Cancer Wisconsin (WBCD) [14] là một trong những tập dữ liệu quan trọng trong nghiên cứu y học và học máy để chẩn đoán ung thư vú. Dữ liệu này được cung cấp bởi Đại học Wisconsin và đã được sử dụng rộng rãi để phát triển các mô hình dự đoán nhằm phân biệt giữa khối u ác tính và lành tính.

Tập dữ liệu được thu thập từ các mẫu sinh thiết bằng kim nhỏ (Fine Needle Aspiration - FNA) của mô vú. Mỗi quan sát trong tập dữ liệu đại diện cho một mẫu khối u, được phân tích dựa trên các đặc trưng hình thái của nhân tế bào.

Bộ dữ liệu bao gồm 569 mẫu với 32 thuộc tính bao gồm cột ID, cột chẩn đoán (M = ác tính, B = lành tính) và 30 cột đặc trưng số. Tập dữ liệu bao gồm 10 đặc điểm chính về hình thái của nhân tế bào, mỗi đặc điểm được biểu diễn bằng ba giá trị: giá trị trung bình, độ lệch chuẩn và giá trị lớn nhất. Các đặc điểm này bao gồm: bán kính, kết cấu, chu vi, diện tích, độ mịn, độ đặc, độ lõm, điểm lõm, độ đối xứng và kích thước fractal.

Tổng cộng 357 được dán nhãn là lành tính, cho thấy các tình trạng không phải ung thư. Ngược lại, 212 người còn lại được dán nhãn là ác tính, biểu thị sự hiện diện của ung thư. Bảng 1 cung cấp thông tin chi tiết về tập dữ liệu, bao gồm các tính năng và lớp của nó. Chia tập dữ liệu thành 70% và 30%. Trong đó 70% tập dữ liệu được sử dụng để đào tạo, trong khi 30% còn lại được giữ lại để thử nghiệm. Sự phân chia này đảm bảo rằng các mô hình phân loại được đánh giá công bằng và toàn diện.

3.1.1 Phân tích và khai phá dữ liệu

Phân tích dữ liệu là một bước quan trọng trong việc hiểu và chuẩn bị dữ liệu cho việc phân tích ung thư vú. Một số bước chính bao gồm:

- Hiểu dữ liệu: Xác định cấu trúc của tập dữ liệu, loại biến, và các đặc trưng có sẵn. Việc này giúp xây dựng nền tảng cho các bước xử lý và phân tích tiếp theo.
- Làm sạch dữ liệu: Phát hiện và xử lý các lỗi hoặc mâu thuẫn trong dữ liệu. Điều này bao gồm việc điền giá trị khuyết, loại bỏ ngoại lệ, và sửa lỗi nhập liệu để đảm bảo tính toàn vẹn của dữ liệu.
- Trực quan hóa dữ liệu: Tạo các biểu đồ trực quan nhằm khám phá đặc điểm và xu hướng của dữ liệu. Các phương pháp phổ biến bao gồm biểu đồ phân bố, biểu đồ hộp (box-plot), và biểu đồ phân tán (scatter plot), giúp nhận diện các mẫu và mối quan hệ giữa các biến.
- Chuyển đổi dữ liệu: Chuẩn bị dữ liệu cho quá trình phân tích bằng cách thực hiện các phép biến đổi như chuẩn hóa (normalization), chuẩn hóa tỷ lệ (scaling), hoặc mã hóa biến phân loại thành biến giả (one-hot encoding).
- Lựa chọn đặc trưng: Xác định các đặc trưng quan trọng nhất cho mô hình phân loại. Quá trình này có thể dựa trên phân tích tương quan hoặc các phương pháp thống kê để chọn ra các biến có ảnh hưởng lớn đến kết quả dự đoán.

Ban đầu, giá trị trung bình của phân phối cho từng đặc trưng được sử dụng để xác định các thống kê của 32 đặc trưng. Trong số đó, có 9 thuộc tính tiêu biểu được lựa chọn và trích xuất để tìm hiểu tổng quan về dữ liệu. Từ đó đưa ra được hướng nghiên cứu. Hình 1 độ lệch chuẩn và khoảng biến thiên của từng thuộc tính đối với các đặc điểm quan trọng nhất từ tập dữ liệu giá trị thực. Nó thể hiện sự phân phối của 9 thuộc tính này cùng với giá trị trung bình. Phân phối này tương đối chuẩn trong phần lớn tập dữ liệu.

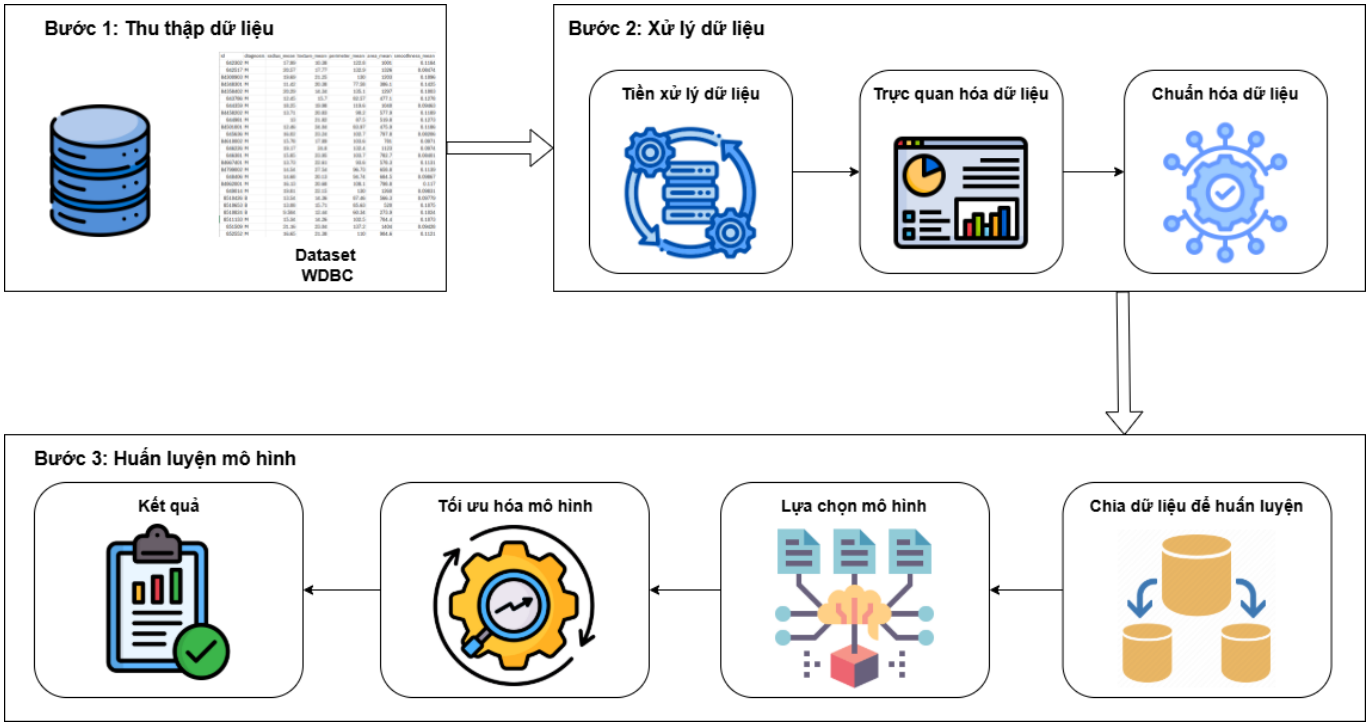

Figure 1: Tổng quan các bước nghiên cứu

Table 1: **Mô tả chi tiết về Dataset**

Phân loại	Tổng số lượng	Số lượng huấn luyện	Số lượng kiểm tra	Nhãn
Lành tính	357	250	107	0
Ác tính	212	148	64	1

3.1.2 Xử lý dữ liệu

Tiền xử lý dữ liệu là một bước quan trọng nhằm đảm bảo chất lượng đầu vào cho các mô hình học máy. Trong nghiên cứu này, bộ dữ liệu Breast Cancer Wisconsin (Diagnostic) được sử dụng để xây dựng mô hình phân loại ung thư vú. Bộ dữ liệu này chứa thông tin về 30 đặc trưng số học liên quan đến hình thái tế bào và một nhãn phân loại gồm hai loại: lành tính (B – Benign) và ác tính (M – Malignant). Các bước tiền xử lý dữ liệu như sau:

- Xử lý dữ liệu bị thiếu và không cần thiết: Bước đầu tiên trong tiền xử lý là kiểm tra sự thiếu hụt dữ liệu. Trong tập dữ liệu này, không có giá trị bị thiếu, do đó không cần áp dụng kỹ thuật bù đắp dữ liệu. Tuy nhiên, cột ID không chứa thông tin có ý nghĩa đối với việc phân loại nên đã bị loại bỏ để tránh ảnh hưởng đến quá trình huấn luyện mô hình.

- Chuyển đổi nhãn thành dạng số: Dữ liệu ban đầu có nhãn 'M' (Malignant - ác tính) và 'B' (Benign - lành tính). Để mô hình học máy có thể xử lý dễ dàng, nhãn được ánh xạ thành giá trị số: M = 1, B = 0.

- Chuẩn hóa dữ liệu: Các đặc trưng trong bộ dữ liệu có phạm vi giá trị khác nhau, có thể gây ảnh hưởng đến hiệu suất của mô hình. Do đó, tập dữ liệu đã được chuẩn hóa bằng StandardScaler, đưa giá trị của mỗi đặc trưng về phân phối có trung bình 0 và độ lệch chuẩn 1. Việc chuẩn hóa giúp cải thiện hiệu suất của mô hình, đặc biệt là đối với các thuật toán như Support Vector Machine (SVM).

- Giảm chiều dữ liệu bằng Phân tích Phương sai Tuyến tính (LDA): Để giảm độ phức tạp của dữ liệu và tăng khả năng phân tách tuyến tính, Linear Discriminant Analysis (LDA) được áp dụng với số thành phần chính. LDA giúp tối ưu hóa không gian đặc trưng bằng cách giữ lại phương sai tối đa giữa các lớp, từ đó cải thiện khả năng phân loại.

- Chia tập dữ liệu thành tập huấn luyện và kiểm tra: Dữ liệu được chia thành tập huấn

Table 2: **Thống kê các đặc trưng quan trọng.**

STT	Đặc trưng	Min	Max	Mean	Độ lệch chuẩn
1	radius_mean	6.981	28.110	14.127	3.524
2	texture_mean	9.710	39.280	19.290	4.301
3	perimeter_mean	43.790	188.500	91.969	24.299
4	area_mean	143.500	2501.000	654.889	351.914
5	smoothness_mean	0.053	0.163	0.096	0.014
6	concavity_mean	0.000	0.427	0.089	0.080
7	concave_points_mean	0.000	0.201	0.049	0.039
8	symmetry_mean	0.106	0.304	0.181	0.027
9	fractal_dimension_mean	0.050	0.097	0.063	0.007

luyện (70%) và tập kiểm tra (30%) để đánh giá hiệu suất mô hình. Việc chia dữ liệu đảm bảo rằng mô hình được học từ một tập hợp đủ lớn và có thể tổng quát hóa tốt trên dữ liệu chưa thấy trước đó.

Quá trình tiền xử lý dữ liệu giúp cải thiện đáng kể hiệu suất của mô hình học máy. Việc chuẩn hóa và giảm chiều dữ liệu không chỉ giúp tăng tốc độ huấn luyện mà còn cải thiện độ chính xác phân loại. Những kỹ thuật này đóng vai trò quan trọng trong việc phát triển mô hình SVM kết hợp với LDA để chẩn đoán ung thư vú một cách chính xác và hiệu quả.

3.2 Phương pháp nghiên cứu

Phương pháp được đề xuất trong nghiên cứu này tập trung vào việc ứng dụng các thuật toán học máy, cụ thể là Support Vector Machine (SVM) và Linear Discriminant Analysis (LDA), để xây dựng một mô hình chẩn đoán ung thư vú ban đầu nhằm phân loại các tế bào thành lành tính hoặc ác tính. Các nghiên cứu gần đây đã khẳng định hiệu quả của SVM trong việc phân loại chính xác dữ liệu y sinh, trong khi LDA đóng vai trò quan trọng trong việc giảm chiều dữ liệu và tăng cường khả năng phân tách giữa các lớp [15]. Sự kết hợp giữa LDA và SVM trong thí nghiệm này mang lại kết quả khả quan trong việc dự đoán ung thư vú, và 2 thuật toán của nghiên cứu được trình bày chi tiết dưới đây để làm rõ cách tiếp cận.

1) **Linear Discriminant Analysis (LDA)**: Linear Discriminant Analysis là một phương pháp phân tích phân biệt có thể được sử dụng trong phân loại và giảm chiều dữ liệu [16, 17, 18]. Mục đích chính của phân tích phân biệt tuyến tính là dự đoán phân loại tốt nhất cho các nhãn đa lớp [19]. Nguyên lý hoạt động chính của LDA là tìm các "đường phân biệt" (discriminant axes) hoặc các thành phần tuyến tính (linear components). Các thành phần này được xác định bằng cách tối đa hóa khoảng cách giữa các trung tâm (means) của các lớp khác nhau sau khi chiếu dữ liệu lên không gian con mới và tối thiểu hóa phương sai (variance) của dữ liệu trong nội bộ mỗi lớp sau khi chiếu dữ liệu lên không gian con mới. Thuật toán được áp dụng như sau:

1. **Tính vector trung bình tổng:**

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2. **Với mỗi lớp $c = 1, 2, \ldots, C$, tính:**

   - *Trung bình lớp:*

$$\mu_c = \frac{1}{n_c} \sum_{x_i \in c} x_i$$

   - *Ma trận scatter trong lớp:*

$$S_W = \sum_{c=1}^{C} \sum_{x_i \in c} (x_i - \mu_c)(x_i - \mu_c)^T$$

   - *Ma trận scatter giữa các lớp:*

$$S_B = \sum_{c=1}^{C} n_c (\mu_c - \mu)(\mu_c - \mu)^T$$

3. **Giải bài toán trị riêng:**

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}$$

   Lấy k vector riêng ứng với k trị riêng lớn nhất để tạo ma trận chiếu $\mathbf{W}$ sao cho:

$$\mathbf{W} \in \mathbb{R}^{d \times k}$$

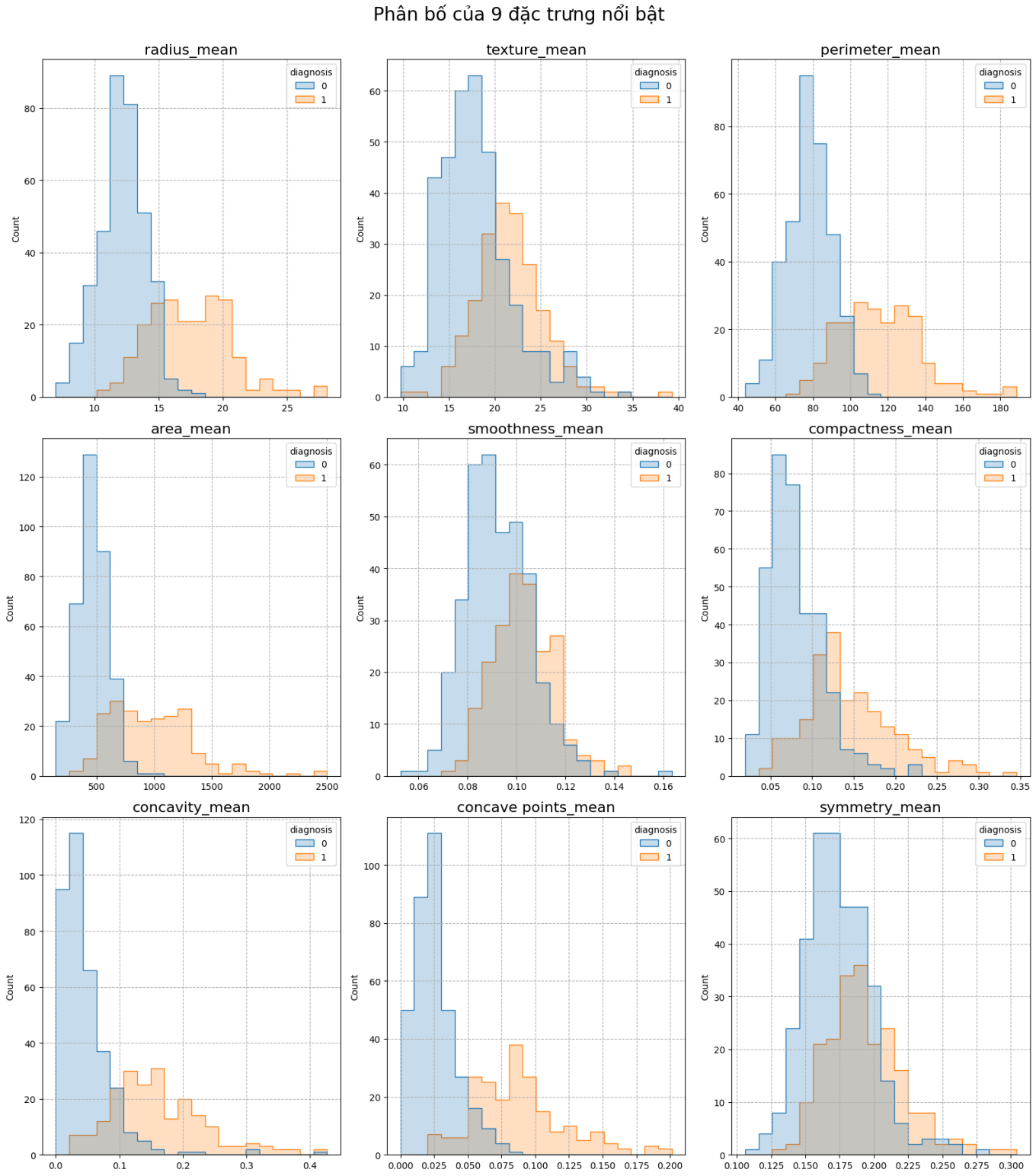

Figure 2: Phân bố của 9 đặc trưng nổi bật

4. **Chiếu dữ liệu ban đầu vào không gian mới:**

$$\mathbf{Z} = \mathbf{X}\mathbf{W}$$

Lưu ý: Với C lớp, số chiều tối đa sau khi giảm bằng LDA là $C - 1$.

Trong đó:

- x_i: Mẫu dữ liệu thứ i, là một vector đặc trưng trong không gian $\mathbb{R}^d$, với d là số chiều của dữ liệu gốc.
- n: Tổng số mẫu dữ liệu trong tập dữ liệu.
- $\mathbf{w}$: Vector riêng, là một hướng chiếu tối ưu trong không gian đặc trưng.
- λ: Trị riêng, biểu thị độ quan trọng của hướng chiếu tương ứng với vector riêng $\mathbf{w}$.

- **X**: Ma trận dữ liệu gốc, trong đó mỗi hàng là một mẫu x_i, có kích thước $\mathbb{R}^{n \times d}$.

2) **Support Vector Machine (SVM)**: Support Vector Machine là một kỹ thuật học máy có giám sát cho các bài toán phân loại và hồi quy, được đề xuất bởi Vapnik *et al.* vào năm 1992. SVM là một thuật toán tính toán học cách gán nhãn cho các đối tượng từ kinh nghiệm và ví dụ. SVM có thể được áp dụng trong chẩn đoán y học [20, 21, 22], dự báo thời tiết, tài chính [23], phân tích thị trường chứng khoán [24, 25] và xử lý ảnh [26]. SVM có đặc điểm cơ bản là tách dữ liệu nhãn nhị phân dựa trên một đường thẳng đạt được khoảng cách tối đa của dữ liệu nhãn [27]. Thực hiện lần lượt các bước như sau:

1. **Tìm siêu phẳng tối ưu có dạng:**

$$f(x) = \mathbf{w}^T x + b = 0$$

sao cho phân tách hai lớp và khoảng cách đến các điểm gần nhất lớn nhất.

2. **Ràng buộc phân lớp đúng:**

$$y_i(\mathbf{w}^T x_i + b) \geq 1, \quad \forall i = 1, \ldots, n$$

3. **Soft Margin SVM:**

Cho phép 1 số điểm vi phạm với biến slack ξ_i:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

với điều kiện:

$$y_i(\mathbf{w}^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Trong đó:

- $\mathbf{w}$: Vector trọng số, vuông góc với siêu phẳng $f(x) = \mathbf{w}^T x + b = 0$. Nó xác định hướng của siêu phẳng trong không gian đặc trưng.
- x: Vector đặc trưng đầu vào, đại diện cho một điểm dữ liệu trong không gian đặc trưng.
- b: Hệ số chặn (bias coefficient), dịch chuyển siêu phẳng ra khỏi gốc tọa độ để cho phép phân tách các lớp một cách phù hợp.
- y_i: Nhãn lớp của điểm dữ liệu thứ i, với $y_i \in \{-1, 1\}$ trong bài toán phân loại nhị phân.
- ξ_i: Biến chùng (slack variable) cho điểm dữ liệu thứ i trong SVM Soft Margin. Nó đo lường mức độ phân loại sai (degree of misclassification) hoặc mức độ vi phạm các ràng buộc lề (margin constraints violation).
- C: Tham số điều chuẩn ($C > 0$) trong SVM Soft Margin. Nó kiểm soát sự đánh đổi giữa việc tối đa hóa lề (tối thiểu hóa $\|\mathbf{w}\|^2$) và giảm thiểu lỗi phân loại (tổng ξ_i).
- $\|\mathbf{w}\|^2$: Bình phương chuẩn Euclidean của vector trọng số, tỷ lệ nghịch với độ rộng lề. Tối thiểu hóa giá trị này sẽ tối đa hóa lề giữa các lớp.

Mã giả của thuật toán SVM-LDA:

1. **Đầu vào:** Tập dữ liệu ung thư vú D chứa các đặc trưng X và nhãn y.
2. Loại bỏ cột định danh (ví dụ: 'id') khỏi tập dữ liệu D.
3. Mã hóa nhãn y thành giá trị nhị phân ('M' $\rightarrow$ 1, 'B' $\rightarrow$ 0).
4. Chia tập dữ liệu D thành tập huấn luyện D_{train} (70%) và tập kiểm tra D_{test} (30%) với phân tầng (stratified sampling).
5. **Bước 1: Chuẩn hóa đặc trưng**
   (a) Khởi tạo StandardScaler và tính toán trung bình, phương sai trên X_{train}.
   (b) Biến đổi X_{train} và X_{test} về phân phối chuẩn (trung bình 0, phương sai 1).
6. **Bước 2: Áp dụng Phân tích Phân biệt Tuyến tính (LDA)**
   (a) Khởi tạo LDA với số thành phần $n_{\text{components}} = 1$.
   (b) Huấn luyện LDA trên X_{train} đã chuẩn hóa và y_{train} để tính vector chiếu.
   (c) Biến đổi X_{train} và X_{test} về không gian đặc trưng 1 chiều, thu được $X_{\text{train}}^{\text{LDA}}, X_{\text{test}}^{\text{LDA}}$.
7. **Bước 3: Huấn luyện bộ phân loại SVM**
   (a) Khởi tạo SVM với kernel tuyến tính, tham số $C = 0.1$, γ = 'scale'.
   (b) Huấn luyện SVM trên $X_{\text{train}}^{\text{LDA}}, y_{\text{train}}$.

8. **Bước 4: Đánh giá mô hình**
   (a) Dự đoán nhãn $\hat{y}_{\text{test}}$ cho $X_{\text{test}}^{\text{LDA}}$.
   (b) Tính toán ma trận nhầm lẫn, độ chính xác, độ nhạy, độ đặc hiệu và F1-score dựa trên $y_{\text{test}}, \hat{y}_{\text{test}}$.
9. **Đầu ra:** Bộ phân loại SVM đã huấn luyện và các chỉ số đánh giá hiệu năng.
10. **Trả về:** Mô hình SVM đã huấn luyện.

4 Kết quả và Thảo luận

Trong học máy, phân tích hiệu suất của mô hình phân loại là bước không thể thiếu để quyết định mức độ chính xác và khả năng tổng quát của mô hình trên dữ liệu chưa biết. Dưới đây là tổng hợp các chỉ số đo hiệu suất (performance metrics) phổ biến thường được sử dụng trong việc kiểm định mô hình phân loại theo các chỉ tiêu: Accuracy, Precision, Sensitivity, Specificity, F1-Score, AUC-ROC và MCC. Mỗi chỉ số đều sẽ được đề cập về định nghĩa, cách tính, chức năng và các tình huống áp dụng hợp lý.

Ma trận nhầm lẫn (Confusion Matrix) là cơ sở để tính toán hầu hết các chỉ số đánh giá hiệu suất của mô hình phân loại. Mỗi ma trận có hai hàng và hai cột, đại diện cho số lượng âm tính thực, dương tính giả, âm tính giả và dương tính thực. Các số lượng này được sử dụng để tính toán các chỉ số hiệu suất như Accuracy, Precision, Sensitivity, Specificity, F1-Score. Ma trận trên đại diện cho một bài toán phân loại nhị phân với hai lớp (0 và 1). Ma trận này bao gồm các thành phần như sau:

- True Positive (T_P): Số lượng mẫu dương được dự đoán đúng.
- True Negative (T_N): Số lượng mẫu âm được dự đoán đúng.
- False Positive (F_P): Số lượng mẫu âm bị dự đoán nhầm là dương.
- False Negative (F_N): Số lượng mẫu dương bị dự đoán nhầm là âm.

Accuracy (Độ chính xác):

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

Accuracy thể hiện tỷ lệ mẫu được phân loại đúng trên tổng số mẫu.

Precision (Độ chính xác dương):

$$\text{Precision} = \frac{T_P}{T_P + F_P}$$

Precision thể hiện tỷ lệ mẫu được dự đoán là dương và thật sự là dương, trong số tất cả các mẫu được dự đoán là dương.

Sensitivity / Recall (Độ nhạy):

$$\text{Sensitivity} = \frac{T_P}{T_P + F_N}$$

Sensitivity thể hiện khả năng của mô hình trong việc phát hiện đúng các mẫu thuộc lớp dương. Chỉ số này đặc biệt quan trọng trong các bài toán mà việc bỏ sót mẫu dương là nghiêm trọng, ví dụ như chẩn đoán bệnh.

Specificity (Độ đặc hiệu):

$$\text{Specificity} = \frac{T_N}{T_N + F_P}$$

Specificity đo lường khả năng của mô hình trong việc phát hiện đúng các mẫu thuộc lớp âm. Chỉ số này quan trọng khi việc nhận nhầm mẫu âm thành dương gây hậu quả lớn, ví dụ như xét nghiệm dương tính giả.

F1-Score:

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

F1-Score là trung bình điều hòa giữa Precision và Recall, nhằm cân bằng giữa hai chỉ số này trong các bài toán mất cân bằng dữ liệu (class imbalance). F1-Score cao cho thấy mô hình đạt được độ chính xác tốt cả trong việc dự đoán dương đúng và giảm thiểu bỏ sót mẫu dương.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve):
Đường cong AUC-ROC là một công cụ thiết yếu được sử dụng để đánh giá hiệu suất của các mô hình phân loại nhị phân. Đường cong này biểu diễn Tỷ lệ dương tính thật (TPR) so với Tỷ lệ dương tính giả (FPR) ở các ngưỡng khác nhau, cho thấy mức độ phân biệt giữa hai lớp như kết quả dương tính và âm tính của mô hình.

- Đường cong ROC : Đường cong ROC biểu diễn TPR so với FPR ở các ngưỡng khác nhau. Nó thể hiện sự đánh đổi giữa độ nhạy và độ đặc hiệu của một bộ phân loại.
- AUC (Diện tích dưới đường cong) : AUC đo diện tích dưới đường cong ROC. Giá trị AUC cao hơn cho thấy hiệu suất mô hình tốt hơn vì nó cho thấy khả năng phân biệt giữa các lớp tốt hơn. Giá trị AUC là 1,0 cho thấy hiệu suất hoàn hảo trong khi 0,5 cho thấy đó là phỏng đoán ngẫu nhiên.

4.1 Kết quả

Bài nghiên cứu trên được thực hiện để chẩn đoán ung thư vú thông qua bộ dữ liệu ung thư vú Wisconsin (WBCD), tập dữ liệu được chia ra làm hai phần: 70% dữ liệu được sử dụng làm tập training và 30% dữ liệu còn lại dùng để làm tập test.

Việc làm trên đảm bảo tính cân bằng khi huấn luyện mô hình. Mô hình trên được huấn luyện và đánh giá dựa trên các chỉ số hiệu suất như Accuracy, Precision, Sensitivity, Specificity, F1-Score.

Trọng tâm là dự đoán các đặc điểm tối ưu để phát hiện ung thư vú hiệu quả. Ma trận nhầm lẫn được sử dụng để đánh giá độ chính xác của mô hình phân loại và xác định các vấn đề tiềm ẩn. Ma trận này có lợi khi xử lý các bộ dữ liệu có phân phối lớp không đồng đều, ngăn chặn những diễn giải sai lệch về độ chính xác của phân loại. Đánh giá liên quan đến việc phân tích Hình 2, cho thấy ba ma trận nhầm lẫn cho các bộ phân loại máy học khác nhau: SVM, Logistic Regression và LDA + SVM.

Trong quá trình thực nghiệm, mô hình kết hợp giữa Linear Discriminant Analysis (LDA) để giảm chiều dữ liệu và Support Vector Machine (SVM) để phân loại đã cho kết quả vượt trội trong việc phân loại ung thư vú là lành tính hay ác tính dựa trên các đặc trưng rút trích từ dữ liệu y sinh học. Sau khi áp dụng LDA để giảm chiều dữ liệu xuống còn 1 chiều (tương ứng với thành phần phân biệt tối ưu giữa các lớp), dữ liệu được biểu diễn trên không gian một chiều mới cho thấy sự phân tách rõ ràng hơn, tạo điều kiện lý tưởng cho mô hình SVM thực hiện phân loại hiệu quả.

Mô hình SVM sau đó được huấn luyện trên dữ liệu đã biến đổi, sử dụng kernel tuyến tính. Kết quả trên tập kiểm tra cho thấy độ chính xác (Accuracy) đạt 98.83%, độ nhạy (Sensitivity) đạt 96.88%, độ đặc hiệu (Specificity) đạt 100% và F1-Score đạt 98.41% được biểu thị ở bảng 3.

Mô hình trên đã dự đoán chính xác 62 trường hợp của lớp ác tính (dương tính thực) và 107 trường hợp của lớp lành tính (âm tính thực) trong khi dự đoán không chính xác 2 trường hợp của lớp ác tính (âm tính giả).

Các mô hình đơn lẻ gồm Logistic Regression và SVM cũng cho kết quả khá tốt. Với mô hình Logistic Regression đạt độ chính xác (Accuracy) đạt 97.08%, độ nhạy (Sensitivity) đạt 93.75%, độ đặc hiệu (Specificity) đạt 99.07% và F1-Score đạt 96.00%. Với mô hình SVM đạt độ chính xác (Accuracy) đạt 96.49%, độ nhạy (Sensitivity) đạt 90.62%, độ đặc hiệu (Specificity) đạt 98.15% và F1-Score đạt 95.16%.

Để trực quan hóa sự so sánh giữa mô hình chúng tôi đề xuất và mô hình đơn lẻ thì hình 4 được cung cấp. Hình này mô tả một biểu diễn cột, nhằm minh họa độ chính xác tổng thể của từng mô hình. Phân tích trong hình này cho phép so sánh nhanh chóng và trực quan về hiệu năng của phương pháp đề xuất so với các mô hình/bộ học cơ sở (baseline learners).

Khi đánh giá hiệu suất mô hình trên một bộ dữ liệu mất cân bằng, các đường cong ROC được sử dụng làm số liệu cụ thể do hiệu quả của chúng trong việc đánh giá khả năng phát hiện dương tính và âm tính giả. Đường cong ROC đặc biệt phù hợp với các đánh giá như vậy. Hình 4 minh họa đường cong ROC của mô hình được đề xuất và mô hình cơ sở. Biểu đồ ROC cung cấp một mô tả đồ họa về khả năng phân biệt giữa các lớp khác nhau của các bộ phân loại, cung cấp thông tin chi tiết về hiệu suất tổng thể của chúng và sự đánh đổi giữa độ nhạy và độ đặc hiệu trên các ngưỡng phân loại khác nhau. Đáng chú ý, kết quả của chúng tôi chỉ ra rằng mô hình được đề xuất đã đạt được giá trị Diện tích dưới đường cong (AUC) cao nhất, đạt điểm tuyệt đối là 1,00.

Như vậy, việc kết hợp LDA với SVM không chỉ giúp giảm chiều dữ liệu và loại bỏ nhiễu mà còn tăng độ chính xác phân loại. Tuy nhiên, để đảm bảo độ tin cậy trong môi trường ứng dụng thực tế, mô hình cần được kiểm thử thêm trên nhiều tập dữ liệu với phân bố đa dạng hơn và có thể kết hợp thêm các phương pháp điều chỉnh mất cân bằng dữ liệu để cải thiện hơn nữa khả năng phát hiện các trường hợp ung thư.

Table 3: Kết quả so sánh các mô hình

Models	Accuracy	Sensitivity	Specificity	F1-score
SVM	96.49%	90.62%	100%	95.08%
Logistic Regression	97.08%	93.75%	99.07%	96.00%
SVM + LDA	**98.83%**	**96.88%**	**100%**	**98.41%**

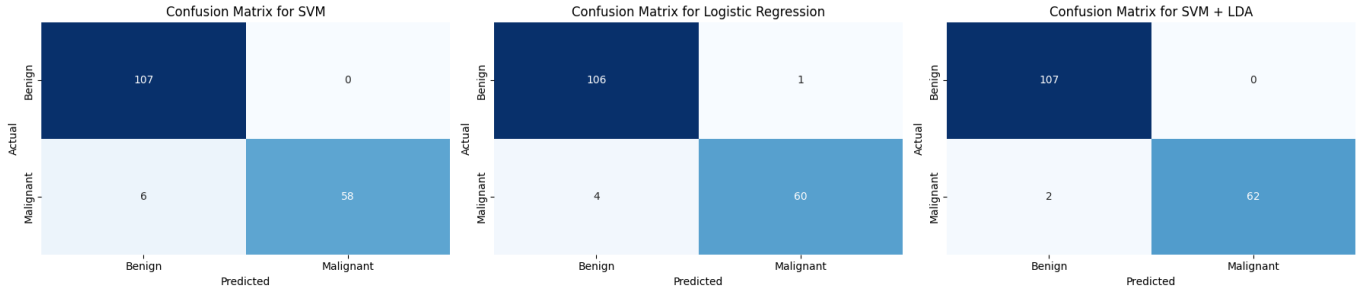

Figure 3: Ma trận nhầm lẫn của từng mô hình thử nghiệm

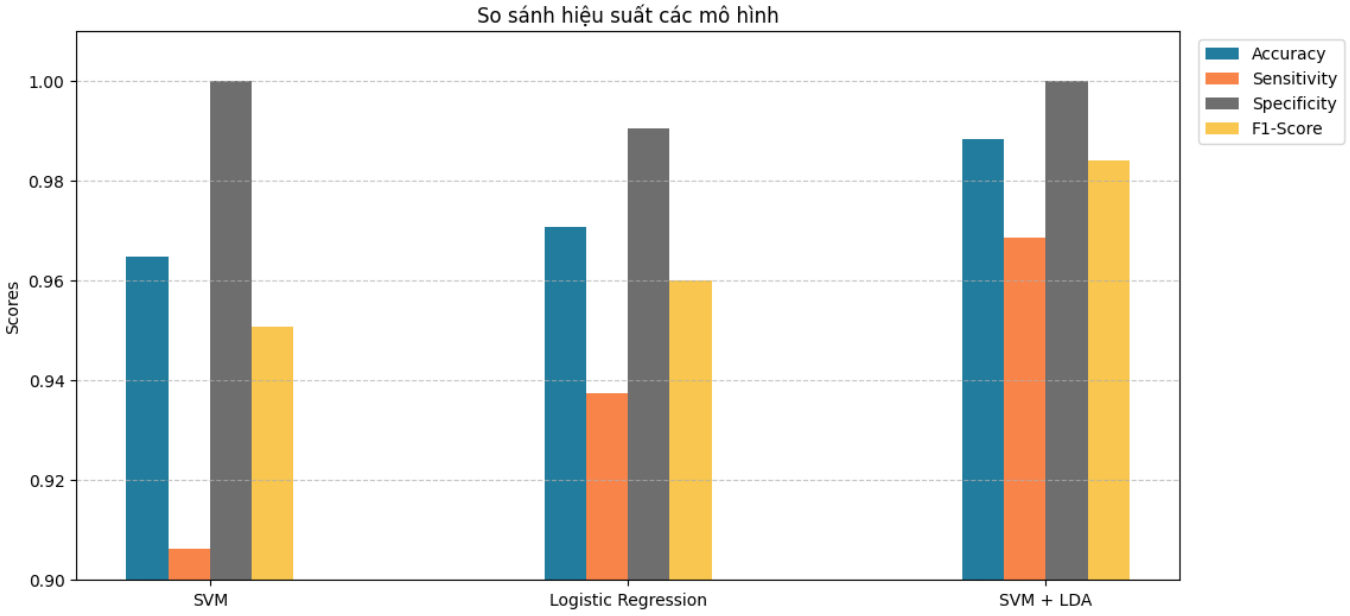

Figure 4: So sánh hiệu suất giữa từng mô hình

4.2 Thảo luận

Mô hình LDA-SVM được đề xuất đã đạt hiệu suất vượt trội trong việc phân loại tế bào ung thư vú thành lành tính hoặc ác tính, với độ chính xác 98,83%, độ nhạy 96,88%, độ đặc hiệu 100% và F1-score 98,41%. Những kết quả này vượt qua SVM đơn lẻ (độ chính xác 96,49%), Logistic Regression đơn lẻ (độ chính xác 97,08%) và các phương pháp truyền thống như Random Forest và Linear Regression, khẳng định lợi ích của việc tích hợp giảm chiều dữ liệu với phân loại mạnh mẽ.

Độ chính xác cao và độ đặc hiệu tuyệt đối cho thấy khả năng của mô hình trong việc xác định chính xác tất cả các trường hợp lành tính, giảm thiểu các trường hợp dương tính giả có thể gây lo lắng không cần thiết hoặc dẫn đến các thủ tục tốn kém. Độ nhạy 96,88% cho thấy mô hình phát hiện gần như toàn bộ các trường hợp ác tính, điều này rất quan trọng để can thiệp sớm. Sự kết hợp giữa khả năng giảm chiều của LDA, tối ưu hóa phương sai giữa các lớp, và khả năng tối ưu ranh giới của SVM có thể đã góp phần giải thích tại sao LDA biến đổi không gian đặc trưng 30 chiều thành không gian dễ phân tách hơn, giúp SVM với kernel tuyến tính đạt được phân loại gần như tối ưu.

So với các nghiên cứu trước đây, mô hình LDA-SVM đạt độ chính xác cao hơn so với SVM đơn lẻ (96,49%) được báo cáo bởi Asri *et al.* (2016) [4] và mô hình K-SVM hybrid (97,38%)

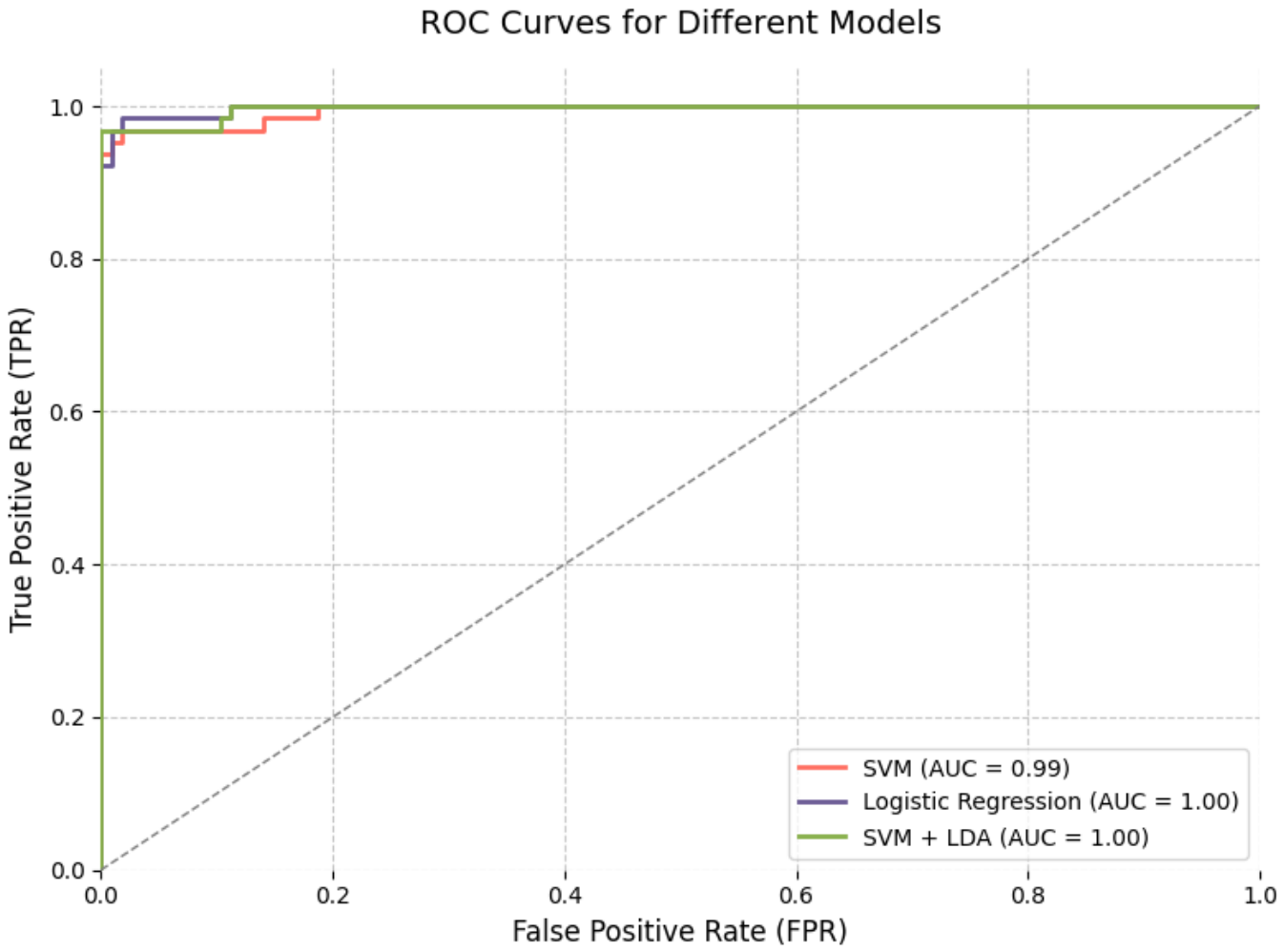

Figure 5: Biểu đồ ROC biểu diễn trực quan về khả năng phân biệt giữa các lớp của các mô hình

được đề xuất bởi Zheng *et al.* (2014) [12]. Không giống như K-means trong nghiên cứu của Zheng *et al.*, vốn phân cụm dữ liệu trước khi phân loại, phương pháp giảm chiều có giám sát của LDA tối ưu hóa trực tiếp khả năng phân tách lớp, có thể giải thích cho hiệu suất vượt trội. Ngoài ra, Rustam *et al.* (2021) [5] đạt được kết quả cạnh tranh với SVM và LDA riêng lẻ, nhưng việc thiếu tích hợp có thể đã giới hạn hiệu suất so với phương pháp kết hợp của chúng tôi. Các bước tiền xử lý, bao gồm chuẩn hóa và loại bỏ cột ID, đã tăng cường thêm độ ổn định của mô hình.

Về mặt lâm sàng, độ đặc hiệu cao của mô hình có thể giảm các ca sinh thiết (biopsy) không cần thiết, trong khi độ nhạy cao đảm bảo hầu hết các trường hợp ác tính được phát hiện, phù hợp với ưu tiên chẩn đoán sớm ung thư vú. Với tình hình ung thư vú gia tăng tại Việt Nam (24.600 ca mới năm 2022, theo WHO), hiệu quả tính toán của mô hình—nhờ giảm chiều của LDA—làm cho nó phù hợp với các cơ sở y tế hạn chế về nguồn lực. Mô hình này có thể đóng vai trò là công cụ hỗ trợ quyết định, đơn giản hóa quy trình chẩn đoán ở những khu vực thiếu thiết bị hình ảnh tiên tiến hoặc chuyên môn bệnh học.

Tuy nhiên, mô hình vẫn có những hạn chế. Việc phụ thuộc vào tập dữ liệu Wisconsin (569 mẫu) làm hạn chế khả năng khái quát hóa cho các quần thể đa dạng. Tính mất cân bằng lớp của tập dữ liệu (357 lành tính so với 212 ác tính) có thể làm tăng nguy cơ quá khớp. Độ nhạy 96,88%, dù cao, cho thấy hai trường hợp ác tính bị phân loại sai thành lành tính, một vấn đề nghiêm trọng do hậu quả của việc bỏ sót chẩn đoán. Ngoài ra, mô hình chưa được kiểm chứng trên các tập dữ liệu bên ngoài hoặc trong môi trường lâm sàng thực tế, nơi chất lượng dữ liệu và sự biến thiên có thể ảnh hưởng đến hiệu suất.

Tóm lại, nghiên cứu này chứng minh hiệu quả của việc kết hợp LDA và SVM trong chẩn đoán ung thư vú, đạt độ chính xác tiên tiến 98,83%. Dù cần thêm xác nhận, mô hình này cung cấp một công cụ hiệu quả tính toán và phù hợp lâm sàng, với tiềm năng nâng cao độ chính xác chẩn đoán và cải thiện kết quả cho bệnh nhân trong quản lý ung thư vú.

5 Kết luận và Hướng phát triển trong tương lai

Ung thư vú là một trong những nguyên nhân hàng đầu gây tử vong ở phụ nữ; do đó, việc phát hiện sớm là vô cùng quan trọng. Việc triển khai các bộ phân loại máy học mạnh mẽ có thể cải thiện đáng kể việc phân loại khối u vú ác tính. Hiệu suất dự đoán phụ thuộc vào nhiều yếu tố mô hình khác nhau. Kết hợp thuật toán giảm chiều và phân loại thường vượt trội hơn một bộ phân loại đơn lẻ vì nó kết hợp nhiều thuật toán giảm chiều dữ liệu và học máy độc lập làm tăng khả năng phân loại đúng. Do đó, nó đã trở nên phổ biến và chứng minh là một phương pháp học máy hiệu quả. Một trong những vấn đề quan trọng nhất là tìm ra cách kết hợp các thuật toán trên để đạt hiệu quả cao nhất. Để giải quyết những vấn đề này, chúng tôi đề xuất áp dụng một mô hình độc đáo được gọi là SVM-LDA. Mô hình trên lựa chọn sự kết hợp việc giảm chiều dữ liệu bằng LDA và phân loại bằng SVM để phân loại ung thư vú một cách chính xác. Kết quả thực nghiệm cho thấy phương pháp SVM-LDA đạt được độ chính xác (Accuracy) đạt được cao nhất 98.83%, độ nhạy (Sensitivity) đạt 96.88%, độ đặc hiệu (Specificity) đạt 100% và F1-Score đạt 98.41%. Hơn nữa, các kết quả thực nghiệm chỉ ra rằng SVM-LDA được đề xuất đã cải thiện độ chính xác so với các mô hình SVM, Logistic regression, Random Forest,... đơn lẻ. Việc phát triển và sử dụng các mô hình kết hợp trên làm tăng khả năng chẩn đoán chính xác và đã cải thiện đáng kể so với các công trình nghiên cứu trước đây.

Hướng tới các nghiên cứu trong tương lai nên xác nhận mô hình trên các tập dữ liệu lớn hơn và đa dạng hơn, như cơ sở dữ liệu SEER hoặc TCGA, để đảm bảo khả năng khái quát hóa. Các kỹ thuật như SMOTE có thể giải quyết mất cân bằng lớp, cải thiện độ nhạy. Việc khám phá các phương pháp giảm chiều khác (ví dụ: PCA, t-SNE) và tinh chỉnh tham số điều chuẩn C của SVM có thể tối ưu hóa thêm hiệu suất. Các thử nghiệm lâm sàng hoặc tích hợp với hệ thống chẩn đoán bệnh viện là cần thiết để đánh giá tính khả thi trong thực tế. Từ đó giúp mô hình được mở rộng và đề xuất trong việc giải quyết các thách thức chăm sóc sức khỏe rộng lớn hơn.

References

[1] B. Zhang *et al.*, "Classification of high dimensional biomedical data based on feature selection using redundant removal," *PLoS One*, vol. 14, no. 4, Art. no. e0214406, Apr. 2019, doi: 10.1371/journal.pone.0214406.

[2] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge, UK: Cambridge Univ. Press, 2000, doi: 10.1017/CBO9780511801389.

[3] A. J. Izenman, "Linear Discriminant Analysis," in *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, New York, NY, USA: Springer, 2008, ch. 8, pp. 233–270, doi: 10.1007/978-0-387-78189-1_8.

[4] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[5] Z. Rustam, Y. Amalia, S. Hartini, and G. Stephani Saragih, "Linear discriminant analysis and support vector machines for classifying breast cancer," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 253–256, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp253-256.

[6] M. Minnoor and V. Baths, "Diagnosis of breast cancer using random forests," *Procedia Computer Science*, vol. 218, pp. 429–437, 2023, doi: 10.1016/j.procs.2023.01.025.

[7] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Busan, South Korea, Dec. 2018, pp. 1–5, doi: 10.1109/ICECOCS.2018.8610632.

[8] A. Batool and Y.-C. Byun, "Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm," *IEEE Access*, vol. 12, pp. 12869–12882, 2024, doi: 10.1109/ACCESS.2024.3356602.

[9] C. Ricciardi *et al.*, "Linear discriminant analysis and principal component analysis to predict coronary artery disease," *Health Informatics Journal*, vol. 26, no. 3, pp. 2181–2192, Jan. 2020, doi: 10.1177/1460458219899210.

[10] K. Sivakami, "Mining big data: Breast cancer prediction using DT-SVM hybrid model," *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, vol. 1, no. 5, pp. 418–429, Aug. 2015.

[11] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *Journal of Personalized Medicine*, vol. 11, no. 2, Art. no. 61, Jan. 2021, doi: 10.3390/jpm11020061.

[12] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014, doi: 10.1016/j.eswa.2013.08.044.

[13] C. Boeri *et al.*, "Machine learning techniques in breast cancer prognosis prediction: A primary evaluation," *Cancer Medicine*, vol. 9, no. 9, pp. 3234–3243, 2020, doi: 10.1002/cam4.2953.

[14] W. H. Wolberg, O. L. Mangasarian, W. N. Street, and K. Street, "Breast Cancer Wisconsin (Diagnostic) Data Set," *UCI Machine Learning Repository*, 1992. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Can

[15] S. Sharma and S. Deshpande, "Breast cancer classification using machine learning algorithms," in *Lecture Notes in Networks and Systems*, Singapore: Springer, Oct. 2020, pp. 571–578, doi: 10.1007/978-981-15-7106-0_56.

[16] S. Wang, J. Lu, X. Gu, H. Du, and J. Yang, "Semi-supervised linear discriminant analysis for dimension reduction and classification," *Pattern Recognition*, vol. 57, pp. 179–189, 2016, doi: 10.1016/j.patcog.2016.02.019.

[17] P. H. Babu and E. S. Gopi, "Medical data classifications using genetic algorithm based generalized kernel linear discriminant analysis," *Procedia Computer Science*, vol. 57, pp. 868–875, 2015, doi: 10.1016/j.procs.2015.07.498.

[18] A. M. Marchevsky, J. A. Tsou, and I. A. Laird-Offringa, "Classification of individual lung cancer cell lines based on DNA methylation markers: Use of linear discriminant analysis and artificial neural networks," *The Journal of Molecular Diagnostics*, vol. 6, no. 1, pp. 28–36, 2004, doi: 10.1016/S1525-1578(10)60488-6.

[19] M. Toğaçar, B. Ergen, and Z. Cömert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," *Medical Hypotheses*, vol. 135, Art. no. 109503, 2020, doi: 10.1016/j.mehy.2019.109503.

[20] Z. Rustam and N. Maghfirah, "Correlated based SVM-RFE as feature selection for cancer classification using microarray databases," *AIP Conference Proceedings*, vol. 2023, no. 1, Art. no. 020235, 2018, doi: 10.1063/1.5064232.

[21] T. Nadira and Z. Rustam, "Classification of cancer data using support vector machines with features selection method based on global artificial bee colony," in *Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Sciences 2017 (ISCPMS2017)*, Bali, Indonesia, 2018, Art. no. 020205, doi: 10.1063/1.5064202.

[22] T. V. Rampisela and Z. Rustam, "Classification of schizophrenia data using support vector machine (SVM)," *Journal of Physics: Conference Series*, vol. 1108, no. 1, Art. no. 012044, 2018, doi: 10.1088/1742-6596/1108/1/012044.

[23] Z. Rustam and N. P. A. A. Ariantari, "Support vector machines for classifying policyholders satisfactorily in automobile insurance," *Journal of Physics: Conference Series*, vol. 1028, no. 1, Art. no. 012005, 2018, doi: 10.1088/1742-6596/1028/1/012005.

[24] Z. Rustam, D. F. Vibranti, and D. Widya, "Predicting the direction of Indonesian stock price movement using support vector machines and fuzzy kernel C-means," in *Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Sciences 2017 (ISCPMS2017)*, Bali, Indonesia, 2018, Art. no. 020207, doi: 10.1063/1.5064204.

[25] D. A. Puspitasari and Z. Rustam, "Application of SVM-KNN using SVR as feature selection on stock analysis for Indonesia stock exchange," in *Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Sciences 2017 (ISCPMS2017)*, Bali, Indonesia, 2018, Art. no. 020207, doi: 10.1063/1.5064204.

[26] Z. Rustam and A. A. Ruvita, "Application support vector machine on face recognition for gender classification," *Journal of Physics: Conference Series*, vol. 1108, no. 1, Art. no. 012067, 2018, doi: 10.1088/1742-6596/1108/1/012067.

[27] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000, doi: 10.1093/bioinformatics/16.10.906.

Phụ lục

A. Phân công công việc

- 23020356 Bùi Hải Đăng cài đặt mã nguồn, huấn luyện mô hình AI, nghiên cứu cải tiến mô hình.
- 23020351 Vũ Nguyên Đan phân tích, trực quan hóa dữ liệu, và lựa chọn mô hình huấn luyện.
- 23020364 Phan Tuấn Hiệp phân tích ưu nhược điểm mô hình, đề xuất phương pháp cải thiện.

Các thành viên trong nhóm đều cùng nhau cố gắng làm việc để hoàn thành công việc đúng tiến độ. Ba thành viên đều cùng nhau viết báo cáo và hoàn thiện slide trình bày.

B. Khó khăn gặp phải

- Có nhiều mô hình học máy dùng để phân loại nên làm cho khâu chọn mô hình huấn luyện gặp khó khăn.
- Tập dữ liệu còn hạn chế nên mô hình chưa có tính tổng quát hóa cao.
- Việc lựa chọn kernel, C và gamma phù hợp cho SVM yêu cầu thử nghiệm nhiều lần và tiêu tốn thời gian tính toán.
- Cần hiểu rõ nguyên lý của LDA và cách kết hợp hợp lý với SVM để tránh làm mất thông tin quan trọng.

C. Đề xuất và hướng phát triển

- Thử và áp dụng các phương pháp chọn đặc trưng khác như PCA, RFE để nâng cao chất lượng phân loại.
- Sử dụng các tập dữ liệu lớn hơn để tăng độ đa dạng và tính tổng quát của mô hình.
- Áp dụng các kỹ thuật cân bằng dữ liệu như SMOTE để xử lý vấn đề mất cân đối giữa hai lớp trong tập dữ liệu.
- Phát triển giao diện ứng dụng (app/web) để triển khai mô hình, hỗ trợ bác sĩ trong công tác chẩn đoán.