Lung cancer classification model using histopathological images

Phạm Thanh Lâm
Khoa: Khoa học Máy tính
Trường Đại học Công nghệ Thông tin
TP. Hồ Chí Minh, Việt Nam
21520055@gm.uit.edu.vn

Đào Duy Thịnh
Khoa: Khoa học Máy tính
Trường Đại học Công nghệ Thông tin
TP. Hồ Chí Minh, Việt Nam
21520463@gm.uit.edu.vn

Nguyễn Nhật Nam
Khoa: Khoa học Máy tính
Trường Đại học Công nghệ Thông tin
TP. Hồ Chí Minh, Việt Nam
21521160@gm.uit.edu.vn

TỔNG QUAN ĐỀ TÀI

Trong nghiên cứu này, nhóm đề xuất Mô hình phân loại ung thư phổi bằng cách sử dụng Hình ảnh mô bệnh học. Ung thư phổi vẫn là nguyên nhân hàng đầu gây tử vong liên quan đến ung thư trên toàn cầu, đòi hỏi các phương pháp phát hiện sớm hiệu quả. Hình ảnh mô bệnh học cung cấp thông tin tế bào chi tiết rất quan trọng để chẩn đoán chính xác. Đề tài của nhóm tận dụng các kỹ thuật học máy để tự động phân tích và phân loại hình ảnh mô bệnh học thành các loại lành tính và ác tính. Thông qua thử nghiệm và xác nhận rộng rãi trên tập dữ liệu đa dạng, mô hình đề xuất của nhóm cho thấy kết quả đầy hứa hẹn về độ chính xác và độ tin cậy. Nghiên cứu này góp phần thúc đẩy sự tiến bộ của chẩn đoán ung thư phổi tự động, cung cấp một công cụ đáng tin cậy cho các chuyên gia chăm sóc sức khỏe trong việc phát hiện sớm và lập kế hoạch điều trị ung thư.

I. GIỚI THIỆU BÀI TOÁN

A. *Mục tiêu cụ thể của đồ án*

Hình ảnh mô bệnh học rất hiệu quả trong việc nghiên cứu tình trạng của các cấu trúc sinh học và chẩn đoán các bệnh như ung thư. Tuy nhiên việc kiểm tra thủ công những hình ảnh này hiện nay tốn thời gian. Do đó cần có các kỹ thuật tự động đáng tin cậy để phân loại hiệu quả các hình ảnh ung thư bình thường và ác tính.

Mục tiêu của đồ án này là huấn luyện các mô hình máy học để phân loại các hình ảnh mô bệnh học của phổi, nhằm hỗ trợ việc chẩn đoán và phát hiện các loại ung thư phổi. Bằng cách sử dụng các kỹ thuật học máy tiên tiến và xử lý hình ảnh, đồ án này không chỉ nhằm mục đích xây dựng một hệ thống phân loại hình ảnh hiệu quả, mà còn hướng đến việc ứng dụng công nghệ vào y học để cải thiện cuộc sống và sức khỏe của con người.

B. *Bài toán máy học*

Bài toán máy học được đặt ra trong đồ án này là bài toán phân loại đa lớp, trong đó mục tiêu là xây dựng các mô hình học máy có khả năng nhận diện và phân loại chính xác các hình ảnh mô bệnh học phổi vào một trong ba loại bệnh lý là: ung thư biểu mô tuyến phổi (Lung adenocarcinoma), ung thư biểu mô tế bào vảy phổi (Lung squamous cell carcinoma), tế bào lành tính (Lung benign tissue).

C. *Ứng dụng*

Các mô hình máy học được phát triển trong đồ án này có tiềm năng ứng dụng rộng rãi trong lĩnh vực y tế, đặc biệt là trong việc chẩn đoán và phát hiện ung thư phổi. Một số ứng dụng cụ thể bao gồm:

Hỗ trợ chẩn đoán: Hệ thống phân loại hình ảnh mô bệnh học có thể được tích hợp vào quy trình chẩn đoán tại các bệnh viện và phòng khám, giúp các bác sĩ phát hiện và phân loại các loại ung thư phổi một cách nhanh chóng và chính xác.

Giảm tải công việc cho bác sĩ: Việc tự động hóa quá trình phân tích hình ảnh giúp giảm tải công việc cho các bác sĩ, tiết kiệm thời gian và nguồn lực y tế, đồng thời tăng hiệu quả và độ chính xác của chẩn đoán.

Cải thiện chất lượng chăm sóc sức khỏe: Bằng cách ứng dụng công nghệ máy học tiên tiến, hệ thống này góp phần nâng cao chất lượng chăm sóc sức khỏe, giúp phát hiện sớm và điều trị kịp thời các bệnh lý về phổi, từ đó cải thiện sức khỏe và cuộc sống của bệnh nhân.

II. DỮ LIỆU

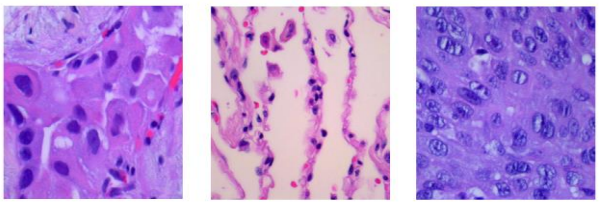
A. *Giới thiệu tập dữ liệu*

Tập dữ liệu được sử dụng trong dự án này bao gồm 15.000 hình ảnh mô bệnh học được phân thành 3 loại. Tất cả các hình ảnh đều có kích thước 768 x 768 pixel và được lưu trữ dưới định dạng tệp JPEG. Có ba loại trong tập dữ liệu, mỗi loại gồm 5.000 hình ảnh:

Lung Adenocarcinoma (ung thư tuyến phổi): Đây là loại ung thư phổi phổ biến nhất, xuất phát từ các tế bào tuyến trong phổi.

Lung Benign Tissue (mô lành tính phổi): Đây là các mô phổi không có dấu hiệu ung thư, cần được phân biệt rõ ràng để tránh chẩn đoán sai lệch.

Lung Squamous Cell Carcinoma (ung thư tế bào vảy phổi): Đây là một loại ung thư phổi khác, phát triển từ các tế bào vảy trong phổi.



Hình 1: Ba mẫu dữ liệu minh họa

B. *Phân tích bộ dữ liệu*

1) Nguồn gốc hình ảnh

Các hình ảnh mô phổi ban đầu được thu thập từ các nguồn đã được xác thực và tuân thủ HIPAA, gồm tổng cộng 1.250 hình ảnh gốc (750 hình ảnh mô phổi và 500 hình ảnh mô đại tràng).

Các hình ảnh này sau đó được tăng cường bằng cách sử dụng gói Augmentor để đạt được tổng số 15.000 hình ảnh, đảm bảo tính đa dạng và phong phú cho tập dữ liệu.

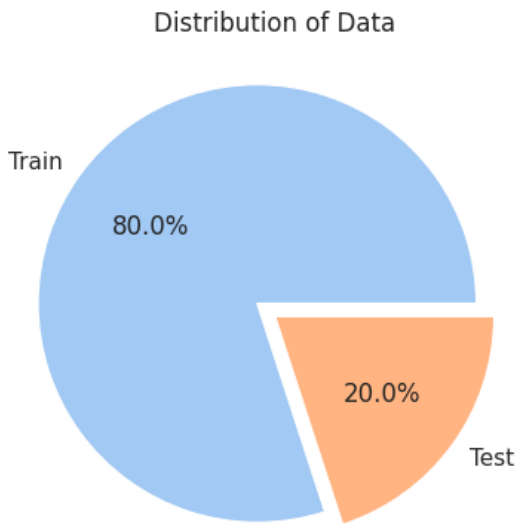
Link dữ liệu: LC25000 LUNG HISTOPATHOLOGICAL IMAGE DATASET

2) Đặc điểm hình ảnh

Các hình ảnh trong tập dữ liệu có độ phân giải cao, giúp cho việc phân tích và chẩn đoán chính xác hơn. Mỗi hình ảnh đều được lưu trữ dưới định dạng tệp JPEG, giúp dễ dàng trong việc xử lý và phân tích bằng các công cụ máy học.

3) Phân chia dữ liệu

Dữ liệu được chia thành hai tập: train và test. Phân chia này được thực hiện theo tỉ lệ 80%, 20%. Tập dữ liệu này sẽ được sử dụng để huấn luyện và đánh giá các mô hình máy học trong dự án, nhằm xây dựng một hệ thống phân loại hình ảnh mô bệnh học chính xác và hiệu quả.



Hình 2: Biểu đồ thể hiện sự phân chia bộ dữ liệu

C. Xử lý dữ liệu

Bắt đầu với thiết lập các thông số cơ bản của ảnh như chiều cao, chiều rộng, số kênh, kích thước batch. Sau đó chuẩn hoá dữ liệu hình ảnh bằng cách chia các giá trị pixel của ảnh cho 255 để đưa về khoảng từ 0 đến 1. Sử dụng GLCM để xử lý dữ liệu huấn luyện. GLCM đo lường cách mà các mức xám của các pixel trong ảnh xuất hiện cùng nhau ở một khoảng cách và góc độ nhất định. Mô hình này sử dụng các đặc trưng như độ tương phản, năng lượng, tính đồng nhất, sự tương quan và entropy để biểu diễn kết cấu của ảnh. Việc tiền xử lý này giúp trích xuất các đặc trưng quan trọng từ ảnh, làm đầu vào cho các mô hình học máy để cải thiện độ chính xác và hiệu suất của quá trình phân loại.

III. PHƯƠNG PHÁP MÁY HỌC

Trong phần này, nhóm trình bày chi tiết các phương pháp máy học được áp dụng trong đồ án, bao gồm mô tả ngắn gọn về từng phương pháp, các cải tiến, công cụ hỗ trợ và các giá trị tham số cụ thể được nhóm sử dụng trong từng phương pháp. Trên cơ sở kết quả thu được, nhóm sử dụng các độ đo để đánh giá kết quả trên.

A. Các phương pháp máy học

1) Support Vector Machine

Support Vector Machine (SVM) là một phương pháp mạnh mẽ trong machine learning được sử dụng cho các bài toán phân loại và hồi quy.

Nguyên lý hoạt động: SVM hoạt động bằng cách tìm một siêu phẳng (hyperplane) trong không gian đặc trưng để tối ưu hóa việc phân tách các lớp dữ liệu khác nhau. Siêu phẳng này được chọn sao cho khoảng cách từ các điểm dữ liệu gần nhất tới siêu phẳng là lớn nhất, giúp giảm thiểu lỗi phân loại.

Tối ưu hóa và tham số:

C: Tham số điều chỉnh mức độ phạt đối với các mẫu phân loại sai. Giá trị nhỏ của C làm mô hình trở nên đơn giản hơn và ít phức tạp hơn, trong khi giá trị lớn của C cố gắng phân loại chính xác hơn các mẫu huấn luyện.

Gamma: Tham số điều chỉnh mức độ ảnh hưởng của một điểm dữ liệu đến đường biên. Giá trị nhỏ của gamma cho phép các điểm dữ liệu xa có ảnh hưởng lớn hơn, trong khi giá trị lớn của gamma tập trung vào các điểm gần hơn.

Ưu điểm nổi bật: SVM có thể sử dụng kernel tuyến tính để phân tách các lớp dữ liệu trong không gian đặc trưng gốc. Kernel tuyến tính đơn giản hóa mô hình và hiệu quả trong các bài toán có dữ liệu có tính tuyến tính hoặc gần tuyến tính, mà không cần phải điều chỉnh nhiều tham số.

SVM là một trong những phương pháp được ưa chuộng nhờ vào khả năng hiệu quả trong phân loại và hồi quy, đặc biệt là khi dữ liệu có cấu trúc rõ ràng và khi sử dụng kernel tuyến tính.

Các tham số của phương pháp SVM:

```
param_grid_svm = {
    'classifier__C': [0.1, 1, 10],
    'classifier__gamma': ['scale', 'auto']
}
```

Hình 3: Giá trị tham số của phương pháp svm

2) eXtreme Gradient Boosting

eXtreme Gradient Boosting là một phương pháp học máy mạnh mẽ dựa trên kỹ thuật boosting, thường được sử dụng cho các bài toán phân loại, hồi quy và xếp hạng. Dưới đây là tóm tắt cách XGBoost hoạt động:

Boosting: Kết hợp nhiều mô hình yếu, thường là cây quyết định, để tạo ra một mô hình mạnh. Mỗi mô hình yếu được huấn luyện để sửa các lỗi của mô hình trước đó.

XGBoost tối ưu hóa hiệu suất, giúp giảm thời gian huấn luyện và dự đoán nhờ sử dụng kỹ thuật cây quyết định và gradient boosting. Có thể xử lý dữ liệu lớn và hỗ trợ tính toán phân tán, làm cho XGBoost phù hợp với các ứng dụng thực tế đòi hỏi khối lượng dữ liệu lớn.

Giảm thiểu overfitting: Tích hợp các cơ chế như regularization để giảm thiểu hiện tượng overfitting, đảm bảo mô hình không quá khớp với dữ liệu huấn luyện.

Hỗ trợ nhiều loại bài toán khác nhau và cho phép tùy chỉnh tham số một cách linh hoạt. XGBoost hoạt động bằng cách liên tục xây dựng các cây quyết định để sửa các lỗi của cây trước đó cho đến khi đạt được độ chính xác mong muốn hoặc khi số vòng lặp tối đa được chỉ định.

Các tham số của phương pháp XGBoost

```
param_grid_xgb = {
    'classifier__n_estimators': [3000, 1000, 500],
    'classifier__max_depth': [10, 15, 7],
    'classifier__learning_rate': [0.5, 0.3, 0.2]
}
```

Hình 4: Giá trị tham số của phương pháp xgboost

3) RandomForestClassifier

RandomForestClassifier là một thuật toán học máy thuộc nhóm học có giám sát, được ứng dụng trong giải quyết các bài toán hồi quy và phân loại.

Thuật toán Random Forest bắt đầu bằng cách lấy mẫu ngẫu nhiên từ bộ dữ liệu gốc bằng kỹ thuật Bootstrapping (random

sampling with replacement), giúp cho phép chọn lại các phần tử đã chọn trước đó sau khi đã lấy mẫu, để tạo ra nhiều mẫu khác nhau từ bộ dữ liệu gốc.

Sau khi tạo mẫu ngẫu nhiên từ bộ dữ liệu, tiến hành xây dựng cây quyết định dựa trên mẫu này. Cây quyết định là một cấu trúc dữ liệu gồm các nút (node) và nhánh (branch), mỗi nhánh đại diện cho một thuộc tính (feature).

Cuối cùng, thuật toán Random Forest kết hợp kết quả dự đoán của nhiều cây quyết định để đưa ra dự đoán cuối cùng, thường bằng cách trung bình hoặc lấy kết quả đa số từ các cây quyết định.

Random Forest giúp giảm thiểu overfitting và mang lại kết quả chính xác và ổn định hơn trong các bài toán phân loại và hồi quy so với việc sử dụng một cây quyết định duy nhất.

Các tham số của phương pháp RandomForestClassifier

```
param_grid = {
    'classifier__n_estimators': [100, 200, 300],
    'classifier__max_depth': [None, 10, 20, 30],
    # Add more hyperparameters as needed
}
```

Hình 5: Giá trị tham số của phương pháp Randomforest Classifier

4) Công cụ sử dụng

Huấn luyện mô hình với các phương pháp nêu trên bằng Kaggle với accelerator là GPU T4x2.

B. Các độ đo đánh giá

1) Accuracy

Độ chính xác là tỷ lệ dự đoán đúng trên tổng số mẫu và là một trong những độ đo phổ biến nhất để đánh giá hiệu suất của mô hình. Công thức tính độ chính xác là:

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

Hình 6: Công thức tính Accuracy

2) Loss Function

Hàm mất mát được sử dụng để đánh giá mức độ sai số của mô hình. Mục tiêu của quá trình huấn luyện là giảm thiểu giá trị của hàm mất mát. Các hàm mất mát phổ biến bao gồm Mean Squared Error (MSE) cho bài toán hồi quy và Cross-Entropy Loss cho bài toán phân loại. Công thức của Mean Squared Error là:

$$\boldsymbol{MSE} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}}$$

Hình 7: Công thức tính loss function

3) F1_score

F1 Score là một độ đo kết hợp giữa độ chính xác (Precision) và độ nhạy (Recall), và là độ đo quan trọng trong các bài toán phân loại, đặc biệt là khi có sự mất cân bằng giữa các lớp. Công thức tính F1 Score là:

$$\mathbf{F1} = \mathbf{2} \times \frac{\mathbf{Precision + Recall}}{\mathbf{Precision \times Recall}}$$

Hình 8: Công thức tính độ đo F1_score

Trong đó, Precision (Độ chính xác) và Recall (Độ nhạy) được tính như sau:

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

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

Hình 9: Công thức tính độ đo Precision và Recall

Độ đo F1 Score cao đồng nghĩa với mô hình có sự cân bằng tốt giữa độ chính xác và độ nhạy.

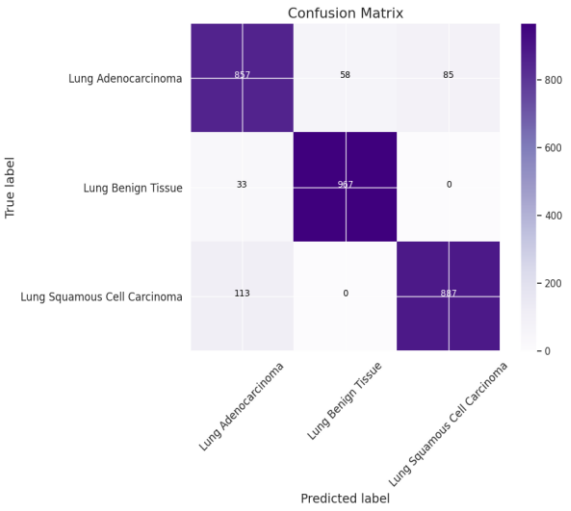
Các phương pháp máy học và các độ đo đánh giá được trình bày ở trên sẽ giúp nhóm xây dựng và đánh giá hiệu quả các mô hình dự đoán, từ đó chọn ra mô hình tốt nhất để áp dụng vào bài toán trong đồ án này.

IV. Các thử nghiệm và tinh chỉnh mô hình

Thực hiện các thử nghiệm và dùng tìm kiếm grid search cross-validation với k-fold cross-validation để thực hiện tinh tỉnh các siêu tham số của các mô hình từ 3 phương pháp mà nhóm đã tìm hiểu

A. Support vector machine

Best model parammeters: { 'C': 10, 'gamma': 'scale'}
Mean Cross-Validation Accuracy: 0.9006666666666666
Đánh giá trên tập kiểm tra:
Accuracy_score: 0.9036666666666666
Confusion matrix:



Hình 10: Confusion matrix (svm)

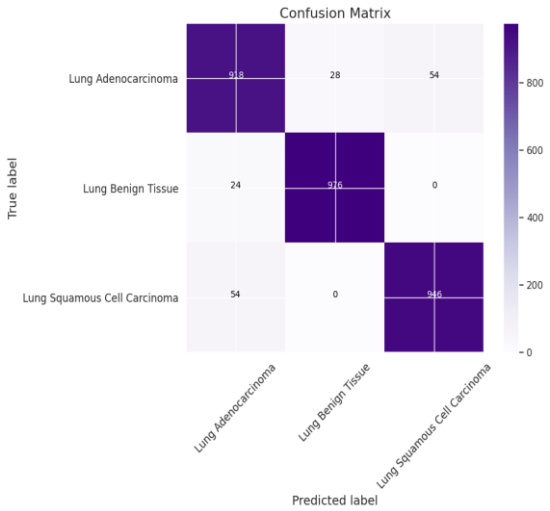
Classification report:

	precision	recall	f1-score	support
0	0.85	0.86	0.86	1000
1	0.94	0.97	0.96	1000
2	0.91	0.89	0.90	1000
accuracy			0.90	3000
macro avg	0.90	0.90	0.90	3000
weighted avg	0.90	0.90	0.90	3000

Hình 11: Classification report (svm)

B. eXtreme Gradient Boosting

Best model parammeters: { 'learning_rate': 0.2, 'max_depth': 15, 'n_estimators': 500}
Mean Cross-Validation Accuracy: 0.9368333333333332
Đánh giá trên tập kiểm tra:
Accuracy_score: 0.9466666666666667
Confusion matrix:



Hình 12: Confusion matrix (xgb)

Classification report:

	precision	recall	f1-score	support
0	0.92	0.92	0.92	1000
1	0.97	0.98	0.97	1000
2	0.95	0.95	0.95	1000
accuracy			0.95	3000
macro avg	0.95	0.95	0.95	3000
weighted avg	0.95	0.95	0.95	3000

Hình 13: Classification report (xgb)

C. *RandomForestClassifier*

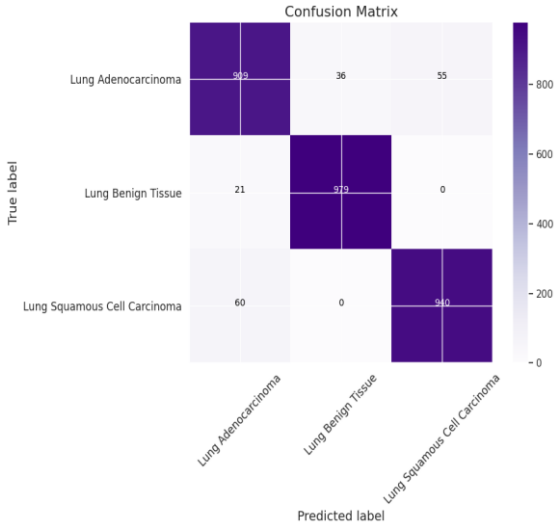
Best model parammeters: { 'max_depth': 30, 'n_estimators': 300}

Mean Cross-Validation Accuracy: 0.9375833333333332

Đánh giá trên tập kiểm tra:

Accuracy_score: 0.9426666666666667

Confusion matrix:



Hình 14: Confusion matrix (randomforestclassifier)

Classification report:

	precision	recall	f1-score	support
0	0.92	0.91	0.91	1000
1	0.96	0.98	0.97	1000
2	0.94	0.94	0.94	1000
accuracy			0.94	3000
macro avg	0.94	0.94	0.94	3000
weighted avg	0.94	0.94	0.94	3000

Hình 15: Classification report (randomforestclassifier)

V. PHÂN TÍCH LỖI VÀ HƯỚNG PHÁT TRIỂN

A. *Phân tích lỗi*

Tuy các phương pháp máy học trên cho kết quả dự đoán phân loại tốt nhưng vẫn còn trong khoảng 5 đến 10% các mẫu dữ liệu kiểm tra bị phân loại sai. Để tiến hành phân tích lỗi mà các mô hình dự đoán sai nhóm đã thực hiện tổng hợp một số thông tin về các đặc trưng của tập dữ liệu huấn luyện.

Giá trị trung bình của các feature độ tương phản, năng lượng, tính đồng nhất, sự tương quan và entropy dựa vào label:

labels	Contrast	Energy	Homogeneity	Correlation	Entropy
0	16.710074	0.038100	0.386385	0.994817	7.442596
1	14.413606	0.068996	0.494097	0.995265	6.782775
2	16.550858	0.034061	0.338671	0.990997	7.359972

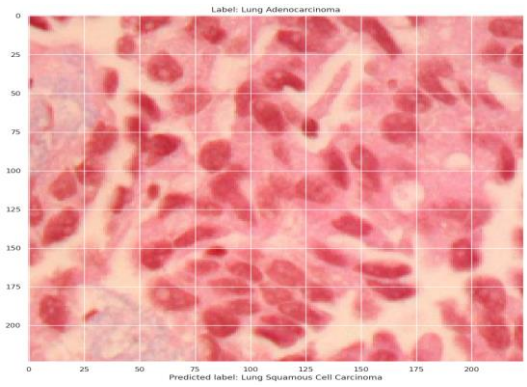
Hình 16: Bảng giá trị trung bình các đặc trưng trong dữ liệu train

Giá trị nhỏ nhất và lớn nhất của các feature độ tương phản, năng lượng, tính đồng nhất, sự tương quan và entropy dựa vào label:

Labels	Contrast		Energy		Homogeneity		Correlation		Entropy	
	min	max	min	max	min	max	min	max	min	max
0	3.958741	49.568238	0.018237	0.109676	0.241039	0.608213	0.983601	0.998760	5.888516	8.427191
1	2.236202	33.123538	0.032220	0.162712	0.360237	0.711069	0.989911	0.998242	4.585474	7.692064
2	2.898670	58.998784	0.018417	0.069802	0.187989	0.555759	0.973075	0.997422	6.035305	8.374600

Hình 17: Giá trị nhỏ nhất và lớn nhất của các feature

Mô hình SVM nhận biết khá nhiều mẫu có nhãn là Lung Adenocarcinoma sai với tổng mẫu sai là 143/1000 mẫu. Dưới đây là 1 mẫu dự đoán sai mà mô hình phân loại nhãn là Lung Squamous Cell Carcinoma tuy nhiên nhãn thực tế là Adenocarcinoma.



Hình 18: Mẫu Lung Adenocarcinoma

Ta có thể thấy rằng các đặc trưng GLCM giữa hai nhãn Lung Adenocarcinoma (0) và Lung Squamous Cell Carcinoma (2) có sự trùng lặp và tương đồng lớn. Giá trị trung bình của các đặc trưng này gần như tương đương và phạm vi giá trị của chúng cũng có sự trùng lặp. Điều này làm cho mô

hình SVM khó phân biệt giữa hai loại ung thư này, dẫn đến việc dự đoán nhầm lẫn..

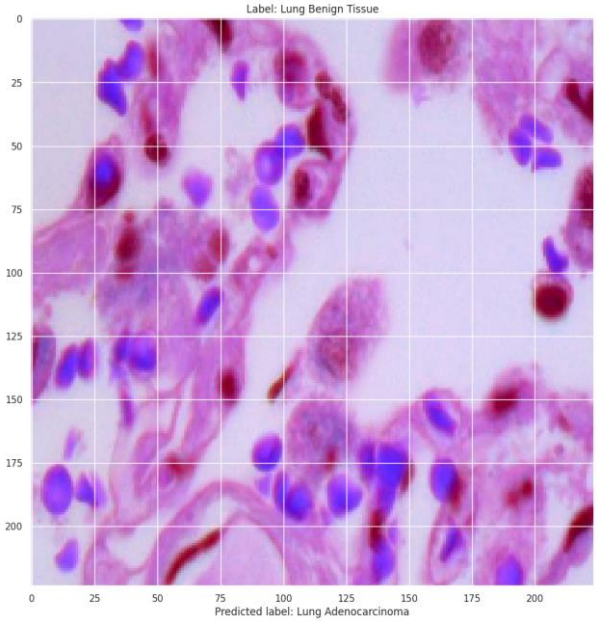
Do các đặc trưng của hai nhãn Lung Adenocarcinoma (0) và Lung Squamous Cell Carcinoma (2) có sự tương đồng lớn và chồng chéo, SVM có thể gặp khó khăn trong việc phân tách chúng một cách chính xác. Hạn chế này chủ yếu xuất phát từ cách SVM hoạt động, cần tìm một siêu phẳng phân tách các lớp trong không gian đặc trưng. Khi dữ liệu chồng chéo và không có đặc trưng phân biệt rõ ràng, SVM có thể không đủ mạnh để phân loại chính xác, dẫn đến các dự đoán nhầm lẫn.

Để cải thiện sự hạn chế của phương pháp svm nhóm đã thực hiện đánh giá thêm trên 2 phương pháp học máy khác đó là Random Forest, Gradient Boosting.

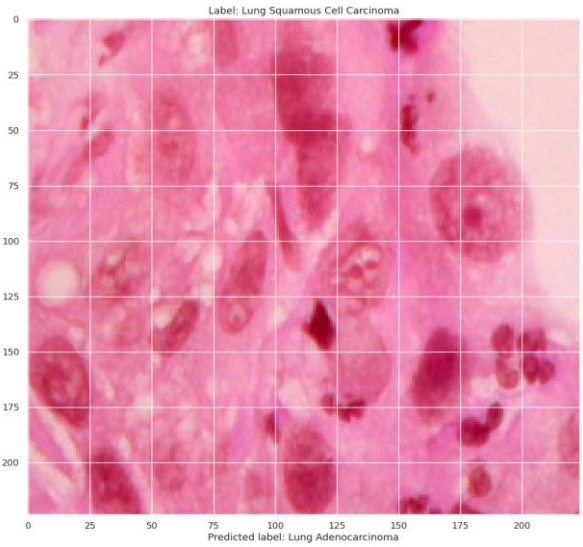
Dựa vào độ chính xác accuracy mà 2 phương pháp này đánh giá trên tập kiểm tra ta thấy rằng 2 mô hình này đã cải thiện khoảng 4% về độ chính xác và 5% độ nhạy đối với lớp 2 lớp là Lung Adenocarcinoma (0) và Lung Squamous Cell Carcinoma.

Sử dụng các mô hình như Random Forest và XGBoost có thể cải thiện độ chính xác (accuracy), độ chính xác (precision), và độ thu hồi (recall) so với SVM trong các bài toán phân loại, đặc biệt là khi dữ liệu có tính chồng chéo và không tuyến tính rõ ràng. Bởi vì với khả năng linh hoạt trong việc học các mối quan hệ phi tuyến tính, xử lý tốt dữ liệu chồng chéo và nhiễu, cùng với việc tối ưu hóa tham số hiệu quả. Điều này làm cho các mô hình ensemble như Random Forest và các mô hình boosting như XGBoost trở nên mạnh mẽ và hiệu quả hơn trong nhiều tình huống phân loại phức tạp.

Tuy nhiên vẫn còn một số ít mẫu dữ liệu vẫn bị mô hình phân loại nhầm lẫn. Nhưng với kết quả của các phương pháp đánh giá trên tập dữ liệu kiểm tra riêng biệt vẫn nhận được kết quả rất tốt cho ba mô hình SVM, XGBosst và Random Forest Classifier lần lượt là 90.37, 94.67, 94.27% .



Hình 19: Mẫu phân loại sai bởi mô hình Random Forest



Hình 20: Mẫu phân loại sai bởi mô hình XGBoost

B. Hướng phát triển

Để tiếp tục phát triển và cải thiện bài toán nhóm chúng em cân nhắc đến các hướng phát triển như sau:

Tăng cường dữ liệu: Sử dụng các kỹ thuật tăng cường dữ liệu như xoay, lật, dịch chuyển, thay đổi độ sáng, và thêm nhiễu để tạo ra nhiều mẫu huấn luyện hơn và cải thiện tính tổng quát của mô hình.

Khai thác đặc trưng mới: Tìm kiếm và trích xuất thêm các đặc trưng từ ảnh mà có thể mang lại thông tin phân biệt tốt hơn giữa các lớp.

Chọn lọc đặc trưng: Sử dụng các kỹ thuật như PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), hoặc các phương pháp chọn lọc đặc trưng khác để giảm chiều dữ liệu và giữ lại những đặc trưng quan trọng nhất.

Sử dụng các mô hình tiên tiến (Advanced Models): Deep Learning: Áp dụng các mô hình học sâu như CNN (Convolutional Neural Network) để tự động trích xuất các đặc trưng và phân loại hình ảnh. Các mô hình như ResNet, Inception, VGGNet, và EfficientNet có thể mang lại hiệu suất cao.

Transfer Learning: Sử dụng các mô hình học sâu đã được huấn luyện trước trên các tập dữ liệu lớn và tinh chỉnh (fine-tuning) trên tập dữ liệu cụ thể của bạn.

Ensemble Learning: Kết hợp nhiều mô hình khác nhau như Random Forest, XGBoost, và các mô hình deep learning để tạo ra một mô hình ensemble mạnh mẽ hơn.

Triển khai và tích hợp (Deployment and Integration) Triển khai mô hình đã huấn luyện vào môi trường thực tế để sử dụng trong ứng dụng phân loại hình ảnh.

VI. KẾT LUẬN

Trong nghiên cứu này, nhóm đã tiến hành so sánh hiệu suất của ba mô hình máy học: SVM, Random Forest và XGBoost trong việc phân loại hình ảnh ung thư phổi dựa trên các đặc trưng GLCM. Kết quả thu được cho thấy sự khác biệt rõ rệt cũng như các thế mạnh riêng của giữa các mô hình.

Đối với mô hình Support Vector Machine (SVM), độ chính xác trung bình khi kiểm tra chéo và tập kiểm tra là khoảng 0.9. Mô hình này đạt Độ chính xác trong khoảng 0.85 đến 0.94 và Độ nhạy trong khoảng 0.86 đến 0.97. Kết quả

trên cho thấy hiệu suất của mô hình SVM khá tốt nhưng vẫn có một số sai số nhất định.

Mô hình Random Forest cho thấy hiệu suất vượt trội hơn với độ chính xác kiểm tra chéo trung bình đạt khoảng 0.9375 và trên tập kiểm tra là xấp xỉ 0.9427. Mô hình này đạt Độ chính xác trong khoảng 0.92 đến 0.96 và Độ nhạy trong khoảng 0.91 đến 0.98. Random Forest có khả năng phân loại chính xác cao hơn và giảm thiểu sai số.

Cuối cùng, mô hình XGBoost đạt hiệu suất cao nhất với độ chính xác kiểm tra chéo trung bình đạt khoảng 0.9368 và trên tập kiểm tra là xấp xỉ 0.9467. Mô hình này có Độ chính xác đạt được trong khoảng 0.92 đến 0.97 và Độ nhạy trong khoảng từ 0.92 đến 0.98. Với kết quả trên nhóm nhận thấy rằng XGBoost không chỉ đạt độ chính xác cao mà còn thể hiện độ ổn định tốt trong phân loại.

Cả hai mô hình Random Forest và XGBoost đều vượt trội hơn so với SVM về tất cả các chỉ số đánh giá, bao gồm Độ chính xác, Độ nhạy và F1 Score. Mặc dù cả hai mô hình đều cho thấy hiệu suất tương tự, Random Forest có lợi thế hơn XGBoost trong việc xử lý các đặc trưng từ dữ liệu hình ảnh ung thư phổi. Tuy nhiên, sự khác biệt này là không đáng kể và cả hai mô hình đều có thể được coi là lựa chọn tốt cho nhiệm vụ phân loại hình ảnh này.

REFERENCES

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232.K. Elissa, “Title of paper if known,” unpublished.

[3] Huynh Chi Trung. Giới thiệu về Support Vector Machine (SVM) https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB

[4] Trí Tuệ Nhân Tạo. https://trituenhantao.io/kien-thuc/svm-qua-kho-hieu-hay-doc-bai-nay/

[5] Random Forest algorithm https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html