Covid-19 Detection From CT Scan Images

Phạm Vĩ, Phạm Nhật Anh, Võ Thị Bích Ngọc
19521101@gm.uit.edu.vn, 19521207@gm.uit.edu.vn, 19520781@gm.uit.edu.vn

Trường Đại học Công Nghệ Thông Tin, Đại học Quốc gia Thành phố Hồ Chí Minh

Tóm tắt nội dung Sự xuất hiện của đại dịch Covid-19 gây ảnh hưởng xấu đến nền kinh tế và y tế trên toàn thế giới. Virus dễ dàng lây lan trong cộng đồng thông qua tiếp xúc, giọt bắn khi nói chuyện hay hắc hơi. Vì thế, việc phát hiện người nhiễm bệnh, ngăn chặn nguồn lây nhiễm là việc rất quan trọng trong công tác phòng chống dịch. Hiện nay,để phát hiện bệnh nhân nhiễm Covid-19, thường sử dụng 2 phương pháp xét nghiệm phổ biến là xét nghiệm phân tử (Molecular Test) và xét nghiệm kháng nguyên (Antigen Test) phát hiện các protein trên bề mặt của virus Covid-19. Tuy nhiên, việc áp dụng 2 biệp pháp này cũng có những điểm hạn chế về mặt thời gian, độ chính xác. Để góp phần chung tay đẩy lùi đại dịch này, chúng tôi tiến hành xây dựng mô hình để phát hiện Covid-19 qua hình ảnh chụp CT phổi của bệnh nhân với các mô hình máy và học sâu trên bộ dữ liệu [4] . Dựa trên các mô hình: Convolutional Neural Network (CNN), Logistic Regresstion (LR), Random Forest Classification (RFC) và K-Nearest Neighbors (KNN). Kết quả thu được khá tốt với độ đo F1-Score lần lượt là 0.98, 0.86, 0.95, 0.98 trên tập dữ liệu test. Sau đó, chúng tôi tiến hành thống kê kết quả trên các mô hình để đưa ra kết luận và chọn ra mô hình học có hiệu quả tốt nhất trên bộ dữ liệu [4] và đưa ra phương hướng phát triển mô hình trong tương lai.

Keywords: Covid-19 · Phân loại · Phương pháp học sâu · Thị giác máy tính · Phương pháp học máy

1 Giới thiệu

COVID-19 là một căn bệnh truyền nhiễm gây viêm phổi cấp do virus SARS-CoV-2 gây ra. Các ca bệnh đầu tiên được phát hiện vào tháng 12 năm 2019 tại Vũ Hán, tỉnh Hồ Bắc, Trung Quốc và sau đó lan ra toàn cầu. Virus lây từ người này sang người kia thông qua tiếp xúc với dịch cơ thể của người bệnh. Việc ho, nói chuyện, hắt hơi hay bắt tay có thể khiến người xung quanh bị phơi nhiễm. Đã có hơn 238 triệu ca nhiễm, hơn 4,82 triệu ca tử vong trên toàn thế giới - tháng 12 năm 2021, và gây không ít ảnh hưởng tiêu cực đến nền kinh tế toàn cầu.

Mục tiêu của chúng tôi là xây dựng mô hình hỗ trợ việc phát hiện nhanh chóng bệnh nhân có nhiễm SARS-CoV-2 hay không thông qua ảnh chụp CT phổi. Từ đó, góp phần đến việc ngăn chặn bệnh chuyển biến nặng và tìm được phát đồ điều trị nhanh chóng.

2 Công trình liên quan

Ozal Yildirim – Chẩn đoán COVID19 tự động từ ảnh chụp X quang, sử dụng Deep Neural network. [1]
Mô hình được phát triển để cung cấp chẩn đoán chính xác cho phân loại nhị phân (COVID so với non-COVID) và phân loại nhiều lớp (COVID so với non-COVID và Viêm phổi). Mô hình tạo ra độ chính xác 98,08% cho các lớp nhị phân và 87,02% cho các trường hợp đa lớp. [2]
Mô hình DarkNet được sử dụng như một bộ phân loại cho hệ thống phát hiện đối tượng thời gian thực (YOLO).

Narendra Kumar Mishra–Chẩn đoán COVID19 từ ảnh chụp X quang sử dụng convolutional neural network(mạng nơ ron phức hợp). [3]

Tạo ra AI bằng cách sử dụng CNN và khả năng chẩn đoán của mô hình deep learning. Phương pháp Học chuyển giao, dựa trên kiến trúc VGG16 và ResNet50, đã được sử dụng để phát triển một thuật toán phát hiện COVID-19 từ hình ảnh chụp CT bao gồm các Khỏe mạnh, COVID-19 và Viêm phổi. Mô hình cho độ chính xác 99% trong trường hợp phân loại nhị phân. Trong trường hợp đa lớp, mô hình đạt độ chính xác 86.74% với VGG16 và 86.74% với ResNes50.

Hammam Alshazly, Thomas Martinetz, sử dụng 2 kiến trúc deep convolutional network mới CovidResNet và CovidDenseNet. [5]

Các mô hình cho phép học chuyển giao giữa các kiến trúc khác nhau, có thể tăng đáng kể hiệu suất chẩn đoán. Mô hình được khởi tạo 1 phần từ các mô hình cơ sở lớn như ResNet50 và DenseNet121. Mô hình dựa trên phân loại nhị phân COVID với khỏe mạnh, COVID với viêm phổi, non-COVID với khỏe mạnh. Mô hình đạt độ chính xác lên tới 93.87% accuracy, 99.13% precision, 95.70% F1-score cho các tác vụ nhị phân và lên tới 83.89% accuracy, 80.36% precision, 81.05% F1-score cho tác vụ phân loại 3 lớp.

SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification của nhóm nghiên cứu Eduardo Soares [6] bằng việc sử dụng eXplainable Deep Learning approach (xDNN) đã đạt được 97,31% điểm F1-score trên bộ dữ liệu SARS-COV-2 Ct-Scan Dataset.

3 Bộ dữ liệu

3.1 Định nghĩa tác vụ

COVID-19 Detection Task: Tác vụ này nhằm mục đích phát hiện xem bệnh nhân có bị nhiễm COVID-19 hay là không thông qua ảnh chụp CT phổi.
Tác vụ được mô tả như sau:
Input: một ảnh chụp CT phổi.
Output: COVID/non-COVID.
- Nhiễm COVID-19 (COVID) là ảnh chụp CT phổi của các bệnh nhân dương

tính với COVID-19.
- Không nhiễm COVID-19 (non-COVID) là ảnh chụp CT phổi của các bệnh nhân không bị nhiễm COVID-19.

3.2 Thông tin dữ liệu

Điểm dữ liệu với hai thuộc tính là ảnh và kết quả (Covid, non-Covid). Bộ dữ liệu được gộp từ 2 bộ dataset [4] và lấy từ Kaggle, bao gồm 16967 điểm dữ liệu với 8845 ảnh CT phổi người mắc Covid-19 và 8122 ảnh không mắc Covid-19 đã được gán nhãn sẵn. Với thuộc tính kết quả là thuộc tính cần được dự đoán. Bộ dữ liệu này được chia thành 2 phần train, test cho mỗi bên Covid và non-Covid theo tỉ lệ 8:2.

	Sample size	Covid	Non-Covid	Features
Training	13573	7084	6489	2
Test	3394	1761	1633	2
Total	16967	8845	8122	2

Bảng 1: Ví dụ về dữ liệu

Hình 1: Ảnh CT phổi người mắc Covid

Hình 2: Ảnh CT phổi người không mắc Covid

3.3 Thách thức bộ dữ liệu

Chất lượng của bộ dữ liệu để giải quyết vấn đề nhận dạng Covid-19 là một trong những thách thức khó khăn nhất mà chúng tôi phải đối mặt. Bộ dữ liệu hiện tại có đủ kích thước và chất lượng là không đủ, trong khi các mô hình thử nghiệm yêu cầu bộ dữ liệu lớn hơn để cải thiện hiệu suất phân loại. Hơn nữa, đề tài nhận biết chẩn đoán Covid-19 dựa trên chụp CT là khó hiểu vì đây vẫn là một đề tài mới với số lượng tài liệu hoặc công trình nghiên cứu liên quan tương đối ít, đặc biệt là các công trình nghiên cứu. Khó khăn của đề tài bắt nguồn từ bản chất của vấn đề đặt ra là xác định Covid-19 dựa trên chụp CT, điều này đòi hỏi người vận hành không chỉ có kiến thức về thực nghiệm.

4 Phương pháp tiếp cận

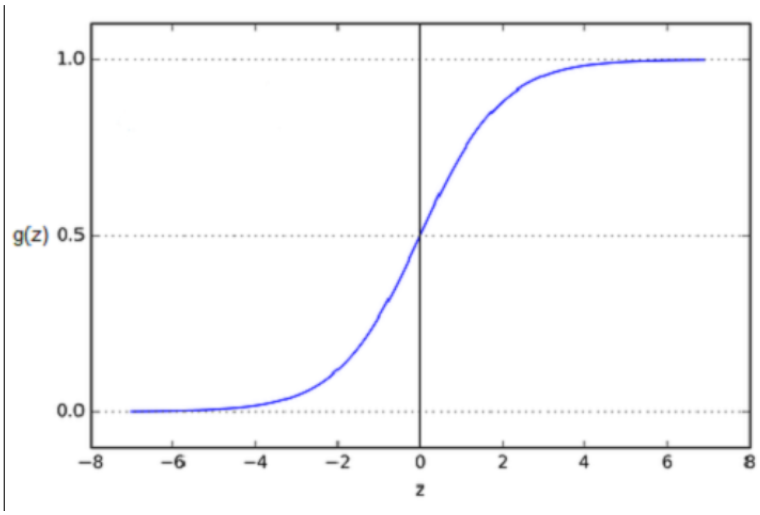
4.1 Tiền xử lý dữ liệu

Việc tiền xử lý dữ liệu là một bước vô cùng cần thiết, có ảnh hưởng tới các bước tiếp theo trong việc huấn luyện các mô hình học máy và học sâu đồng thời ảnh hưởng tới độ chính xác của mô hình. Bộ dữ liệu được lấy từ Kaggle về, sau khi hoàn thành việc gán nhãn (Covid, non-Covid) cho từng điểm dữ liệu, chúng tôi vẫn còn những khó khăn ảnh hưởng tới quá trình huấn luyện như là một số ảnh không cùng kích thước,... Do đó, chúng tôi đã áp dụng một số kỹ thuật tiền xử lý ảnh để giải quyết thách thức trên.
Như trên đã đề cập, chúng tôi xử lý định dạng lại kích thước toàn bộ các ảnh trong bộ dữ liệu vì sau nhiều lần huấn luyện mô hình chúng ta nhận thấy rằng việc huấn luyện với nhiều kích thước không đồng nhất sẽ dẫn đến độ chính xác của mô hình thấp.

4.2 Mô hình học máy

Logistic Regression là một thuật toán phân loại được dùng để phân lớp các đối tượng cho 1 tập hợp giá trị rời rạc (như 0, 1, 2, ...). Điển hình là bài toán phân

loại vật thể (con chó, con mèo,..). Thuật toán trên dùng hàm sigmoid logistic để đưa ra đánh giá theo xác suất. Learning là quá trình tìm, định lượng các hệ số W0, W1,... trong mô hình Logistic Regression từ tập dữ liệu có sẵn (training data set). Trong Machine Learning nói riêng và AI nói chung rất khó có thể năng tìm ra đúng, chính xác tuyệt đối các hệ số này mà chỉ cố gắng định lượng giá trị có khả năng cao nhất gần đúng hay nói cách khác là giảm thiểu tối đa sai sót, chênh lệch khi lắp ghép các hệ số B0, B1 này vào model so với kết quả thực tế. Công thức tổng quát Logistic Regression: g(z)= $\frac{1}{1+e-z}$

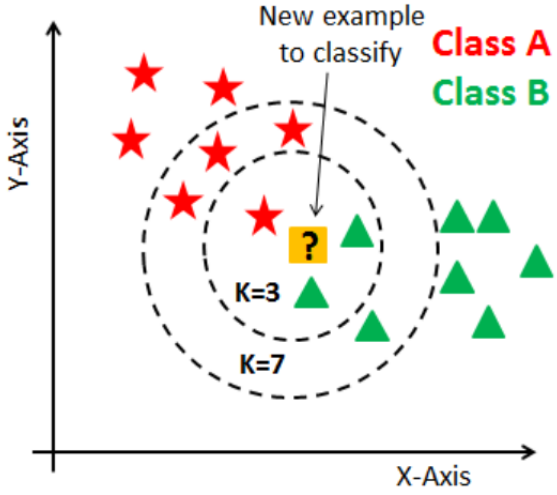


Hình 3: Đồ thị Logistic Regression

Prediction = 0 nếu g(z) $<$ 0.5 và Prediction = 1 nếu g(z) $>=$ 0.5 .

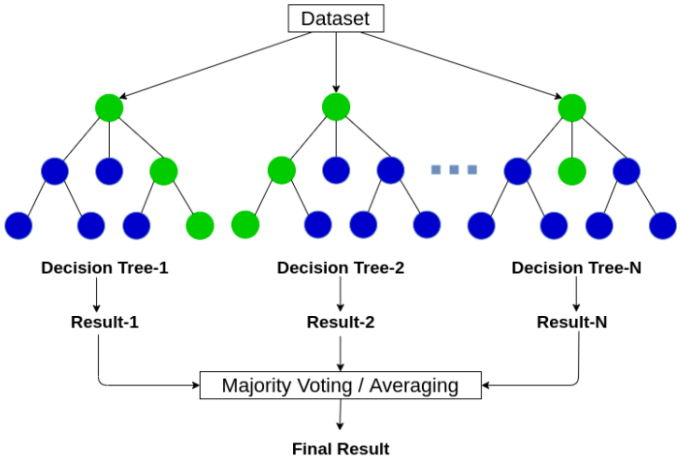
K-nearest neighbors (KNN) là một kĩ thuật học có giám sát (supervised learning) dùng để phân loại dữ liệu mới bằng cách tìm điểm tương đồng giữa dữ liệu mới này với dữ liệu sẵn có. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới.

Với KNN, trong bài toán Classification, label của một điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng major voting (bầu chọn theo số phiếu) giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó. KNN có thể áp dụng được vào cả hai bài toán phân lớp và hồi quy.



Hình 4: KNN

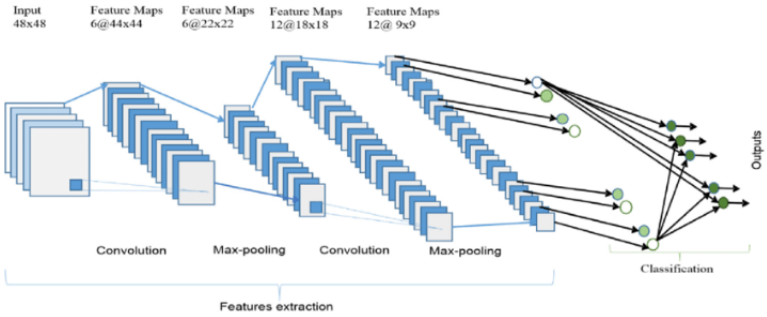
Random forest classification là thuật toán học có giám sát(supervised learning). Nó có thể được sử dụng cho cả phân lớp và hồi quy. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu.



Hình 5: Minh họa cho Random Forest Classification

4.3 Mô hình học sâu

Convolutional Neural Network (CNN) là một trong những mô hình Deep Learning tiên tiến. Được sử dụng nhiều trong các bài toán nhận dạng các object trong ảnh. CNN là một kiến trúc mạng nơ-ron đa tầng được phát triển để phục vụ các nhiệm vụ phân loại. Nó có thể phát hiện các tính năng kết hợp và đưa ra kết quả được phân loại bằng các phép tính và hàm kích hoạt (Activation Function) ở các lớp chập.



Hình 6: Convolutional Neural Network (CNN)

5 Thử nghiệm và thảo luận

5.1 Chuẩn bị dữ liệu

Như đã đề cập ở mục 4.1, chúng tôi đã tiến hành tiền xử lý dữ liệu chuẩn bị cho quá trình huấn luyện các mô hình kiến trúc mạng. Bộ dữ liệu sau đó sẽ được chia lại thành hai tập là tập huấn luyện và tập kiểm thử với kích cỡ như bộ dữ liệu ban đầu.

5.2 Cài đặt và tinh chỉnh tham số mô hình

Logistics Regression: Chúng tôi đã sử dụng tham số đầu vào solver='lbfgs', random state=0 .

KNN: trong quá trình huấn luyện model K-Nearest Neighbors chúng tôi đã khảo sát các tham số là n-neighbors và sử dụng n-neighbors=1 .

Random Forest Classifiication: chúng tôi đã sử dụng tham số đầu vào là n-estimators=100 .

Convolutional Neural Network (CNN): Chúng tôi sử dụng CNN với 4 Conv2D Layer với activation = relu, 4 MaxPool2D Layer, 4 BatchNormalization Layer, 1 Flatten Layer và 2 cấu trúc Dense với activation = relu và softmax. Tổng số tham số chúng tôi sử dụng cho kiến trúc mạng này là 1,898,202 .

5.3 Kết quả thử nghiệm

5.3.1 Các độ đo Độ đo Recall: chỉ số này còn được gọi là độ bao phủ tức là xem xét xem mô hình tìm được có khả năng tổng quát hóa như thế nào. Recall càng cao, tức là số điểm là positive bị bỏ sót càng ít. Recall = 1, tức là tất cả số điểm có nhãn là Positive đều được mô hình nhận ra. Recall=$\frac{TP}{TP+FN}$.

Precision (độ chính xác) là tỉ lệ thực sự positive trên tổng số các trường hợp được mô hình dán nhãn "Positive". Precision càng cao, tức là số điểm mô hình dự đoán là positive đều là positive càng nhiều. Precision = 1, tức là tất cả số điểm mô hình dự doán là Positiveđều đúng, hay không có điểm nào có nhãn là Negative mà mô hình dự đoánnhầm là Positive. Precision= $\frac{TP}{TP+FP}$.

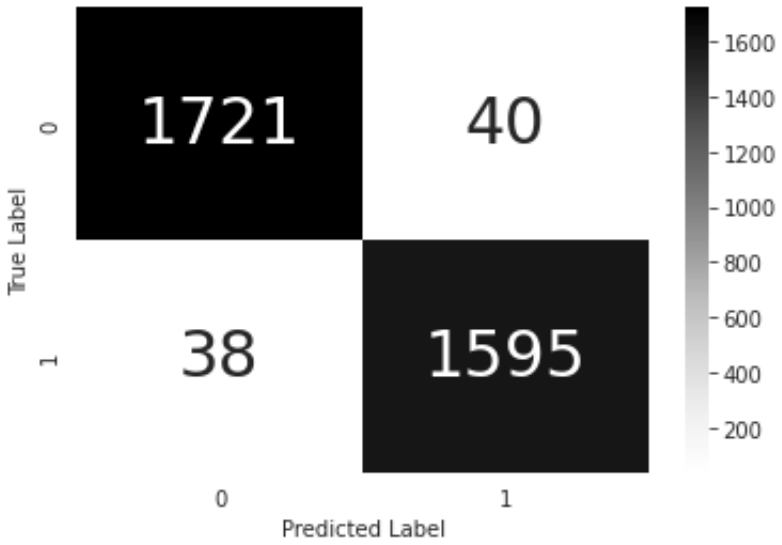
Độ đo Accuracy: Cách đơn giản và hay được sử dụng nhất là accuracy (độ chính xác). Cách đánh giá này đơn giản tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử. Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$.

Độ đo F1: tuy nhiên, chỉ có Precision hay chỉ có Recall thì không đánh giá được chất lượngmô hình. Chỉ dùng Precision, mô hình chỉ đưa ra dự đoán cho một điểm mà nóchắc chắn nhất. Khi đó Precision = 1, tuy nhiên ta không thể nói là mô hình nàytốt. Chỉ dùng Recall, nếu mô hình dự đoán tất cả các điểm đều là positive. Khi đó Recall = 1, tuy nhiên ta cũng không thể nói đây là mô hình tốt.

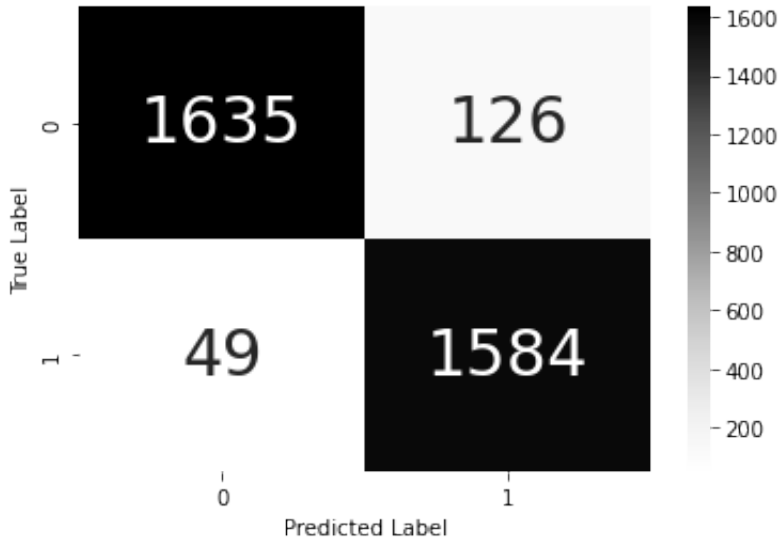
Khi đó F1-score được sử dụng. F1-score là trung bình điều hòa (harmonicmean) của precision và recall (giả sử hai đại lượng này khác 0). F1-score được tính theo công thức: F1=$\frac{2*(Precision*Recall)}{Precision+Recall}$.

Kết quả

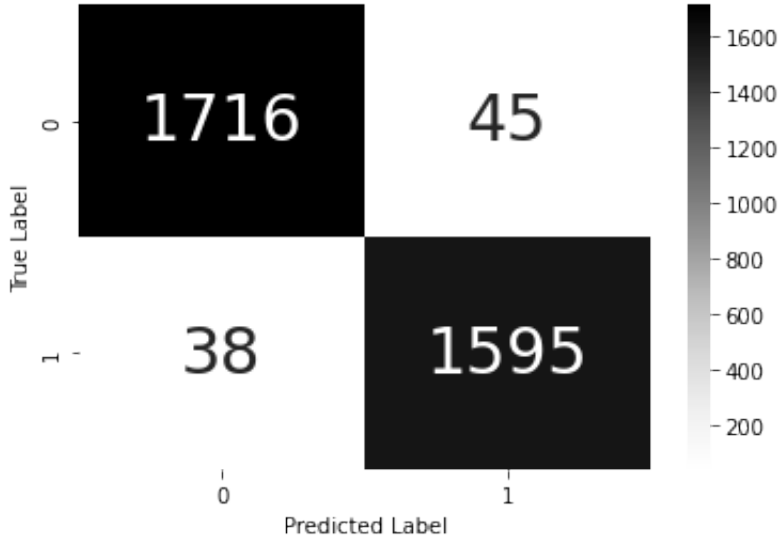
Hình 7: Logistics Regression



Hình 8: KNN



Hình 9: RFC



Hình 10: CNN

Độ đo	Logistic Regression	KNN	RFC	CNN
Precision	0.86	0.98	0.95	0.98
Recall	0.86	0.98	0.95	0.98
F1-Score	0.86	0.98	0.95	0.98
Accuracy	0.86	0.98	0.95	0.98

Bảng 2: Kết quả đánh giá các mô hình trên bộ dữ liệu Sar-Cov-2 CT Scan.

Đối với các bài toán phân loại nhãn và đặc biệt là có sự mất cân bằng về các lớp trong bộ dữ liệu thì Recall, F1 - score là những độ đo hiệu quả mang đến sự chính xác trong việc đánh giá hiệu suất của mô hình. Các kết quả này càng cao, càng gần giá trị 1 thì mô hình phân loại càng tốt.

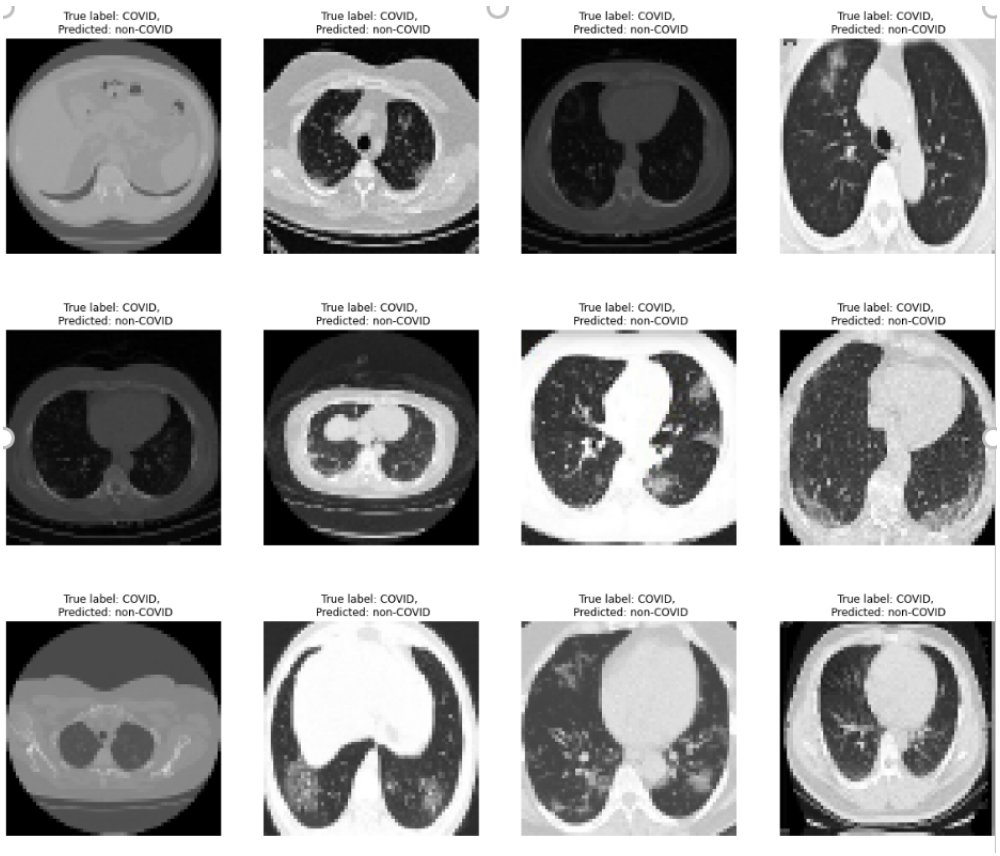
	Logistic Regression		KNN		RFC		CNN	
	Recall	F1-Score	Recall	F1-Score	Recall	F1-Score	Recall	F1-Score
Covid	0.87	0.87	0.98	0.98	0.93	0.95	0.97	0.98
Non-Covid	0.86	0.86	0.98	0.98	0.97	0.95	0.98	0.97

Bảng 3: Kết quả phân loại trên từng nhãn của các mô hình.

Từ các kết quả thử nghiệm với kết quả phân loại trên từng nhãn được ghi lại trên Bảng trên, chúng ta có thể nhận thấy 4 mô hình đều phân loại khá tốt cho cả 2 nhãn "Covid" và "non-Covid". Do bộ dữ liệu có sự cân bằng khá tốt giữa các nhãn (gần như tỉ lệ 1:1) nên việc các mô hình đạt được độ hiệu quả trong việc phân loại 2 nhãn cũng là 1 điều dễ hiểu. Theo kết quả thử nghiệm ở trên, độ chính xác của 3 mô hình CNN, KNN, RFC ($> 94\%$) đều hiệu quả cao hơn hẳn so với LR. Trong đó, đặc biệt độ chính xác của KNN, CNN cao gần như là tuyệt đối của một mô hình lý tưởng. Từ bảng kết quả trên cho thấy rằng, KNN là một lựa chọn vô cùng hiệu quả cho bài toán này.

5.3.2 Phân tích lỗi

Chúng ta nhận thấy rằng có một số ảnh bị dự đoán sai là do ảnh bị mờ ảnh, biến dạng, bị thang màu xám phủ lên làm cho mô hình không tìm được đặc trưng và xảy ra sai sót.
Ngoài ra còn có độ chênh lệch đáng kể giữa chiều rộng và chiều cao, sau khi resize về cùng một kích thước cũng ảnh hưởng đáng kể đến những tính chất ban đầu của ảnh.



Hình 11: Phân tích lỗi trên mô hình tốt nhất

5.3.3 So sánh với các công trình trước đó Đối với dataset của công trình trước đó, số lượng ảnh nhãn COVID và non-COVID khá tương đồng. Điều đó cũng xảy ra ở bộ data của chúng tôi.
Bộ data chúng tôi sử dụng cũng bao gồm SARS-COV-2 CT-Scan dataset nên sẽ có khá nhiều điểm tương đồng về đặc trưng của ảnh do kế thừa lại từ bộ data ban đầu.

Dataset	Image Source	Images	COVID	non-COVID
SARS-COV-2 Ct-Scan	Kaggle	2481	1252	1229
Our dataset	Kaggle	16967	8845	8122

Bảng 4: Sự tương quan của bộ dataset của công trình có liên quan trước đó và bộ dataset mà chúng tôi sử dụng.

Model	F1-score
COVID-19 Detection(Hammam Alshazly)	95.7%
SARS-CoV-2 (Eduardo Soares)	97.31%
Our model(KNN,CNN)	98.0%

Bảng 5: So sánh với nghiên cứu trước đây trên tập dữ liệu SARS-COV-2 CT-Scan Dataset.

6 Kết luận và hướng phát triển

6.1 Kết quả đạt được

Tìm hiểu và tiếp cận được các phương pháp phân loại ảnh cho bài toán object detection cụ thể ở đây là Covid 19 Detection.
Biết ứng dụng các thuật toán máy học trong việc phân loại và đánh giá dữ liệu.
Tìm hiểu được phương pháp Convolutional neural network.
Kết quả của phương pháp học máy là vô cùng khả quan.Với mô hình tốt nhất là KNN với các độ đo đánh giá mô hình Recall, F1-Score lần lượt là 0.98, 0.98.

6.2 Khó khăn gặp phải

Sử dụng các phương pháp tiền xử lý ảnh như xoay, lật ảnh, tăng độ tương phản. Nhưng không đạt được kết quả như mong đợi.
Hạn chế về thời gian nên chưa tìm hiểu nhiều phương pháp Deep Learning.

6.3 Hướng phát triển

Tiếp tục nghiên cứu và tìm hiểu về phương pháp Deep Learning.
Có thể ứng dụng được mô hình này trong tương lai, nhưng chúng ta cần bổ sung thêm số lượng dữ liệu, phong phú và tập dữ liệu phải được cập nhật thường xuyên, kịp thời. Điều đó sẽ giúp cải thiện độ chính xác với mô hình và khi vào thực tế thì mô hình sẽ dễ dàng thích nghi và đưa ra kết quả tốt.

Tài liệu

1. Ozturk Tulin, Talo Muhammed, Yildirimc Eylul Azra, Baloglu Ulas B, Yildirim Ozal, Acharya U.Rajendra. Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-Ray Images - Computers in Biology and Medicine. 2020 - Vol. 121.
2. COVID-NetTeam-COVID-Net Open Source Initiative
   github.com/lindawangg/COVID-Net
3. Dandi Yang, Cristhian Martinez, Lara Visuña, Hardev Khandhar, Chintan Bhatt, Jesus Carretero. Sohrabi, Detection and analysis of COVID-19 in medical images using deep learning techniques - World health organization declares global emergency: A review of the 2019 novel coronavirus (2020), [71–76].
4. Dữ liệu được lấy từ : kaggle.com/maedemaftouni/large-covid19-ct-slice-dataset và kaggle.com/plameneduardo/sarscov2-ctscan-dataset
5. https://pubmed.ncbi.nlm.nih.gov/34401477/
6. Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. medRxiv, 2020.
7. Machine Learning cơ bản
   machinelearningcoban.com
8. Random Forest algorithm
   machinelearningcoban/random-forest
9. Làm quen với Keras
   viblo.asia/lam-quen-voi-keras
10. Tìm hiểu về thư viện Keras trong Deep Learning | Thor Pham Blog
    https://thorpham.github.io/keras