**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

**VIỆN KỸ THUẬT – CÔNG NGHỆ**

**\*\*\*\*\*\*\*\*\*\***



**BÁO CÁO TIỂU LUẬN**

**SỬ DỤNG TRÍ TUỆ NHÂN TẠO TRONG VIỆC PHÁT HIỆN VÀ DỰ ĐOÁN BỆNH VIÊM PHỔI QUA HÌNH ẢNH X–RAY**

**Sinh viên thực hiện : Đặng Trọng Đại**

**MSSV : 2024802010398**

**Lớp : D20CNTT04**

**Ngành : Công Nghệ Thông Tin**

**Giảng viên hướng dẫn : ThS. Nguyễn Hải Vĩnh Cường**

*Bình Dương, Tháng 3/2023*

**TÓM TẮT**

Đề tài xây dựng trí tuệ nhân tạo dự đoán bệnh viêm phổi qua hình ảnh X-ray

Mục tiêu: Đề tài này nhằm xây dựng một mô hình AI (trí tuệ nhân tạo) để dự đoán và phát hiện bệnh viêm phổi thông qua việc phân tích hình ảnh X-ray của bệnh nhân.

Ý tưởng chính: Sử dụng các kỹ thuật học sâu (deep learning), đặc biệt là các mạng nơ-ron tích chập (CNN - Convolutional Neural Networks) để đạt được khả năng phân loại hình ảnh X-ray có bệnh viêm phổi hay không.

Các bước thực hiện:

1. Thu thập dữ liệu: Tập hợp các hình ảnh X-ray của bệnh nhân, bao gồm cả những người bị viêm phổi và những người không bị.
2. Tiền xử lý dữ liệu: Làm sạch, định dạng và chia dữ liệu thành tập huấn luyện, tập xác thực và tập kiểm tra.
3. Xây dựng mô hình: Thiết kế mạng nơ-ron tích chập (CNN) để học từ dữ liệu và nhận diện các đặc trưng quan trọng trong hình ảnh X-ray.
4. Huấn luyện mô hình: Sử dụng tập huấn luyện để cập nhật trọng số của mô hình, đồng thời kiểm soát quá trình học để tránh hiện tượng quá khớp (overfitting).
5. Đánh giá mô hình: Sử dụng tập kiểm tra để đánh giá hiệu quả của mô hình AI trong việc phát hiện bệnh viêm phổi từ hình ảnh X-ray.
6. Tối ưu hóa và điều chỉnh mô hình: Thực hiện các thay đổi và tinh chỉnh để cải thiện hiệu suất của mô hình trên dữ liệu thực tế.
7. Triển khai mô hình: Áp dụng mô hình AI vào thực tế, giúp bác sĩ và nhân viên y tế dễ dàng phát hiện bệnh viêm phổi thông qua hình ảnh X-ray.

Kết quả mong đợi: Mô hình AI dự đoán bệnh viêm phổi qua hình ảnh X-ray sẽ giúp nâng cao khả năng phát hiện và chẩn đoán bệnh viêm phổi, đồng thời giảm thiểu thời gian và nguồn lực y tế trong việc xử lý các ca bệnh.

**SUMMARY**

Topic: Developing an AI model for predicting pneumonia through X-ray images

Objective: This project aims to build an AI (Artificial Intelligence) model to predict and detect pneumonia through analyzing X-ray images of patients.

Main idea: Using deep learning techniques, especially Convolutional Neural Networks (CNN), to achieve the ability to classify X-ray images with or without pneumonia.

Implementation steps:

1. Data collection: Collect X-ray images of patients, including both those with and without pneumonia.
2. Data preprocessing: Clean, format, and divide data into training, validation, and testing sets.
3. Model building: Design a Convolutional Neural Network (CNN) to learn from data and recognize important features in X-ray images.
4. Model training: Use the training set to update the model's weights while controlling the learning process to avoid overfitting.
5. Model evaluation: Use the testing set to evaluate the effectiveness of the AI model in detecting pneumonia from X-ray images.
6. Model optimization and adjustment: Make changes and adjustments to improve the model's performance on real-world data.
7. Model deployment: Apply the AI model in practice, helping doctors and healthcare staff easily detect pneumonia through X-ray images.

Expected results: The AI model for predicting pneumonia through X-ray images is expected to improve the ability to detect and diagnose pneumonia, as well as reduce time and healthcare resources in handling pneumonia cases.

**MỤC LỤC**

[LỜI MỞ ĐẦU 1](#_Toc130985770)

[CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI 2](#_Toc130985771)

[1.1. Giới thiệu về đề tài 2](#_Toc130985772)

[1.2. Lý do chọn đề tài 3](#_Toc130985773)

[1.3. Mục tiêu nghiên cứu 3](#_Toc130985774)

[1.4. Đối tượng nghiên cứu 3](#_Toc130985775)

[1.5. Phạm vi nghiên cứu 3](#_Toc130985776)

[1.6. Cơ sở lý thuyết 3](#_Toc130985777)

[1.6.1. Ngôn ngữ python 3](#_Toc130985778)

[1.6.2. Các thư viện sử dụng 4](#_Toc130985779)

[1.6.3. Các công nghệ sử dụng 5](#_Toc130985780)

[CHƯƠNG 2. TIỀN XỬ LÝ DỮ LIỆU 6](#_Toc130985781)

[2.1. Thu thập dữ liệu 6](#_Toc130985782)

[2.2. Mô tả dữ liệu 6](#_Toc130985783)

[2.3. Trực quan hóa dữ liệu 7](#_Toc130985784)

[2.4. Tăng cường dữ liệu 8](#_Toc130985785)

[CHƯƠNG 3. THIẾT KẾ MÔ HÌNH 9](#_Toc130985786)

[3.1. Giới thiệu mô hình 9](#_Toc130985787)

[3.2. Giai đoạn huấn luyện (training) 10](#_Toc130985788)

[3.3. Giai đoạn đánh giá (testing) 11](#_Toc130985789)

[CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM 13](#_Toc130985790)

[4.1. Kết quả thực nghiệm 13](#_Toc130985791)

[4.2. Đánh giá mô hình 13](#_Toc130985792)

[4.2.1. Đánh giá mô hình bằng biểu đồ huấn luyện và đánh giá 13](#_Toc130985793)

[4.2.2. Đánh giá mô hình qua biểu đồ nhiệt 14](#_Toc130985794)

[4.2.3. Đánh giá mô hình thông qua Confusion Matrix 14](#_Toc130985795)

[4.2.4. Đánh giá mô hình thông qua classification report 16](#_Toc130985796)

[CHƯƠNG 5. KẾT LUẬN 17](#_Toc130985797)

[5.1. Kết quả đạt được 17](#_Toc130985798)

[5.2. Hạn chế 17](#_Toc130985799)

[5.3. Kết luận 17](#_Toc130985800)

[TÀI LIỆU THAM KHẢO 18](#_Toc130985801)

**DANH MỤC BẢNG**

[Bảng 4.1. Bảng đánh giá mô hình thông qua confusion matrix 15](#_Toc130984730)

[Bảng 4.2. Bảng đánh giá mô hình thông qua classification report 16](#_Toc130984731)

**DANH MỤC HÌNH ẢNH**

[Hình 1. Bệnh viêm phổi 2](#_Toc130984671)

[Hình 2. Xem trước hình ảnh của mỗi lớp 7](#_Toc130984672)

[Hình 3. Thống kê về tập dữ liệu 7](#_Toc130984673)

[Hình 4. Đánh giá mô hình thông qua biểu đồ training and validation plot 13](#_Toc130984674)

[Hình 5. Đánh giá mô hình thông qua đồ thị heatmap 14](#_Toc130984675)

**DANH MỤC TỪ VIẾT TẮT**

|  |  |
| --- | --- |
| **Từ viết tắt** | **Giải thích** |
| AI | Artificial Intelligence (Trí tuệ nhân tạo) |
| CNN | Convolutional Neural Network |
| Pneumonia | Bệnh Viêm Phổi |
| Normal | Người bình thường |
| Training | Huấn luyện |
| Testing | Đánh giá |
| Loss | Sự mất mát (dữ liệu) |
| Accuracy | Độ Chính Xác |

# LỜI MỞ ĐẦU

Bệnh viêm phổi (Pneumonia) là một trong những bệnh lý nguy hiểm nhất đối với con người. Theo thống kê của Tổ chức Y tế Thế giới, mỗi năm có khoảng 8 triệu người mắc bệnh viêm phổi và gần 3 triệu người tử vong vì bệnh này. Do đó, việc phát hiện và dự đoán bệnh viêm phổi là một vấn đề được quan tâm đặc biệt trong lĩnh vực y tế.

Bệnh viêm phổi được xác định là một loại bệnh lý về đường hô hấp, tác động trực tiếp lên phổi, gây ra tình trạng viêm nhiễm ở phổi và những khối u khí độc. Bệnh thường gây nên các triệu chứng như sốt cao, ho, khó thở, đau ngực, mệt mỏi,… Nếu không được phát hiện và điều trị đúng cách sẽ dẫn đến những biến chứng nguy hiểm, thậm chí có thể gây tử vong.

Khi bệnh viêm phổi được phát hiện sớm, tỷ lệ điều trị thành công sẽ cao và giúp đưa ra quyết định điều trị chính xác hơn. Việc ứng dụng các công nghệ tiên tiến trong việc phát hiện và dự đoán bệnh viêm phổi cũng đóng vai trò rất quan trọng trong việc giúp cho bệnh nhân được điều trị kịp thời.

Trong những năm gần đây, Trí tuệ nhân tạo (AI) đã phát triển rất nhanh chóng và được ứng dụng rộng rãi trong các lĩnh vực khác nhau, trong đó có y tế. Sử dụng AI trong việc phát hiện và dự đoán bệnh viêm phổi là một trong những nghiên cứu mới nhất và đầy triển vọng.

Do đó, em chọn đề tài phát hiện và chuẩn đoán bệnh viêm phổi để áp dụng công nghệ AI vào việc tìm kiếm và chuẩn đoán bệnh viêm phổi từ hình ảnh X quang.

Đề tài gồm các phần sau:

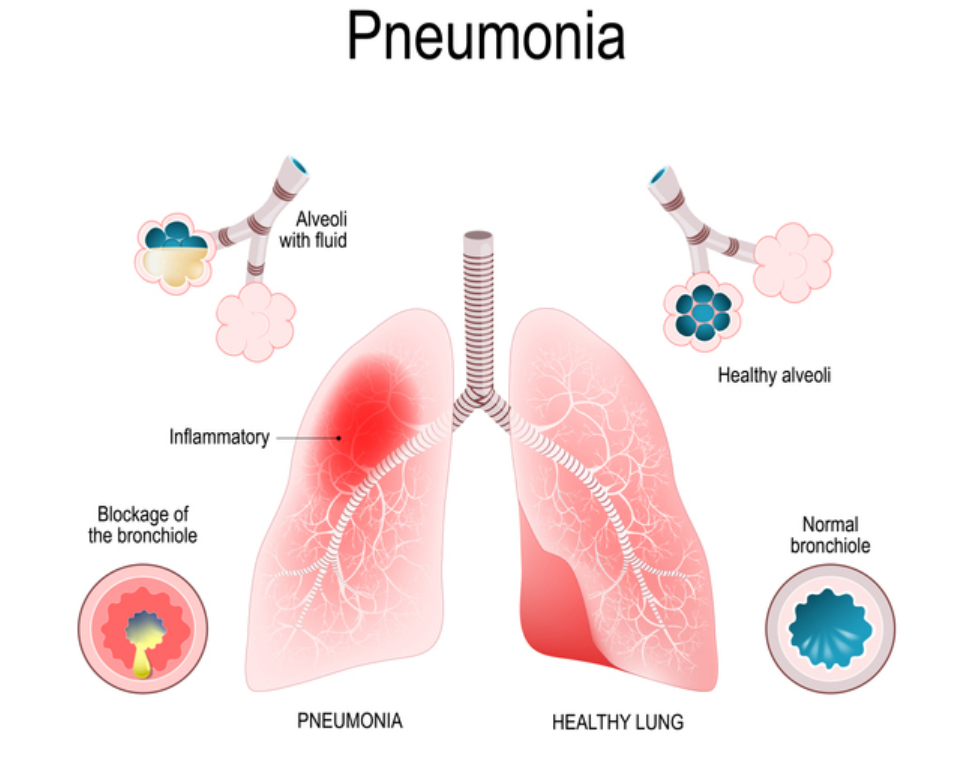
* Tổng quan về đề tài
* Tiền xử lý dữ liệu
* Thiết kế mô hình
* Đánh giá thực nghiệm
* Kết luận

# TỔNG QUAN VỀ ĐỀ TÀI

## Giới thiệu về đề tài

Tên đề tài: “Pneumonia Detection” – Ứng dụng nhận biết bệnh viêm phổi

Viêm phổi là một tình trạng viêm nhiễm của phổi ảnh hưởng chủ yếu đến các túi khí nhỏ được gọi là phế nang. Các triệu chứng thường bao gồm một số kết hợp ho có đờm hoặc ho khan, đau ngực, sốt và khó thở. Mức độ nghiêm trọng của tình trạng này có thể thay đổi. Viêm phổi thường do nhiễm vi-rút hoặc vi khuẩn và ít phổ biến hơn do các vi sinh vật khác, một số loại thuốc hoặc tình trạng bệnh như bệnh tự miễn dịch. Các yếu tố nguy cơ bao gồm xơ nang, bệnh phổi tắc nghẽn mãn tính (COPD), hen suyễn, tiểu đường, suy tim, tiền sử bệnh hút thuốc, khả năng ho kém như sau đột quỵ và hệ thống miễn dịch yếu. Chẩn đoán thường dựa trên các triệu chứng và khám thực thể. Chụp X-quang ngực, xét nghiệm máu và nuôi cấy đờm có thể giúp xác định chẩn đoán. Bệnh có thể được phân loại theo nơi mắc phải, chẳng hạn như mắc tại cộng đồng hoặc bệnh viện hoặc liên quan đến chăm sóc sức khỏe bệnh viêm phổi.



Hình 1. Bệnh viêm phổi

## Lý do chọn đề tài

Bệnh viêm phổi là một trong những bệnh lý nguy hiểm và phổ biến trên toàn cầu, đặc biệt là trong bối cảnh đại dịch COVID-19 đang diễn ra. Việc phát hiện và chẩn đoán bệnh viêm phổi sớm là rất quan trọng để cung cấp cho bệnh nhân những liệu pháp điều trị kịp thời, từ đó giảm tỷ lệ tử vong và tăng khả năng phục hồi của bệnh nhân. Do đó, nghiên cứu và phát triển một hệ thống trí tuệ nhân tạo để phát hiện bệnh viêm phổi sớm có thể mang lại nhiều lợi ích cho ngành y tế và cộng đồng.

## Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu là phát triển một hệ thống trí tuệ nhân tạo để phát hiện bệnh viêm phổi từ hình ảnh chụp X-quang của bệnh nhân. Mô hình sẽ được huấn luyện trên tập dữ liệu lớn chứa các hình ảnh của người bị bệnh viêm phổi và những người khỏe mạnh. Mục tiêu là tạo ra một hệ thống có độ chính xác cao, nhận dạng chính xác bệnh viêm phổi và hạn chế số lượng dương tính giả.

## Đối tượng nghiên cứu

Đối tượng của nghiên cứu là các bệnh nhân bị viêm phổi và những người khỏe mạnh. Dữ liệu hình ảnh được sử dụng để huấn luyện và kiểm tra mô hình sẽ được thu thập từ các bệnh viện và các cơ sở y tế khác nhau.

## Phạm vi nghiên cứu

Nghiên cứu sẽ tập trung vào phát triển một hệ thống trí tuệ nhân tạo để phát hiện bệnh viêm phổi từ hình ảnh chụp X-quang. Mô hình sẽ được huấn luyện trên tập dữ liệu lớn và được kiểm tra trên các tập dữ liệu kiểm tra khác nhau để đánh giá độ chính xác của nó. Nghiên cứu sẽ tập trung vào các kỹ thuật học máy và xử lý hình ảnh để tạo ra một hệ thống phát hiện bệnh viêm phổi có hiệu quả và đáng tin cậy.

## Cơ sở lý thuyết

### Ngôn ngữ python

Python là một ngôn ngữ lập trình đa mẫu hình, nó hỗ trợ hoàn toàn mẫu lập trình hướng đối tượng và lập trình cấu trúc; ngoài ra về mặt tính năng, Python cũng hỗ trợ lập trình hàm và lập trình hướng khía cạnh. Nhờ vậy mà Python có thể làm được rất nhiều thứ, sử dụng trong nhiều lĩnh vực khác nhau.

Python là ngôn ngữ được ứng dụng đa dạng trong các lĩnh vực:

* Làm Website
* Tool tự động hóa
* Khoa học máy tính
* Lĩnh vực IoT
* Làm game

### Các thư viện sử dụng

* OS: là thư viện chuẩn cung cấp các phương thức để tương tác với hệ thống file và thư mục của máy tính, là một công cụ hữu ích cho các tác vụ liên quan đến xử lý file và thư mục trên máy tính.
* Matplotlib: thư viện giúp tạo đồ thị, vẽ biểu đồ, hiển thị dữ liệu dưới dạng đồ họa (2D và 3D) cho các ứng dụng khoa học máy tính, là một công cụ rất hữu ích cho các ứng dụng khoa học và kỹ thuật.
* Seaborn: là một thư viện được sử dụng để tạo ra các đồ thị và biểu đồ hấp dẫn và chuyên nghiệp. Nó cung cấp một số tính năng và chức năng để tùy chỉnh các biểu đồ để phù hợp với nhu cầu của người dùng.
* Tensorflow: là một thư viện mã nguồn mở của Google, được sử dụng rộng rãi trong lĩnh vực trí tuệ nhân tạo và học sâu. Thư viện này cung cấp một loạt các công cụ để xây dựng, đào tạo và triển khai các mô hình học máy và mạng nơ-ron sâu.
* Keras: là một thư viện mã nguồn mở được sử dụng để xây dựng, đào tạo và triển khai các mô hình mạng nơ-ron sâu. Nó cung cấp một API đơn giản và dễ sử dụng cho việc xây dựng các mô hình mạng nơ-ron sâu, bao gồm cả mạng nơ-ron tích chập (CNN), mạng nơ-ron truyền thẳng (feedforward neural network) và mạng nơ-ron tái cấu trúc (recurrent neural network).
* Cv2: là một thư viện của Python được sử dụng trong xử lý ảnh và thị giác máy tính. Thư viện này là một giao diện cho thư viện **OpenCV** - một thư viện mã nguồn mở có chức năng tương tự như **cv2**. **cv2** cung cấp các chức năng xử lý ảnh như lọc ảnh, biến đổi màu sắc, phát hiện đối tượng, xử lý video và các chức năng khác.
* Pandas: thư viện giúp đọc, ghi, xử lý dữ liệu cho các tác vụ phân tích dữ liệu và machine learning.
* NumPy: thư viện giúp thực hiện các phép toán đại số tuyến tính.
* Scikit-learn (Sklearn) là thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling gồm: classification, regression, clustering, và dimensionality reduction.

### Các công nghệ sử dụng

* Sử dụng các mô hình mạng nơ-ron sâu như Convolutional Neural Network (CNN) để huấn luyện mô hình phát hiện bệnh viêm phổi từ hình ảnh X-quang.
* Computer vision: sử dụng các kỹ thuật xử lý hình ảnh như phân đoạn hình ảnh (image segmentation), phát hiện cạnh (edge detection) và đặc trưng hóa hình ảnh (feature extraction) để trích xuất thông tin từ hình ảnh X-quang.
* Transfer learning: sử dụng các mô hình đã được huấn luyện trước đó trên tập dữ liệu lớn, như VGG16, ResNet50 và InceptionV3, để giúp tăng tốc quá trình huấn luyện mô hình và cải thiện độ chính xác của mô hình.

# TIỀN XỬ LÝ DỮ LIỆU

## Thu thập dữ liệu

Để phát triển một hệ thống trí tuệ nhân tạo để phát hiện bệnh viêm phổi từ hình ảnh chụp X-quang, ta cần thu thập các dữ liệu hình ảnh của bệnh nhân bị viêm phổi và những người khỏe mạnh để huấn luyện và kiểm tra mô hình. Dữ liệu cần được thu thập từ các nguồn đáng tin cậy và có độ đại diện cao cho các trường hợp bệnh viêm phổi và không bị bệnh viêm phổi.

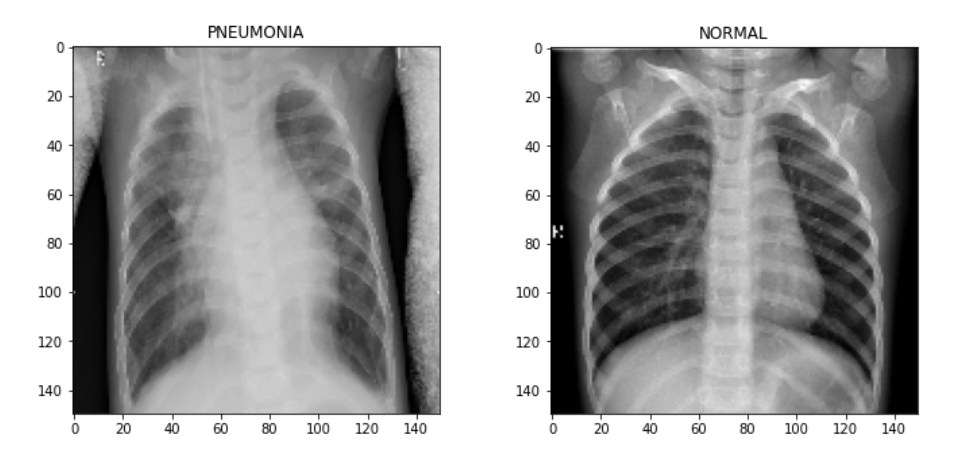
Dữ liệu cần thu thập bao gồm: Hình ảnh chụp X-quang của bệnh nhân và người bình thường: đây là phương pháp khảo sát phổ biến để chẩn đoán bệnh viêm phổi. Các hình ảnh này cần được thu thập từ bệnh viện hoặc các cơ sở y tế có đầy đủ trang thiết bị và kỹ thuật để chụp X-quang.

Sau khi thu thập được dữ liệu, cần kiểm tra tính đúng đắn và sự đại diện của các mẫu dữ liệu để đảm bảo tính đáng tin cậy của kết quả huấn luyện và kiểm tra của mô hình.

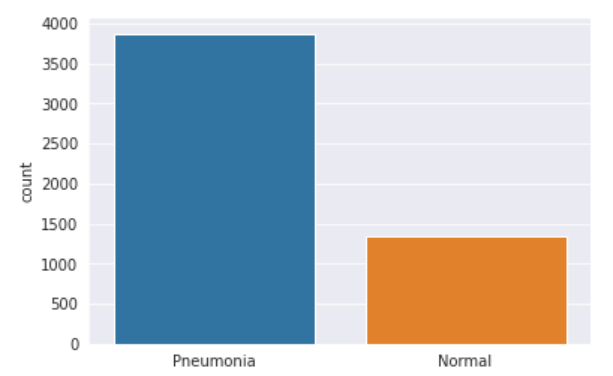
## Mô tả dữ liệu

Bộ dữ liệu này được tổ chức thành 3 thư mục (train, test, val) và chứa các thư mục con cho mỗi danh mục hình ảnh (Pneumonia/Normal). Có tổng cộng 5.863 hình ảnh X-Quang (JPEG) và 2 danh mục (Pneumonia/Normal). Hình ảnh X-Quang phổi (trước-sau) được lựa chọn từ các nhóm bệnh nhân nhi khoa từ 1 đến 5 tuổi từ Trung tâm Y tế Phụ nữ và Trẻ em Quảng Châu, tỉnh Quảng Châu. Tất cả các hình ảnh X-Quang phổi được thực hiện trong quá trình chăm sóc lâm sàng thông thường cho bệnh nhân. Để phân tích hình ảnh X-Quang phổi, tất cả các bức ảnh chụp phổi đều được sàng lọc ban đầu để kiểm soát chất lượng bằng cách loại bỏ tất cả các quét quá thấp hoặc không đọc được. Sau đó, chẩn đoán cho các hình ảnh được đánh giá bởi hai bác sĩ chuyên gia trước khi được lựa chọn để huấn luyện hệ thống trí tuệ nhân tạo. Để đảm bảo tính chính xác của chẩn đoán, tập dữ liệu đánh giá cũng được kiểm tra bởi một bác sĩ chuyên gia thứ ba.

## Trực quan hóa dữ liệu



Hình 2. Xem trước hình ảnh của mỗi lớp



Hình 3. Thống kê về tập dữ liệu

Hình ảnh trên mô tả tập dữ liệu của bộ training. Chúng ta thấy rằng ở đây tập dữ liệu không đồng đều dẫn tới tình trạng mất cân bằng dữ liệu. Để giảm bớt tình trạng mất cân bằng này, chúng ta cần có biện pháp sử dụng tăng cường dữ liệu.

## Tăng cường dữ liệu

Để tránh vấn đề overfitting, chúng ta cần mở rộng bộ dữ liệu một cách nhân tạo. Chúng ta có thể làm cho bộ dữ liệu hiện tại trở nên lớn hơn. Ý tưởng là sửa đổi dữ liệu huấn luyện với các biến đổi nhỏ để tái tạo các biến thể. Các phương pháp sửa đổi dữ liệu huấn luyện theo cách thay đổi biểu diễn mảng trong khi giữ nguyên nhãn được gọi là các kỹ thuật tăng cường dữ liệu. Một số kỹ thuật tăng cường phổ biến mà mọi người sử dụng là grayscale, lật ngang, lật dọc, cắt ngẫu nhiên, độ rung màu sắc, dịch chuyển, xoay và nhiều hơn nữa. Bằng cách áp dụng chỉ một vài biến đổi này vào dữ liệu huấn luyện của chúng ta, chúng ta có thể dễ dàng làm tăng gấp đôi hoặc gấp ba số lượng các ví dụ huấn luyện và tạo ra một mô hình rất mạnh mẽ.

Để tăng cường dữ liệu, chúng ta cần phải thực hiện một vài thứ trên tập dữ liệu:

1. Ngẫu nhiên xoay một số hình ảnh huấn luyện 30 độ.

2. Ngẫu nhiên phóng to 20% một số hình ảnh huấn luyện.

3. Ngẫu nhiên dịch chuyển hình ảnh sang ngang 10% chiều rộng.

4. Ngẫu nhiên dịch chuyển hình ảnh lên xuống 10% chiều cao.

5. Ngẫu nhiên lật hình ảnh theo chiều ngang.

Sau khi mô hình đã sẵn sàng, chúng sẽ lấp được bộ dữ liệu huấn luyện.

# THIẾT KẾ MÔ HÌNH

## Giới thiệu mô hình

Mạng nơron tích chập (còn gọi là ConvNet / CNN) là một thuật toán Deep Learning có thể lấy hình ảnh đầu vào, gán độ quan trọng (các trọng số - weights và độ lệch - bias có thể học được) cho các đặc trưng/đối tượng khác nhau trong hình ảnh và có thể phân biệt được từng đặc trưng/đối tượng này với nhau. Công việc tiền xử lý được yêu cầu cho mạng nơron tích chập thì ít hơn nhiều so với các thuật toán phân loại khác. Trong các phương thức sơ khai, các bộ lọc được thiết kế bằng tay (hand - engineered), với một quá trình huấn luyện để chọn ra các bộ lọc/đặc trưng phù hợp thì mạng nơron tích chập lại có khả năng tự học để chọn ra các bộ lọc/ đặc trưng tối ưu nhất.

Kiến trúc của nơron tích chập tương tự như mô hình kết nối của các nơron trong bộ não con người và được lấy cảm hứng từ hệ thống vỏ thị giác trong bộ não (visual cortex). Các nơ-ron riêng lẻ chỉ phản ứng với các kích thích trong một khu vực hạn chế của trường thị giác được gọi là Trường tiếp nhận (Receptive Field). Một tập hợp các trường như vậy chồng lên nhau để bao phủ toàn bộ khu vực thị giác..

CNN là một trong những kiến trúc mạng học sâu được sử dụng phổ biến nhất trong các ứng dụng xử lý hình ảnh, bao gồm cả việc phát hiện bệnh trong hình ảnh y tế.

Trong việc phát hiện bệnh viêm phổi qua hình ảnh X-quang, chúng ta cần phân loại các hình ảnh là bình thường hay bị nhiễm bệnh. Mô hình CNN có thể học được các đặc trưng từ hình ảnh để phân loại chúng vào các lớp tương ứng. Các lớp này có thể là "normal" và "pneumonia".

Trong quá trình huấn luyện mô hình CNN, chúng ta sử dụng tập dữ liệu X-quang của các bệnh nhân được chẩn đoán là bình thường và bị nhiễm bệnh để huấn luyện mô hình phân loại. Sau đó, chúng ta đánh giá hiệu suất của mô hình trên một tập dữ liệu kiểm tra độc lập.

## Giai đoạn huấn luyện (training)

Giai đoạn huấn luyện của một mô hình máy học là quá trình học từ dữ liệu đầu vào để tối ưu hóa các tham số trong mô hình để có thể đưa ra dự đoán chính xác trên tập dữ liệu huấn luyện. Trong việc sử dụng mô hình CNN để phát hiện bệnh viêm phổi qua hình ảnh X-quang, giai đoạn huấn luyện sẽ bao gồm các bước sau:

1. Chuẩn bị dữ liệu: Chuẩn bị tập dữ liệu hình ảnh X-quang của các bệnh nhân được chẩn đoán là bình thường và bị nhiễm bệnh để sử dụng trong quá trình huấn luyện. Dữ liệu cần được chuẩn hóa và tiền xử lý để phù hợp với mô hình.

2. Xây dựng mô hình: Xây dựng mô hình CNN và thiết lập các siêu tham số, bao gồm số lượng lớp, kích thước bộ lọc, hàm kích hoạt, tốc độ học và số lượng epoch để huấn luyện.

3. Huấn luyện mô hình: Sử dụng tập dữ liệu huấn luyện để huấn luyện mô hình. Quá trình này sẽ sử dụng các kỹ thuật tối ưu hóa như SGD (Stochastic Gradient Descent), Adam optimizer để điều chỉnh các tham số của mô hình để giảm thiểu hàm mất mát.

4. Đánh giá mô hình: Sau khi huấn luyện xong, ta cần đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra độc lập để đánh giá khả năng tổng quát hóa của mô hình.

5. Tinh chỉnh mô hình: Dựa trên kết quả đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra, ta có thể điều chỉnh các siêu tham số của mô hình để cải thiện hiệu suất của nó. Quá trình này được lặp lại cho đến khi đạt được một mô hình tối ưu cho bài toán.

Khi huấn luyện mô hình, chúng ta cần lưu ý đến vấn đề overfitting - mô hình quá tập trung vào dữ liệu huấn luyện và không thể tổng quát hóa trên dữ liệu mới. Để giảm thiểu tình trạng overfitting, chúng ta có thể sử dụng các kỹ thuật như regularization, dropout, early stopping hoặc thu thập thêm dữ liệu kiểm tra độc lập để đánh giá khả năng tổng quát hóa của mô hình.

Ngoài ra, khi huấn luyện mô hình, ta cũng cần lưu ý đến vấn đề imbalance data - tập dữ liệu huấn luyện có số lượng mẫu bệnh nhân bình thường nhiều hơn số lượng mẫu bệnh nhân bị nhiễm bệnh. Điều này có thể gây ra hiện tượng mô hình không nhận diện được các mẫu bệnh nhân bị nhiễm bệnh. Để giải quyết vấn đề này, ta có thể sử dụng các kỹ thuật như oversampling hoặc undersampling để cân bằng số lượng mẫu giữa các lớp.

Sau khi huấn luyện mô hình, ta có thể sử dụng mô hình này để phân loại các hình ảnh X-quang mới và đưa ra dự đoán về khả năng bệnh nhân bị nhiễm bệnh viêm phổi hay không.

## Giai đoạn đánh giá (testing)

Giai đoạn đánh giá là giai đoạn quan trọng để đánh giá chất lượng của mô hình phát hiện bệnh viêm phổi qua hình ảnh X-quang sử dụng mạng CNN. Để đánh giá mô hình, ta cần sử dụng một tập dữ liệu kiểm tra độc lập, tương tự như giai đoạn huấn luyện. Tập dữ liệu kiểm tra này sẽ được sử dụng để đánh giá khả năng tổng quát hóa của mô hình - mức độ mà mô hình có thể phát hiện được bệnh viêm phổi trên các hình ảnh X-quang mới, chưa từng được sử dụng trong quá trình huấn luyện.

Giai đoạn đánh giá bao gồm các bước sau đây:

1. Đưa tập dữ liệu kiểm tra vào mô hình để phân loại các hình ảnh X-quang thành hai lớp: bệnh nhân bị nhiễm bệnh viêm phổi và bệnh nhân bình thường.

2. Tính toán các chỉ số đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra, bao gồm: độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc hiệu (specificity), và F1-score. Độ chính xác tính tỷ lệ giữa số lượng dự đoán đúng và tổng số lượng mẫu kiểm tra. Độ nhạy (sensitivity) tính tỷ lệ giữa số lượng bệnh nhân bị nhiễm bệnh viêm phổi được phát hiện đúng và tổng số lượng bệnh nhân bị nhiễm bệnh viêm phổi trong tập kiểm tra. Độ đặc hiệu (specificity) tính tỷ lệ giữa số lượng bệnh nhân bình thường được phát hiện đúng và tổng số lượng bệnh nhân bình thường trong tập kiểm tra. F1-score là trung bình điều hòa giữa độ nhạy và độ đặc hiệu của mô hình, tính theo công thức: 2 \* (precision \* recall) / (precision + recall).

3. Đánh giá sự khác biệt giữa kết quả dự đoán của mô hình và kết quả thực tế trên tập dữ liệu kiểm tra, thông qua việc xem xét các trường hợp dự đoán sai và phân tích nguyên nhân của các lỗi này.

4. Nếu kết quả đánh giá không đạt yêu cầu, ta có thể điều chỉnh các thông số của mô hình hoặc thêm dữ liệu mới để cải thiện hiệu suất của mô hình.

Ngoài ra, trong giai đoạn đánh giá, ta cũng có thể sử dụng các phương pháp trực quan hóa để hiểu rõ hơn về cách mà mô hình đưa ra dự đoán. Ví dụ, ta có thể sử dụng ma trận confusion (confusion matrix) để xem xét số lượng dự đoán đúng và sai trong mỗi lớp, hoặc sử dụng các đường cong ROC (ROC curve) và Precision-Recall (PR) curve để đánh giá độ chính xác của mô hình trên các ngưỡng khác nhau của đầu ra dự đoán.

Tóm lại, giai đoạn đánh giá là giai đoạn quan trọng để đánh giá chất lượng của mô hình phát hiện bệnh viêm phổi qua hình ảnh X-quang sử dụng mạng CNN. Để đánh giá mô hình, ta cần sử dụng tập dữ liệu kiểm tra độc lập, tính toán các chỉ số đánh giá hiệu suất và sử dụng các phương pháp trực quan hóa để hiểu rõ hơn về cách mà mô hình đưa ra dự đoán. Nếu kết quả đánh giá không đạt yêu cầu, ta có thể điều chỉnh mô hình hoặc thêm dữ liệu mới để cải thiện hiệu suất của mô hình.

# ĐÁNH GIÁ THỰC NGHIỆM

## Kết quả thực nghiệm

Sau khi chạy chương trình huấn luyện mô hình học máy CNN vào việc phát hiện và dự đoán bệnh viêm phổi thì chương trình đưa ra được độ đánh giá hiệu suất của mô hình trong quá trình training và testing như sau:

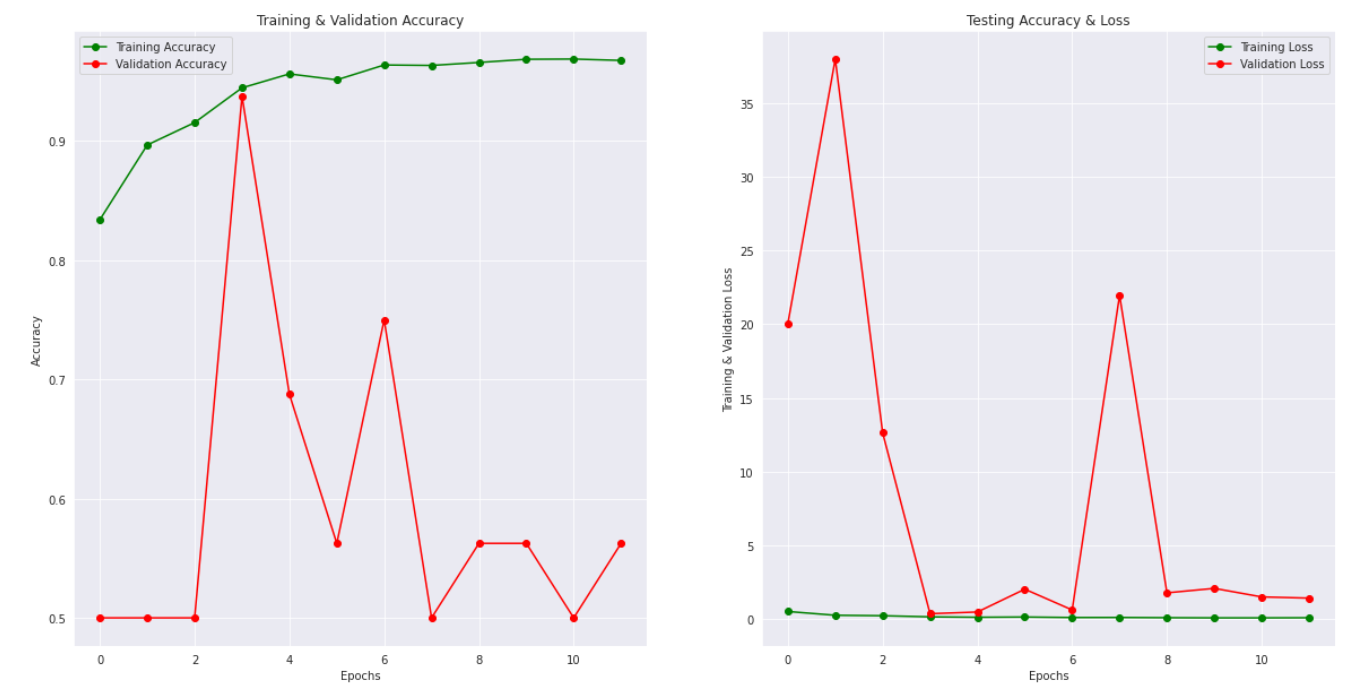
* Loss: 0.2656453164915244
* Accuracy: 91.50640964508057 %

Đánh giá độ hiệu suất của mô hình trên tập dữ liệu được cung cấp cho thấy rằng:

* Loss của mô hình là 0.2656453164915244, có nghĩa là mức độ khác biệt giữa đầu ra dự đoán và đầu ra thực tế trên tập dữ liệu đang được đánh giá khá thấp.
* Accuracy của mô hình là 91.50640964508057%, có nghĩa là mô hình đã dự đoán chính xác lớp của hơn 91% số mẫu trên tập dữ liệu đang được đánh giá.

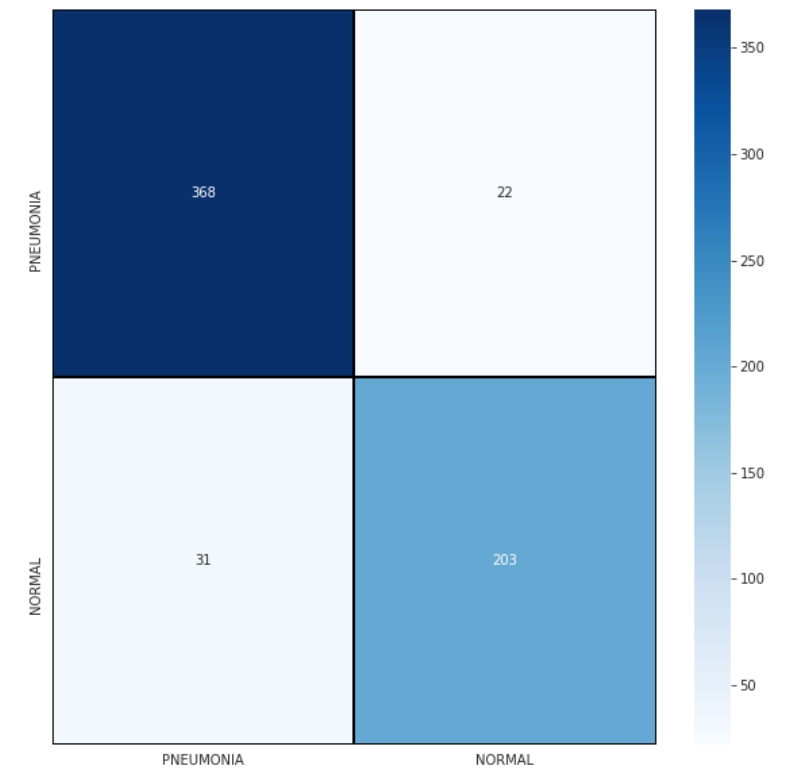
## Đánh giá mô hình

### Đánh giá mô hình bằng biểu đồ huấn luyện và đánh giá



Hình 4. Đánh giá mô hình thông qua biểu đồ training and validation plot

### Đánh giá mô hình qua biểu đồ nhiệt



Hình 5. Đánh giá mô hình thông qua đồ thị heatmap

### Đánh giá mô hình thông qua Confusion Matrix

Bảng 4.1. Bảng đánh giá mô hình thông qua confusion matrix

|  |  |  |
| --- | --- | --- |
|  | Dự đoán không có bệnh viêm phổi | Dự đoán có bệnh viêm phổi |
| Thực tế có bệnh viêm phổi | 368 | 22 |
| Thực tế không có bệnh viêm phổi | 31 | 203 |

Confusion matrix này liên quan đến việc dự đoán xem một người có bệnh viêm phổi hay không dựa trên một bộ dữ liệu kiểm tra. Các ô trong ma trận biểu thị số lượng dự đoán chính xác và sai lệch.

* Ô trên cùng bên trái (368) biểu thị số lượng trường hợp bệnh nhân được dự đoán có bệnh viêm phổi và thực tế cũng có bệnh viêm phổi.
* Ô phía bên phải (22) biểu thị số lượng trường hợp bệnh nhân được dự đoán có bệnh viêm phổi nhưng thực tế lại không có bệnh viêm phổi.
* Ô phía dưới bên trái (31) biểu thị số lượng trường hợp bệnh nhân được dự đoán không có bệnh viêm phổi nhưng thực tế lại có bệnh viêm phổi.
* Ô phía dưới bên phải (203) biểu thị số lượng trường hợp bệnh nhân được dự đoán không có bệnh viêm phổi và thực tế cũng không có bệnh viêm phổi.

Dựa trên confusion matrix, chúng ta có thể tính toán độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc hiệu (specificity) và F1 score của mô hình dự đoán.

* Độ chính xác: (368+203)/(368+22+31+203) = 0.915
* Độ nhạy (sensitivity): 368/(368+31) = 0.922
* Độ đặc hiệu (specificity): 203/(203+22) = 0.902
* F1 score: 2\*(precision \* recall)/(precision + recall)

Với confusion matrix này, độ chính xác của mô hình là 0.915, có nghĩa là nó đã phân loại đúng 91.5% các bệnh nhân trong tập dữ liệu kiểm tra. Độ nhạy và độ đặc hiệu của mô hình cũng khá cao, 0.922 và 0.902 tương ứng, có nghĩa là mô hình đã đưa ra kết quả dự đoán tốt cho cả hai nhóm bệnh nhân (có bệnh viêm phổi và không có bệnh viêm phổi). Tuy nhiên, để đánh giá một mô hình dự đoán, chúng ta cần xem xét nhiều hơn về cách sử dụng mô hình và giá trị ngưỡng để quyết định nếu một bệnh nhân được xác định là có hoặc không có bệnh.

### Đánh giá mô hình thông qua classification report

Bảng 4.2. Bảng đánh giá mô hình thông qua classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1 – score | support |
| Pneumonia (Class 0) | 0.92 | 0.94 | 0.93 | 390 |
| Normal (Class 1) | 0.90 | 0.87 | 0.88 | 234 |
| Accuracy |  |  | 0.92 | 624 |
| Macro avg | 0.91 | 0.91 | 0.91 | 624 |
| Weighted avg | 0.91 | 0.92 | 0.91 | 624 |

Đối với lớp "Pneumonia" (Class 0), mô hình có độ chính xác (precision) là 0.92, tức là trong các trường hợp được dự đoán là bệnh viêm phổi, có 92% là đúng. Độ bao phủ (recall) là 0.94, tức là trong tổng số các trường hợp bệnh viêm phổi, mô hình có thể dự đoán chính xác 94%. F1-score là 0.93, là một đánh giá tổng hợp về độ chính xác và độ bao phủ của mô hình. Với lớp này, mô hình có khả năng phân loại tốt.

Đối với lớp "Normal" (Class 1), mô hình có độ chính xác (precision) là 0.90, tức là trong các trường hợp được dự đoán là bình thường, có 90% là đúng. Độ bao phủ (recall) là 0.87, tức là trong tổng số các trường hợp bình thường, mô hình có thể dự đoán chính xác 87%. F1-score là 0.88, là một đánh giá tổng hợp về độ chính xác và độ bao phủ của mô hình. Với lớp này, mô hình có khả năng phân loại tốt nhưng thấp hơn so với lớp "Pneumonia".

Accuracy của mô hình là 0.92, tức là tỉ lệ các dự đoán chính xác trên tổng số dữ liệu là 92%. Tuy nhiên, accuracy không phải là một chỉ số đủ tốt để đánh giá mô hình, đặc biệt khi các lớp dữ liệu không cân bằng về số lượng.

Macro avg (trung bình không trọng số) của precision, recall và F1-score là 0.91, tức là mô hình có khả năng phân loại tốt trung bình cho cả hai lớp dữ liệu.

Weighted avg (trung bình có trọng số) của precision, recall và F1-score cũng là 0.91, do số lượng dữ liệu của hai lớp là gần bằng nhau, nên trọng số của các lớp đều là 0.5. Chỉ số này cũng đánh giá khả năng phân loại trung bình của mô hình.

# KẾT LUẬN

## Kết quả đạt được

Trong quá trình thực hiện đề tài, em đã tìm hiểu và xây dựng được một số kỹ thuật trong lĩnh vực thị giác máy tính và học máy, cụ thể là mạng nơ ron tích chập. Trong việc phát hiện và dự đoán bệnh viêm phổi. Mô hình được xây dựng với độ chính xác lên đến hơn 90% cùng với độ mất mát được giảm đến mức tối thiểu.

## Hạn chế

Tuy nhiên trong quá trình thực hiện vẫn gặp phải một số vấn đề sai sót như: Vẫn có trường hợp sai lệch trong việc dự đoán; Tỉ lệ tập huấn luyện không đồng đều; Sự hạn chế của mô hình AI.

## Kết luận

Sử dụng công nghệ AI trong việc phát hiện và dự đoán bệnh viêm phổi là một hướng đi mới trong lĩnh vực y tế. Việc sử dụng AI giúp tăng độ chính xác, tiết kiệm thời gian và giảm thiểu sai lầm trong quá trình phân tích và đánh giá bệnh viêm phổi. Hình thức này sẽ mang lại nhiều lợi ích cho bệnh nhân và bác sĩ. Tuy nhiên, để đạt được kết quả tốt nhất, cần phải kết hợp sử dụng công nghệ AI với sự tư vấn của các chuyên gia y tế.

# TÀI LIỆU THAM KHẢO

1. David Beazley & Brian K. Jones (2013), *Python Cookbook 3rd edition*. O’Reilly Media.

2. Stef Maruch & Aahz Maruch (2006), *Python for dummy.* John Wiley & Sons.

3. Sebastian Raschka & Vahid Mirjalili (2019), *Python Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow 2 3rd edition*. Packt.

4. Andreas C. Muller & Sarah Guido (2013), *Introduction to Machine Learning with Python*. O’Reilly Media.