BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG

PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI

BỘ MÔN CÔNG NGHỆ THÔNG TIN

TÊN ĐỀ TÀI

Lung Sound Classification - Phân loại âm thanh phổi

Giảng viên:	**Cô Vũ Thị Hạnh**
Sinh viên thực hiện:	**2351267273 Nguyễn Hồng Nguyên** **2351267264 Trần Mạnh Hùng** **2351267259 Phạm Thị Quỳnh Giao**
Lớp:	**S26-65TTNT**

TP. Hồ Chí Minh, ngày … tháng … năm 2026

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày … tháng … năm 2026

Chữ ký của giảng viên

LỜI CẢM ƠN

Trước tiên, với tình cảm sâu sắc và chân thành nhất, nhóm em xin bày tỏ lòng biết ơn đến tất cả các cá nhân và tổ chức đã tạo điều kiện, hỗ trợ và giúp đỡ nhóm em trong suốt quá trình học tập và thực hiện đề tài này. Trong suốt thời gian học tập tại trường, nhóm em đã nhận được rất nhiều sự quan tâm, chỉ bảo tận tình của quý thầy cô và sự giúp đỡ quý báu từ bạn bè.

Với lòng biết ơn sâu sắc, nhóm em xin gửi lời cảm ơn chân thành đến quý thầy cô Bộ môn Công nghệ Thông tin – Phân hiệu Trường Đại học Thủy Lợi, những người đã truyền đạt cho nhóm em những kiến thức quý báu trong suốt quá trình học tập. Nhờ sự giảng dạy, hướng dẫn và động viên tận tình của quý thầy cô mà nhóm em có thể hoàn thành tốt đề tài này.

Đặc biệt, nhóm em xin gửi lời cảm ơn sâu sắc đến Cô Vũ Thị Hạnh, người đã trực tiếp hướng dẫn, tận tình giúp đỡ và định hướng cho nhóm em trong suốt quá trình thực hiện bài báo cáo.

Do thời gian có hạn và kiến thức còn nhiều hạn chế, bài báo cáo chắc chắn không tránh khỏi những thiếu sót. Nhóm em rất mong nhận được những ý kiến đóng góp quý báu của quý thầy cô để có thể hoàn thiện hơn kiến thức và kinh nghiệm của mình trong lĩnh vực này.

Nhóm em xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC ẢNH

DANH MỤC BẢNG

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

Viết tắt	Nghĩa tiếng Anh	Nghĩa tiếng Việt
AI	Artificial Intelligence	Trí tuệ nhân tạo
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
EDA	Exploratory Data Analysis	Phân tích dữ liệu khám phá
FFT	Fast Fourier Transform	Biến đổi Fourier nhanh
STFT	Short-Time Fourier Transform	Biến đổi Fourier ngắn hạn
Mel	Mel scale	Thang tần số Mel
MFCC	Mel-Frequency Cepstral Coefficients	Hệ số cepstral theo thang Mel
ROC	Receiver Operating Characteristic	Đường cong ROC
AUC	Area Under Curve	Diện tích dưới đường cong
ROC-AUC	Receiver Operating Characteristic – Area Under Curve	Chỉ số phân biệt lớp
TP	True Positive	Dự đoán đúng dương
FP	False Positive	Dự đoán sai dương
FN	False Negative	Dự đoán sai âm
ReLU	Rectified Linear Unit	Hàm kích hoạt chỉnh lưu
Grad-CAM	Gradient-weighted Class Activation Mapping	Bản đồ chú ý theo gradient
ImageNet	ImageNet Dataset	Tập dữ liệu ảnh chuẩn
BP	Baseline	Ghi âm chuẩn
DP	Disease	Ghi âm bệnh lý
EP	Emphysema	Khí phế thũng
COPD	Chronic Obstructive Pulmonary Disease	Bệnh phổi tắc nghẽn mạn tính
E W	Expiratory Wheeze	Khò khè khi thở ra
I E W	Inspiratory Expiratory Wheeze	Khò khè khi thở vào và ra
BRON	Bronchitis	Viêm phế quản
Crep	Crepitations	Ran khô
P	Posterior	Phía sau

Viết tắt	Nghĩa tiếng Anh	Nghĩa tiếng Việt
A	Anterior	Phía trước
L	Left	Trái
R	Right	Phải
U	Upper	Trên
M	Middle	Giữa
L (Lower)	Lower	Dưới

Ký hiệu và đại lượng kỹ thuật:

Ký hiệu	Ý nghĩa
Hz	Hertz – đơn vị tần số
kHz	Kilohertz = 1000 Hz
dB	Decibel – đơn vị cường độ âm
n_fft	Kích thước FFT
hop_length	Bước nhảy giữa các khung thời gian
n_mels	Số dải tần Mel
mean	Giá trị trung bình
std	Độ lệch chuẩn
SEED	Hạt giống ngẫu nhiên để tái lập kết quả
TP, FP, FN	Các thành phần trong Confusion Matrix
f(x) = max(0, x)	Hàm ReLU
(mel_db - mean) / std	Công thức chuẩn hóa dữ liệu
Sigmoid	Hàm kích hoạt cho phân loại 2 lớp
Softmax	Hàm kích hoạt cho phân loại nhiều lớp
BatchNormalization	Chuẩn hóa theo batch
MaxPooling	Gộp cực đại
Dense	Lớp kết nối đầy đủ

Viết tắt nhãn bệnh trong dữ liệu:

Mã	Bệnh lý
N	Normal – Bình thường
Asthma	Hen suyễn
C	Heart Failure – Suy tim

Mã	Bệnh lý
COPD	Phổi tắc nghẽn mạn tính
Pneumonia	Viêm phổi
Crep	Xơ phổi
Bronchial	Viêm phế quản
B	Pleural Effusion – Tràn dịch màng phổi

MỞ ĐẦU

Trong y học lâm sàng, việc nghe phổi đóng vai trò quan trọng trong chẩn đoán ban đầu các bệnh lý liên quan đến hệ hô hấp và tim mạch. Tuy nhiên, phương pháp này phụ thuộc nhiều vào kinh nghiệm của bác sĩ và có thể mang tính chủ quan. Với sự phát triển của học máy và học sâu, việc tự động phân tích tín hiệu âm thanh y sinh đang trở thành một hướng nghiên cứu tiềm năng, giúp hỗ trợ bác sĩ trong quá trình chẩn đoán.

Trong báo cáo này, nhóm tập trung xây dựng mô hình học máy nhằm phân loại âm thanh phổi thành hai nhóm: bình thường và bất thường. Dữ liệu âm thanh được tiền xử lý và chuyển đổi sang dạng Mel-spectrogram, sau đó sử dụng mạng nơ-ron tích chập để thực hiện phân loại.

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1.1. Lý do chọn đề tài

Chẩn đoán bệnh phổi thông thường dựa vào khám lâm sàng, trong đó nghe phổi là một phương pháp quan trọng. Tuy nhiên, phương pháp này phụ thuộc nhiều vào kinh nghiệm của bác sĩ, dễ xảy ra sai sót và khó chuẩn hóa. Ngoài ra, không phải ai cũng có cơ hội tiếp cận các chuyên gia y tế, đặc biệt ở các vùng sâu, vùng xa.

Trong bối cảnh đó, việc ứng dụng trí tuệ nhân tạo để tự động phân loại âm thanh phổi là một hướng đi có tính thực tiễn cao. Với sự phát triển mạnh mẽ của học sâu, đặc biệt là mạng nơ-ron tích chập (CNN), các hệ thống phân loại âm thanh có thể đạt độ chính xác rất cao và hoạt động ổn định trên nhiều loại dữ liệu khác nhau.

Bên cạnh đó, các mô hình học chuyển giao như MobileNetV2 cho phép tận dụng các mô hình đã được huấn luyện trên tập dữ liệu lớn như ImageNet, từ đó giảm đáng kể thời gian huấn luyện và cải thiện độ chính xác khi làm việc với tập dữ liệu có quy mô vừa và nhỏ.

Đề tài còn đưa vào thử nghiệm mô hình EfficientNet-B0. Trong khai phá dữ liệu hình ảnh (Image Mining), EfficientNet-B0 nổi bật nhờ kỹ thuật Compound Scaling, giúp tối ưu hóa đồng thời ba yếu tố: chiều sâu, chiều rộng và độ phân giải của mạng.

Từ những phân tích trên, nhóm quyết định lựa chọn đề tài xây dựng hệ thống phân loại âm thanh phổi dựa trên mạng nơ-ron tích chập nhằm nghiên cứu, so sánh hiệu quả của các mô hình khác nhau và xây dựng một ứng dụng minh họa có khả năng sử dụng trong thực tế.

1.2. Mục tiêu của đề tài

Mục tiêu chung của đề tài là xây dựng một hệ thống hoàn chỉnh có khả năng tự động phân loại âm thanh phổi thành hai nhóm chính: Bình thường (Normal) và Bất thường (Abnormal), dựa trên dữ liệu âm thanh đầu vào.

Để đạt được mục tiêu chung đó, đề tài hướng đến các mục tiêu cụ thể sau:

Thứ nhất, nghiên cứu và áp dụng các kỹ thuật xử lý tín hiệu âm thanh và tiền xử lý dữ liệu nâng cao nhằm đảm bảo dữ liệu đầu vào có chất lượng tốt và phù hợp cho việc huấn luyện mô hình học sâu. Điều này bao gồm giảm nhiễu, lọc dải tần số, trừ nhiễu phổ, nén dải động, nâng cao tần số cao, cắt bỏ khoảng lặng, chuẩn hóa độ dài và trích xuất Mel Spectrogram.

Thứ hai, xây dựng và huấn luyện hai mô hình khác nhau, bao gồm một mô hình CNN cơ bản được xây dựng từ đầu và mô hình MobileNetV2, mô hình efficientnet-b0 sử dụng học chuyển giao, từ đó so sánh hiệu quả của từng mô hình trên cùng một tập dữ liệu.

Thứ ba, đánh giá kết quả huấn luyện thông qua các chỉ số như Accuracy, Precision, Recall, F1-Score và ROC-AUC để có cái nhìn toàn diện về chất lượng mô hình.

Thứ tư, xây dựng chức năng trực quan hóa Grad-CAM với cấu trúc 3 tầng nhằm giải thích quá trình ra quyết định của mô hình, giúp người dùng hiểu được mô hình tập trung vào vùng nào trên Mel Spectrogram khi đưa ra kết quả phân loại.

Thứ năm, xây dựng một ứng dụng minh họa cho phép người dùng tải file âm thanh lên và nhận kết quả dự đoán (Bình thường hoặc Bệnh lý) một cách trực quan.

Các mục tiêu trên không chỉ nhằm xây dựng một hệ thống hoạt động đúng mà còn hướng đến việc tạo ra một sản phẩm có khả năng mở rộng và ứng dụng trong thực tế.

1.3. Đ tượng nghiên cứu

Đề tài tập trung nghiên cứu bài toán phân loại âm thanh phổi dựa trên tín hiệu âm thanh, với mục tiêu xây dựng một hệ thống có khả năng tự động nhận dạng tình trạng phổi thông qua dữ liệu âm thanh đầu vào.

Đối tượng nghiên cứu của đề tài là các file âm thanh phổi thuộc hai nhóm chính: Bình thường (Normal) và Bất thường (Abnormal).

Nhóm Bất thường bao gồm các loại bệnh như: Hen suyễn (Asthma), Bệnh phổi tắc nghẽn mãn tính (COPD), Suy tim (Heart Failure), Viêm phổi (Pneumonia), Xơ phổi (Lung Fibrosis), Viêm phế quản (Bronchitis) và Tràn dịch màng phổi (Pleural Effusion).

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Tổng quan về xử lý tín hiệu âm thanh

Xử lý tín hiệu âm thanh là một lĩnh vực thuộc kỹ thuật điện tử và xử lý tín hiệu số, nghiên cứu các phương pháp để thu nhận, xử lý, phân tích và tái tạo tín hiệu âm thanh. Âm thanh là một dạng sóng cơ học lan truyền qua các chất lỏng, chất rắn hoặc khí, được đặc trưng bởi tần số, biên độ và pha.

Tần số âm thanh được đo bằng Hertz (Hz), thể hiện số lần dao động trong một giây. Tai người có thể nghe được tần số từ khoảng 20 Hz đến 20.000 Hz. Biên độ thể hiện độ lớn của sóng, liên quan trực tiếp đến cường độ âm thanh. Pha thể hiện vị trí của sóng tại một thời điểm nhất định.

Trong lĩnh vực y tế, xử lý tín hiệu âm thanh được ứng dụng để phân tích âm thanh tim, âm thanh phổi, âm thanh dạ dày và các âm thanh sinh học khác. Việc phân tích âm thanh phổi giúp bác sĩ phát hiện các bệnh lý như hen suyễn, viêm phổi, xơ phổi và các bệnh khác.

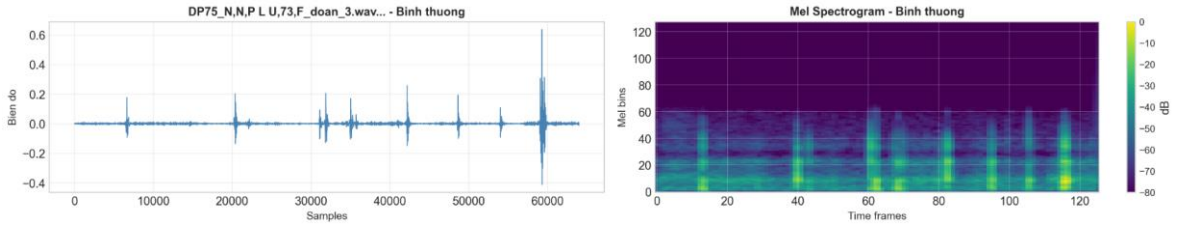
Các đặc trưng cơ bản của tín hiệu âm thanh bao gồm tần số, biên độ, thời gian và năng lượng. Từ các đặc trưng này, có thể trích xuất các đặc trưng cao hơn như Mel Spectrogram, MFCC (Mel-Frequency Cepstral Coefficients), và các đặc trưng khác.

2.2. Mel Spectrogram

Mel Spectrogram là một biểu diễn thời gian-tần số của tín hiệu âm thanh, được tính toán dựa trên thang đo Mel. Thang đo Mel là một thang đo phi tuyến tính phản ánh cách tai người cảm nhận tần số. Tai người cảm nhận tần số không phải theo cách tuyến tính mà theo cách phi tuyến, tức là sự thay đổi tần số ở vùng tần số thấp được cảm nhận rõ hơn so với vùng tần số cao.

Mel Spectrogram được sử dụng rộng rãi trong các bài toán xử lý âm thanh như nhận dạng giọng nói, phân loại âm thanh, phát hiện sự kiện âm thanh và các bài toán khác.

Ưu điểm của Mel Spectrogram là nó phản ánh cách cảm nhận tần số của tai người, giúp mô hình học sâu học được các đặc trưng phù hợp hơn.



Ảnh 1 Ảnh sóng và ảnh phổ của phổi bình thường

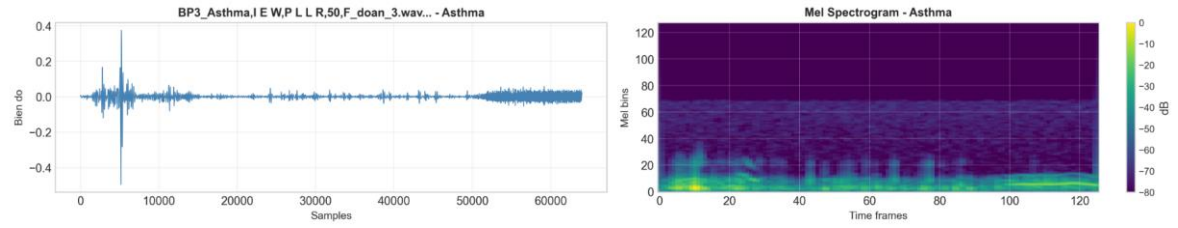
Phổi bình thường (Normal)
Ảnh sóng:

- Biên độ nhỏ, đều
- Ít đỉnh nhọn bất thường
- Nhịp thở tương đối ổn định

Mel Spectrogram:

- Năng lượng tập trung chủ yếu ở tần số thấp
- Màu sắc khá đồng đều, không có "vệt sáng" kéo dài
- Ít nhiễu ở tần số cao

Đặc trưng: Âm thở mềm, liên tục, ít biến động.



Ảnh 2 Ảnh sóng và ảnh phổ của bệnh phổi Asthma(Hen suyễn)

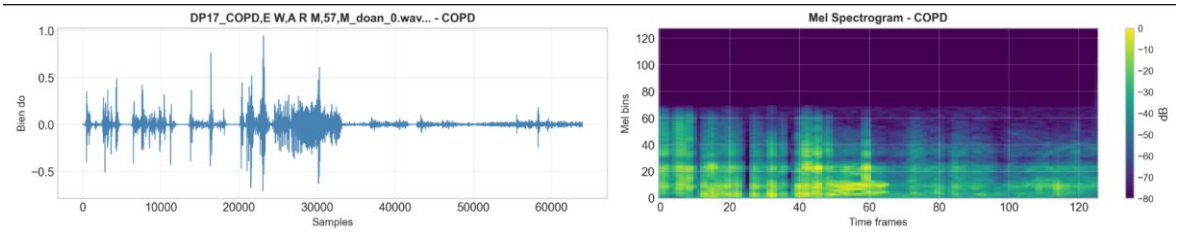
Hen suyễn (Asthma)
Ảnh sóng:

- Có những đoạn biên độ tăng mạnh đột ngột
- Xuất hiện các chuỗi dao động kéo dài

Mel Spectrogram:

- Có nhiều dải sáng ở tần số trung bình
- Các vệt sáng kéo dài theo trục thời gian → biểu hiện wheeze (khò khè)

Đặc trưng: Âm khò khè cao, kéo dài, xuất hiện rõ ở tần số trung–cao.



Ảnh 3 Ảnh sóng và ảnh phổ của bệnh phổi phổi tắc nghẽn mãn tính (COPD)

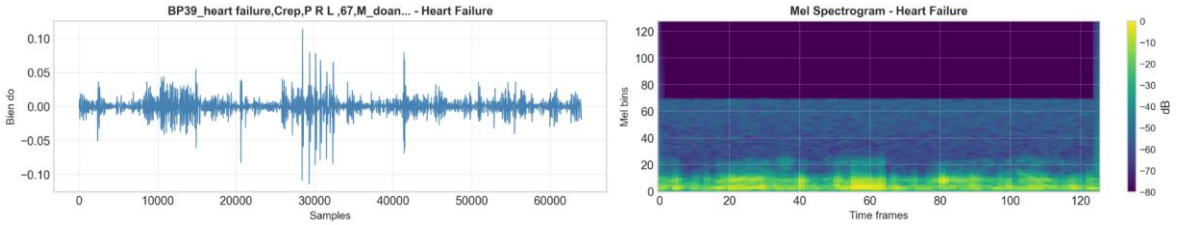
COPD – Phổi tắc nghẽn mạn tính
Ảnh sóng:

- Dao động mạnh, không đều
- Biên độ lớn trong nhiều đoạn
- Nhịp thở nặng, gấp

Mel Spectrogram:

- Năng lượng trải rộng từ thấp đến trung
- Nhiều vùng sáng liên tục
- Phổ "dày", không sạch như phổi bình thường

Đặc trưng: Âm thở nặng, kéo dài, có wheeze nhưng thấp hơn hen.



Ảnh 4 Ảnh sóng và ảnh phổ của bệnh phổi Suy tim (Heart Failure)

Suy tim (Heart Failure)
Ảnh sóng:

- Nhiều đỉnh nhọn ngắn, rời rạc
- Biên độ không đều, kiểu "lốm đốm"

Mel Spectrogram:

- Nhiều chấm sáng ngắn ở tần số thấp–trung
- Không kéo dài thành dải như hen
- Phổ không đều theo thời gian

Đặc trưng: Ran ẩm (crackles): âm ngắn, rời rạc, như tiếng nổ lách tách.

2.3. Mạng nơ-ron tích chập (CNN)

Mạng nơ-ron tích chập là một loại mạng nơ-ron nhân tạo được thiết kế đặc biệt cho dữ liệu dạng lưới, trong đó hình ảnh và Mel Spectrogram là các dạng dữ liệu phổ biến nhất. CNN khai thác cấu trúc không gian của dữ liệu thông qua các phép tích chập để trích xuất đặc trưng cục bộ.

Một mô hình CNN điển hình bao gồm các thành phần cơ bản sau:

Lớp tích chập (Convolutional Layer) sử dụng các bộ lọc nhỏ trượt trên dữ liệu đầu vào để tạo ra các bản đồ đặc trưng. Mỗi bộ lọc sẽ học một loại đặc trưng cụ thể. Số lượng bộ lọc tăng dần từ các lớp đầu đến các lớp sau, giúp mô hình học được các đặc trưng ngày càng trừu tượng hơn.

Hàm kích hoạt phi tuyến, ReLU (Rectified Linear Unit), giúp mô hình học được các mối quan hệ phi tuyến trong dữ liệu. ReLU được định nghĩa là $f(x) = \max(0, x)$, tức là nó giữ nguyên các giá trị dương và đặt các giá trị âm thành 0.

Lớp gộp (Pooling Layer), thường là MaxPooling, có nhiệm vụ giảm kích thước không gian của bản đồ đặc trưng, từ đó giảm số tham số và hạn chế hiện tượng quá khớp. MaxPooling chọn giá trị lớn nhất trong một cửa sổ nhỏ.

Lớp làm phẳng (Flatten Layer) chuyển đổi bản đồ đặc trưng 2D thành một vector 1D để chuẩn bị cho các lớp kết nối đầy đủ.

Các lớp kết nối đầy đủ (Fully Connected Layers) tổng hợp thông tin từ các đặc trưng đã học để đưa ra kết quả phân loại cuối cùng. Lớp cuối cùng sử dụng hàm Sigmoid (cho phân loại 2 lớp) hoặc Softmax (cho phân loại nhiều lớp) để chuyển đổi đầu ra thành xác suất cho mỗi lớp.

CNN có ưu điểm nổi bật là tận dụng được tính cục bộ của dữ liệu và khả năng chia sẻ trọng số, giúp giảm đáng kể số lượng tham số so với các mạng nơ-ron truyền thống.

Nhờ đó, CNN đạt hiệu quả cao trong các bài toán phân loại và nhận dạng hình ảnh, cũng như phân loại âm thanh.

2.4. Học chuyển giao (Transfer Learning)

Học chuyển giao là kỹ thuật sử dụng tri thức đã học từ một bài toán hoặc một tập dữ liệu lớn để áp dụng cho một bài toán khác có liên quan. Trong học sâu, học chuyển giao thường được thực hiện bằng cách sử dụng các mô hình đã được huấn luyện trước trên những tập dữ liệu quy mô lớn như ImageNet.

Các mô hình tiền huấn luyện đã học được các đặc trưng tổng quát của hình ảnh như cạnh, màu sắc, hình dạng và cấu trúc. Khi áp dụng cho bài toán mới, các lớp đầu của mô hình thường được giữ nguyên, chỉ huấn luyện lại hoặc tinh chỉnh các lớp cuối để phù hợp với số lớp và đặc điểm của dữ liệu mới.

Phương pháp này mang lại nhiều lợi ích:

Thứ nhất, giảm thời gian huấn luyện. Vì các lớp đầu đã được huấn luyện trên tập dữ liệu lớn, chúng ta chỉ cần huấn luyện lại các lớp cuối, tiết kiệm thời gian đáng kể.

Thứ hai, hạn chế yêu cầu về dung lượng dữ liệu. Các mô hình tiền huấn luyện đã học được các đặc trưng tổng quát, nên chúng ta không cần tập dữ liệu quá lớn để đạt độ chính xác cao.

Thứ ba, giúp mô hình đạt độ chính xác cao hơn so với việc huấn luyện từ đầu, đặc biệt khi tập dữ liệu có quy mô nhỏ.

Do đó, học chuyển giao đặc biệt phù hợp với các bài toán có tập dữ liệu vừa và nhỏ như trong đề tài này.

2.5. Kiến trúc MobileNetV2

MobileNetV2 là một mô hình học sâu được thiết kế với mục tiêu tối ưu hóa về tốc độ và bộ nhớ, phù hợp cho các thiết bị có tài nguyên hạn chế như điện thoại di động hoặc hệ thống nhúng.

Mô hình sử dụng kỹ thuật tích chập tách biệt theo chiều sâu (depthwise separable convolution) nhằm giảm số phép tính so với tích chập thông thường. Thay vì áp dụng một bộ lọc 3x3 với tất cả các kênh đầu vào, tích chập tách biệt chia thành hai bước: tích chập theo chiều sâu (depthwise convolution) áp dụng một bộ lọc 3x3 cho mỗi kênh đầu vào, sau đó tích chập điểm (pointwise convolution) sử dụng bộ lọc 1x1 để kết hợp các kênh.

Ngoài ra, MobileNetV2 còn áp dụng kiến trúc khối đảo ngược (inverted residual block) kết hợp với lớp bottleneck tuyến tính, giúp duy trì khả năng biểu diễn đặc trưng trong khi vẫn giữ được số lượng tham số ở mức thấp.

Trong đề tài, MobileNetV2 được sử dụng như một mô hình học chuyển giao. Phần backbone của mô hình được giữ nguyên trọng số đã huấn luyện trước trên ImageNet, sau đó bổ sung các lớp phân loại mới để phù hợp với bài toán phân loại 2 lớp âm thanh phổi.

2.6. Kiến trúc efficientnet-b0

Nếu như mobilenetv2 tập trung vào việc giảm thiểu số lượng phép tính thông qua tích chập tách biệt, thì efficientnet-b0 đại diện cho một bước tiến mới trong việc tối ưu hóa hiệu suất mô hình thông qua phương pháp compound scaling (thay đổi tỷ lệ đồng thời).

Nguyên lý Compound Scaling: Thay vì chỉ tăng chiều sâu (số lớp), chiều rộng (số kênh) hoặc độ phân giải của ảnh đầu vào một cách riêng lẻ như các mô hình truyền thống, efficientnet-b0 sử dụng một hệ số kép để mở rộng cả ba chiều này một cách cân bằng. Điều

này giúp mô hình đạt được độ chính xác cao hơn trong khi vẫn duy trì được hiệu suất tính toán cực kỳ ấn tượng.

Cấu trúc MBConv và Squeeze-and-Excitation (SE): Thành phần cốt lõi của efficientnet-b0 là khối mbconv (mobile inverted bottleneck), tương tự như mobilenetv2 nhưng được tích hợp thêm cơ chế squeeze-and-excitation. Cơ chế này cho phép mô hình tự động gán trọng số cho từng kênh đặc trưng, giúp mạng "tập trung" vào những thông tin quan trọng nhất trên mel spectrogram (ví dụ: các dải tần số đặc trưng của tiếng ran phổi) và loại bỏ các thông tin nhiễu.

Ứng dụng trong đề tài: Trong hệ thống này, efficientnet-b0 đóng vai trò là mô hình đối trọng với mobilenetv2. Với cấu trúc tối ưu, mô hình này được kỳ vọng sẽ khai phá được các đặc trưng phi tuyến tính phức tạp hơn trong âm thanh bệnh lý, từ đó cải thiện các chỉ số như recall và f1-score cho nhóm bệnh nhân có biểu hiện lâm sàng không rõ ràng.

2.7. Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM là một phương pháp để trực quan hóa và giải thích quyết định của các mô hình học sâu. Phương pháp này tính toán gradient của lớp dự đoán đối với các bản đồ đặc trưng của lớp tích chập cuối cùng, sau đó kết hợp để tạo ra một heatmap thể hiện vùng dữ liệu mà mô hình tập trung khi đưa ra quyết định.

Grad-CAM giúp tăng tính giải thích của mô hình, cho phép người dùng hiểu được mô hình tập trung vào vùng nào khi đưa ra quyết định. Điều này đặc biệt quan trọng trong các ứng dụng y tế, nơi cần phải hiểu rõ lý do của kết quả dự đoán.

2.8. Các chỉ số đánh giá mô hình phân loại

Để đánh giá chất lượng của mô hình phân loại, đề tài sử dụng nhiều chỉ số khác nhau nhằm phản ánh toàn diện hiệu quả hoạt động của hệ thống.

Accuracy là tỉ lệ giữa số mẫu được dự đoán đúng và tổng số mẫu trong tập dữ liệu. Chỉ số này đơn giản và dễ hiểu nhưng có thể gây sai lệch khi dữ liệu giữa các lớp không cân bằng.

Precision thể hiện mức độ chính xác của các dự đoán dương, cho biết trong số các mẫu được mô hình dự đoán thuộc một lớp nhất định, có bao nhiêu mẫu thực sự đúng. Precision được tính bằng công thức:

Precision = TP / (TP + FP), trong đó TP là True Positive và FP là False Positive.

Recall phản ánh khả năng phát hiện đầy đủ các mẫu thuộc một lớp, cho biết trong số các mẫu thực sự thuộc lớp đó, mô hình dự đoán đúng được bao nhiêu. Recall được tính bằng công thức:
Recall = TP / (TP + FN), trong đó FN là False Negative.

F1-Score là trung bình điều hòa của Precision và Recall, giúp đánh giá sự cân bằng giữa hai chỉ số này. F1-Score được tính bằng công thức:
F1 = 2 * (Precision * Recall) / (Precision + Recall).

ROC-AUC là diện tích dưới đường cong ROC (Receiver Operating Characteristic), thể hiện khả năng phân biệt giữa các lớp của mô hình. Đối với bài toán phân loại 2 lớp, chỉ số này thường được tính trực tiếp.

Confusion Matrix là một bảng thể hiện số lượng mẫu được dự đoán đúng và sai cho mỗi lớp. Từ Confusion Matrix, có thể tính toán các chỉ số khác như Precision, Recall và F1-Score cho mỗi lớp.

Việc sử dụng đồng thời nhiều chỉ số giúp đánh giá mô hình một cách khách quan và toàn diện hơn, tránh phụ thuộc vào một thước đo duy nhất.

CHƯƠNG 3: MÔ TẢ DỮ LIỆU VÀ TIỀN XỬ LÝ

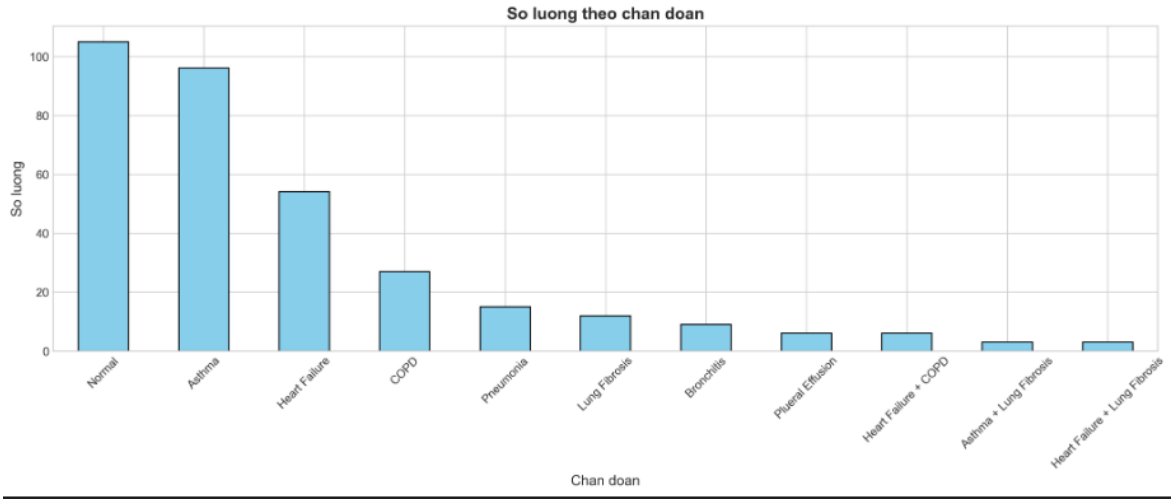
3.1. Nguồn và đặc điểm tập dữ liệu

Tập dữ liệu được sử dụng trong đề tài là tập âm thanh phổi phục vụ cho bài toán phân loại tình trạng phổi. Dữ liệu được tổ chức trong thư mục data/Audio Files/, trong đó mỗi file tương ứng với một bản ghi âm thanh phổi của một bệnh nhân.

Theo kết quả thống kê, tổng số file âm thanh trong toàn bộ tập dữ liệu là 336 file. Tần số lấy mẫu của tất cả các file là 16 kHz, độ dài cắt về 5 giây.

Quan trọng là cần lưu ý rằng bài toán phân loại trong đề tài được quy về 2 lớp chính: Bình thường (Normal) và Bất thường (Abnormal). Điều này có ý nghĩa y tế quan trọng vì mục tiêu chính là phát hiện sớm các bệnh lý phổi, không nhất thiết phải phân biệt chi tiết từng loại bệnh.

Phân bố cụ thể theo từng chẩn đoán như sau:



Ảnh 5 Phân bố các loại bệnh

Bình thường (Normal - N): 105 file, chiếm 31.3% tổng số dữ liệu.

Hen suyễn (Asthma): 96 file, chiếm 28.6%.

Suy tim (Heart Failure): 63 file, chiếm 18.8%.

Bệnh phổi tắc nghẽn mãn tính (COPD): 33 file, chiếm 9.8%.

Viêm phổi (Pneumonia): 15 file, chiếm 4.5%.

Xơ phổi (Lung Fibrosis): 15 file, chiếm 4.5%.

Viêm phế quản (Bronchitis - BRON): 9 file, chiếm 2.7%.

Tràn dịch màng phổi (Pleural Effusion): 6 file, chiếm 1.8%.

Từ số liệu trên có thể thấy dữ liệu không cân bằng giữa các lớp. Cụ thể, lớp Bình thường chiếm 31.3% trong khi lớp Bất thường chiếm 68.7%. Sự mất cân bằng này có thể dẫn đến mô hình thiên lệch về các lớp có số lượng file nhiều hơn. Do đó, cần áp dụng các kỹ thuật như oversampling để cân bằng dữ liệu trong quá trình huấn luyện.

3.2. Phân loại bệnh lý và đặc trưng âm thanh phổi

Bảng dưới đây tóm tắt các bệnh lý được nghiên cứu, tên tiếng Anh, tên tiếng Việt, đặc trưng âm thanh và số lượng file thực tế:

Bảng 1 Tóm tắt các bệnh lý được nghiên cứu

Mã	Bệnh lý (Tiếng Anh)	Bệnh lý (Tiếng Việt)	Đặc trưng âm thanh	Số file	Tỷ lệ
N	Normal	Bình thường	Âm thanh phế nang mềm, liên tục (100–400 Hz)	105	31.3%
Asthma	Asthma	Hen suyễn	Tiếng khò khè cao, kéo dài (400–2000 Hz)	96	28.6%
C	Heart Failure	Suy tim	Tiếng ran ẩm, rời rạc (100–500 Hz)	63	18.8%

Mã	Bệnh lý (Tiếng Anh)	Bệnh lý (Tiếng Việt)	Đặc trưng âm thanh	Số file	Tỷ lệ
COPD	COPD	Bệnh phổi tắc nghẽn	Tiếng khò khè kéo dài, âm thanh yếu (200–1000 Hz)	33	9.8%
Pneumonia	Pneumonia	Viêm phổi	Tiếng ran ẩm rõ ràng, nhiều (100–800 Hz)	15	4.5%
Crep	Lung Fibrosis	Xơ phổi	Tiếng ran khô sắc nét, giống Velcro (200–1000 Hz)	15	4.5%
Bronchial	Bronchitis	Viêm phế quản	Tiếng khò khè + ran ẩm kết hợp (300–1500 Hz)	9	2.7%
B	Pleural Effusion	Tràn dịch màng phổi	Âm thanh yếu, bị che phủ (50–200 Hz)	6	1.8%

Tổng cộng: Bình thường 105 file(31,3%), bất thường 231file (68,7%)

3.3. Cấu trúc thư mục dữ liệu

Dữ liệu được tổ chức trong thư mục data/Audio Files/.

Cách đặt tên file tuân theo quy tắc:
[Prefix]_[ChanDoan],[AmThanh],[ViTri],[Tuoi],[GioiTinh].wav

Trong đó:

Prefix: BP (Baseline), DP (Disease) hoặc EP (Emphysema) - phân loại ban đầu của dữ liệu.

ChanDoan: Chẩn đoán (N, Asthma, COPD, C, Pneumonia, Crep, Bronchial, B) - loại bệnh hoặc tình trạng bình thường.

AmThanh: Loại âm thanh được ghi âm. Các ký hiệu phổ biến gồm:

- N: Normal (Bình thường)

- E W: Expiratory Wheeze (Tiếng khò khè khi thở ra)

- I E W: Inspiratory Expiratory Wheeze (Tiếng khò khè khi thở vào và ra)

- C: Crackles (Tiếng ran ẩm)

- Crep: Crepitations (Tiếng ran khô)

- Bronchial: Tiếng phế quản

ViTri: Vị trí nghe trên cơ thể bệnh nhân. Các ký hiệu phổ biến gồm:

- P: Posterior (Phía sau)

- A: Anterior (Phía trước)

- L: Left (Bên trái)

- R: Right (Bên phải)

- U: Upper (Phía trên)

- M: Middle (Phía giữa)

- L: Lower (Phía dưới)

Ví dụ: P L L = Posterior Left Lower (Phía sau, bên trái, phía dưới)

Tuoi: Tuổi bệnh nhân (tính bằng năm).

GioiTinh: M (Nam) hoặc F (Nữ).

Ví dụ: BP1_Asthma,I E W,P L L,70,M.wav là file âm thanh của bệnh nhân nam 70 tuổi, chẩn đoán hen suyễn, loại âm thanh tiếng khò khè khi thở vào và ra, vị trí nghe phía sau bên trái phía dưới.

3.4. Phân tích dữ liệu khám phá (EDA)

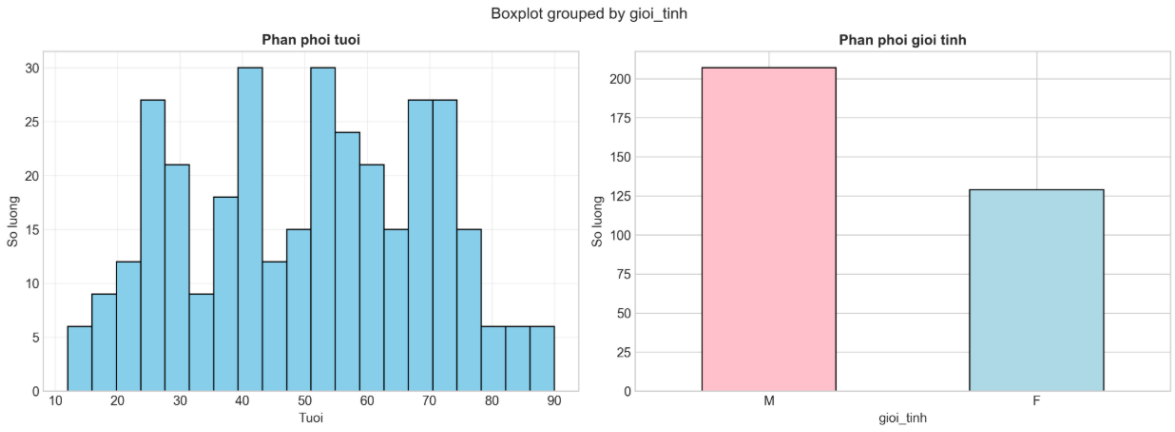
Phân tích dữ liệu khám phá được thực hiện để hiểu rõ hơn về đặc điểm của tập dữ liệu. Các chức năng chính của bước này gồm:

Thứ nhất: thống kê số lượng file theo từng bệnh lý. Kết quả cho thấy dữ liệu không cân bằng giữa các lớp, với số lượng file dao động từ 6 đến 105 file cho mỗi lớp.

Thứ hai: phân tích phân bố theo tuổi và giới tính. Tuổi của bệnh nhân dao động từ 12 đến 90 tuổi, với trung bình khoảng 50 tuổi. Giới tính được chia thành nam (M) và nữ (F), với tỷ lệ tương đối cân bằng.

Thứ ba: thống kê độ dài của các file âm thanh. Độ dài cắt về 4 giây.

Thứ tư: vẽ biểu đồ thể hiện sự phân bố dữ liệu giữa các lớp, phân bố theo tuổi, giới tính và độ dài.



Ảnh 6 Phân bổ tuổi và giới tính

Các biểu đồ EDA giúp nhóm hiểu rõ hơn về đặc điểm của dữ liệu và xác định các bước tiền xử lý cần thiết.

3.5. Tiền xử lý âm thanh

Tiền xử lý âm thanh là bước quan trọng để chuẩn bị dữ liệu cho huấn luyện mô hình. Các bước tiền xử lý được thực hiện theo quy trình nâng cao được cài đặt trong CNN_preprocessing.py, gồm các bước sau:

Thứ nhất, giảm nhiễu (Noise Reduction). Sử dụng thư viện noisereduce để giảm nhiễu nền từ tín hiệu âm thanh. Phương pháp này ước tính profile nhiễu từ các phần yên tĩnh và trừ nó khỏi tín hiệu, giúp làm sạch âm thanh mà vẫn giữ được các đặc trưng quan trọng.

Thứ hai, lọc dải tần số (Bandpass Filter). Áp dụng bộ lọc Butterworth bậc 3 để giữ lại chỉ các tần số trong khoảng 50-4000 Hz, đây là dải tần số chính của âm thanh phổi. Các tần số ngoài khoảng này được loại bỏ, giúp giảm nhiễu và tập trung vào thông tin hữu ích.

Thứ ba, trừ nhiễu phổ (Spectral Subtraction). Kỹ thuật này tính toán STFT (Short-Time Fourier Transform) của tín hiệu, ước tính profile nhiễu từ các khung đầu tiên, sau đó trừ nó khỏi phổ tín hiệu. Mức độ trừ được điều chỉnh (0.4) để tránh làm mất thông tin quan trọng.

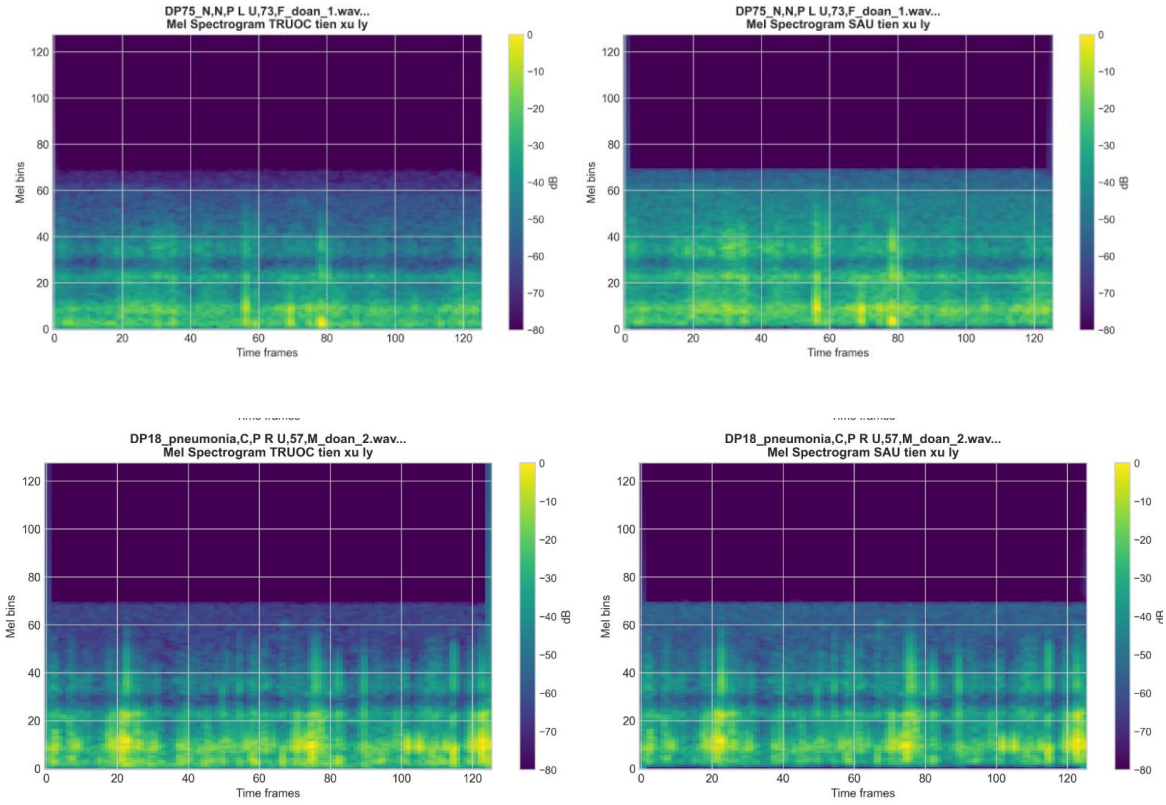
Thứ tư, nén dải động (Dynamic Range Compression). Kỹ thuật này giảm khoảng cách giữa các phần yếu và phần mạnh của tín hiệu. Bằng cách áp dụng compression ratio 3:1 với ngưỡng -35 dB, các phần yếu trở nên rõ ràng hơn và các phần mạnh không quá lớn, giúp mô hình học được các đặc trưng ở tất cả các mức độ biên độ.

Thứ năm, nâng cao tần số cao (Pre-emphasis). Áp dụng bộ lọc pre-emphasis với hệ số 1.2 để nâng cao các tần số cao. Điều này giúp làm nổi bật các đặc trưng ở tần số cao như tiếng khò khè, giúp mô hình phân biệt tốt hơn giữa các loại bệnh lý.

Thứ sáu, chuẩn hóa biên độ (Amplitude Normalization). Các file âm thanh có biên độ khác nhau. Bước này chuẩn hóa biên độ bằng cách chia cho giá trị biên độ cực đại, đưa tất cả các file về cùng một phạm vi giá trị từ -1 đến 1.

Thứ bảy, cắt bỏ khoảng lặng (Silence Trimming). Nhiều file âm thanh có khoảng lặng ở đầu hoặc cuối, không chứa thông tin hữu ích. Bước này phát hiện và loại bỏ các khoảng lặng này bằng cách sử dụng ngưỡng 20 dB, giúp tập trung vào phần âm thanh có ý nghĩa.

Thứ tám, chuẩn hóa độ dài (Duration Normalization). Các file âm thanh có độ dài khác nhau, từ 4 đến 10 giây. Để phù hợp với đầu vào của mô hình, tất cả các file được chuẩn hóa về độ dài 4 giây. Nếu file ngắn hơn 4 giây, nó được padding (thêm các mẫu 0 ở cuối). Nếu file dài hơn 4 giây, nó được cắt ngắn.



Ảnh 7 Ảnh ví dụ về trước và sau tiền xử lý

3.6. Trích xuất đặc trưng

Trích xuất đặc trưng là bước chuyển đổi tín hiệu âm thanh thô thành các đặc trưng có ý nghĩa cho mô hình học sâu. Bước này được thực hiện trong file feature_engineering.py, hàm chuyen_sang_mel(), gồm các bước sau:

Thứ nhất, chuẩn hóa độ dài. Tất cả file âm thanh được chuẩn hóa về 4 giây (64,000 mẫu ở 16kHz). Nếu file ngắn hơn, nó được padding (thêm 0). Nếu file dài hơn, nó được cắt ngắn.

Thứ hai, tăng cường dữ liệu (nếu cần). Áp dụng các phép biến đổi như time shift (dịch thời gian), thêm noise (nhiễu ngẫu nhiên), pitch shift (dịch cao độ), time stretch (kéo dãn thời gian) với mức độ khác nhau.

Thứ ba, tính toán Mel Spectrogram. Sử dụng librosa.feature.melspectrogram() với các tham số:

- n_fft=2048: Kích thước FFT, quyết định độ phân giải tần số

- hop_length=512: Bước nhảy giữa các khung, quyết định độ phân giải thời gian

- n_mels=128: Số mel bins, tương ứng với 128 dải tần số

Kết quả là ma trận 128 × 126 (128 mel bins × 126 khung thời gian)

Thứ tư, chuyển sang thang đo dB. Sử dụng librosa.power_to_db(mel, ref=np.max) để chuyển từ thang đo tuyến tính sang thang đo dB (decibel), khoảng từ -80 đến 0 dB. Việc này giúp dữ liệu phù hợp hơn với cách cảm nhận của tai người (tai người cảm nhận âm thanh theo thang đo logarit).

Thứ năm, chuẩn hóa dữ liệu. Sử dụng mean/std từ tập train để chuẩn hóa: mel_chuan = (mel_db - mean_train) / (std_train + 1e-6). Việc chuẩn hóa này đảm bảo dữ liệu có trung bình 0 và độ lệch chuẩn 1, giúp mô hình học tốt hơn.

Thứ sáu, reshape output. Thêm một chiều mới để tạo shape (128, 126, 1) - 128 mel bins × 126 khung thời gian × 1 channel, phù hợp với đầu vào của CNN.

3.7. Tăng cường dữ liệu (Data Augmentation)

Do số lượng file âm thanh không quá lớn (336 file) và điều kiện ghi âm trong thực tế rất đa dạng, nhóm áp dụng các kỹ thuật tăng cường dữ liệu nhằm tạo ra nhiều biến thể khác nhau của cùng một file trong quá trình huấn luyện. Việc này giúp mô hình học được tính đa dạng của dữ liệu, giảm hiện tượng quá khớp và tăng khả năng tổng quát hóa.

Các kỹ thuật tăng cường dữ liệu được áp dụng gồm:

Thứ nhất, Spectral Subtraction (Trừ nhiễu phổ). Kỹ thuật này loại bỏ nhiễu nền từ tín hiệu âm thanh bằng cách tính toán STFT, ước tính profile nhiễu từ các khung đầu tiên, sau đó trừ nó khỏi phổ tín hiệu. Mức độ trừ được điều chỉnh (0.4) để tránh làm mất thông tin quan trọng. Phương pháp này giúp tạo ra các biến thể âm thanh với mức nhiễu khác nhau.

Thứ hai, Dynamic Range Compression (Nén dải động). Kỹ thuật này chuẩn hóa phạm vi động của tín hiệu bằng cách áp dụng compression ratio 3:1 với ngưỡng -35 dB. Điều này giúp các phần yếu trở nên rõ ràng hơn và các phần mạnh không quá lớn. Kỹ thuật này giúp mô hình học được các đặc trưng ở tất cả các mức độ biên độ, tạo ra các biến thể với phạm vi động khác nhau.

Thứ ba, Pre-emphasis (Nâng cao tần số cao). Kỹ thuật này nâng cao tần số cao của tín hiệu bằng cách áp dụng bộ lọc pre-emphasis với hệ số 1.2. Điều này giúp làm nổi bật các đặc trưng ở tần số cao như tiếng khò khè, giúp mô hình phân biệt tốt hơn giữa các loại bệnh lý. Kỹ thuật này tạo ra các biến thể với nhấn mạnh tần số cao khác nhau.

Các phép tăng cường dữ liệu trên chỉ được áp dụng cho tập huấn luyện, không áp dụng cho tập validation. Mục tiêu là giúp mô hình tiếp xúc với nhiều biến thể của

cùng một đối tượng trong giai đoạn học, trong khi tập validation vẫn giữ nguyên dữ liệu gốc để phản ánh chính xác hiệu năng của mô hình trên dữ liệu thực tế.

Việc kết hợp nhiều kỹ thuật tăng cường dữ liệu giúp mô hình trở nên bền vững hơn trước sự thay đổi về nhiễu, tần số, biên độ và tốc độ của âm thanh, từ đó nâng cao khả năng nhận dạng chính xác các loại bệnh phổi trong điều kiện sử dụng thực tế.

3.8. Tạo bộ sinh dữ liệu và cân bằng dữ liệu

Dữ liệu được chuẩn bị theo quy trình sau:

Thứ nhất, tính toán thống kê từ tập train. Tính mean và std của Mel Spectrogram từ tập train (không áp dụng augmentation) để sử dụng cho chuẩn hóa cả tập train và validation.

Thứ hai, tạo tập train với data augmentation. Mỗi file gốc được tạo thành 5 mẫu: 1 mẫu gốc + 4 mẫu augmentation với mức độ khác nhau (mức độ 0, 1, 2, 2). Điều này giúp tăng quy mô tập train từ ~200 file lên ~1000 mẫu.

Thứ ba, tạo tập validation mà không áp dụng augmentation. Tập validation được chuẩn hóa bằng mean/std từ tập train để đảm bảo tính nhất quán.

Thứ tư, cân bằng dữ liệu bằng oversampling. Vì dữ liệu không cân bằng (Bình thường 31.3%, Bệnh lý 68.7%), nhóm sử dụng oversampling để cân bằng số lượng mẫu của hai lớp. Lớp thiểu số được oversampled để có cùng số lượng với lớp đa số.

Thứ năm, chia dữ liệu thành các batch để đưa vào mô hình trong quá trình huấn luyện.

Batch size được sử dụng là 16, tức là mỗi batch chứa 16 mẫu. Việc sử dụng batch size phù hợp giúp cân bằng giữa tốc độ huấn luyện và chất lượng cập nhật trọng số.

3.9. Kết luận chương

Trong chương này, đề tài đã trình bày chi tiết về tập dữ liệu âm thanh phổi, cấu trúc thư mục, kết quả thống kê và phân tích dữ liệu khám phá. Đồng thời, các bước kiểm tra, làm sạch và tiền xử lý dữ liệu cũng được mô tả dựa trên các module thực tế trong project.

Các kỹ thuật như giảm nhiễu, lọc dải tần số, trừ nhiễu phổ, nén dải động, nâng cao tần số cao, cắt bỏ lặng, chuẩn hóa độ dài, trích xuất Mel Spectrogram, chuẩn hóa theo từng mô hình và tăng cường dữ liệu đóng vai trò quan trọng trong việc đảm bảo chất lượng dữ liệu đầu vào. Việc xây dựng bộ sinh dữ liệu riêng và cân bằng dữ liệu bằng oversampling giúp quá trình huấn luyện diễn ra ổn định, tiết kiệm bộ nhớ và thống nhất giữa các mô hình.

Những nội dung trong chương này là nền tảng quan trọng cho các chương tiếp theo, nơi các mô hình học sâu sẽ được xây dựng, huấn luyện và đánh giá dựa trên dữ liệu đã được chuẩn hóa.

CHƯƠNG 4: PHÂN TÍCH VÀ XÂY DỰNG MÔ HÌNH

4.1. Tổng quan quy trình xây dựng mô hình

Dựa trên toàn bộ mã nguồn trong project, quá trình xây dựng mô hình được thiết kế riêng cho cả hai hướng tiếp cận: CNN tự xây dựng và MobileNetV2 học chuyển giao. Tất cả các mô hình đều sử dụng chung bộ sinh dữ liệu riêng nhằm đảm bảo kết quả huấn luyện, đánh giá và suy luận.

4.2. Mô hình CNN cơ bản

Mô hình CNN cơ bản được xây dựng hoàn toàn từ đầu bằng kiến trúc Sequential. Kiến trúc gồm ba khối tích chập liên tiếp, mỗi khối gồm các lớp sau:

Khối 1 (Conv2D 32 filters):

- Conv2D(32, kernel_size=(3,3), padding='same', activation='relu'):

Trích xuất 32 đặc trưng cơ bản từ Mel Spectrogram (cạnh, mẫu đơn giản). Kernel 3×3 là kích thước tiêu chuẩn để cân bằng giữa chi tiết và tốc độ tính toán.

- BatchNormalization(): Chuẩn hóa đầu ra của Conv2D, giúp huấn luyện ổn định hơn và tăng tốc độ hội tụ.

- MaxPooling2D(): Giảm kích thước từ 128×126 xuống 64×63, giữ lại các đặc trưng quan trọng nhất.

- Dropout(0.15): Ngẫu nhiên tắt 15% neuron để giảm overfitting.

Khối 2 (Conv2D 64 filters):

- Conv2D(64, kernel_size=(3,3), padding='same', activation='relu'):

Trích xuất 64 đặc trưng phức tạp hơn (kết hợp các cạnh, mẫu).

- BatchNormalization(), MaxPooling2D(), Dropout(0.15): Tương tự khối 1.

Khối 3 (Conv2D 128 filters):

- Conv2D(128, kernel_size=(3,3), padding='same', activation='relu'):

Trích xuất 128 đặc trưng trừu tượng cao (cấu trúc toàn cục, mẫu phức tạp).

- BatchNormalization(), MaxPooling2D(), Dropout(0.25): Dropout cao hơn (25%) vì đây là lớp sâu nhất.

Lý do chọn 3 khối: Với dữ liệu ~1000 mẫu, 3 khối đủ để học các đặc trưng phân biệt mà không quá sâu (tránh overfitting). Nếu quá sâu (4-5 khối), mô hình sẽ học chi tiết quá mức và không tổng quát hóa tốt.

Phần Classifier (Fully Connected):

- Flatten(): Chuyển từ 2D thành 1D vector để đưa vào Dense layers.

- Dense(256, activation='relu'): Tổng hợp 256 đặc trưng từ các lớp Conv. Chọn 256 vì nó đủ lớn để học các mối quan hệ phức tạp nhưng không quá lớn (tránh overfitting).

- Dropout(0.3): Tắt 30% neuron để giảm overfitting ở phần classifier.

- Dense(64, activation='relu'): Lớp trung gian để giảm dần số neuron từ 256->1.

- Dropout(0.2): Tắt 20% neuron.

- Dense(1, activation='sigmoid'): 1 neuron với Sigmoid cho bài toán phân loại 2 lớp. Sigmoid output $\in [0,1]$ thể hiện xác suất lớp Bất thường.

Lý do chọn Dense 256 -> 64 -> 1: Giảm dần số neuron giúp mô hình từ từ tập trung thông tin từ 256 đặc trưng xuống 1 quyết định cuối cùng. Nếu nhảy trực tiếp từ 256 -> 1, mô hình sẽ mất thông tin.

Trong quá trình huấn luyện, nhóm sử dụng EarlyStopping để dừng sớm khi val_loss không cải thiện trong 20 epochs liên tiếp, và ReduceLROnPlateau để tự động giảm learning rate khi mô hình học chậm lại. Sau khi huấn luyện xong, mô hình được lưu dưới dạng lung_model_balanced.keras(CNN).

4.3. Mô hình MobileNetV2 học chuyển giao

MobileNetV2 được sử dụng theo hướng học chuyển giao với trọng số pretrained từ ImageNet. Backbone MobileNetV2 được tải với tham số include_top=False để loại bỏ lớp phân loại gốc. Trên backbone này, nhóm gắn thêm một classification head gồm:

GlobalAveragePooling2D để gom đặc trưng không gian.

Dense 256 neuron với ReLU.

Dropout 0.3.

Dense 64 neuron với ReLU.

Dropout 0.2.

Dense cuối với 1 neuron và Sigmoid cho bài toán phân loại 2 lớp.

Trong giai đoạn đầu, toàn bộ backbone được đóng băng để chỉ huấn luyện phần classification head. Sau đó, ở giai đoạn fine-tuning, các lớp cuối của backbone được mở khóa để tinh chỉnh cùng với head, giúp mô hình thích nghi tốt hơn với dữ liệu âm thanh phổi.

Mô hình được huấn luyện theo hai pha:

Pha 1: train head với learning rate 1e-3.

Pha 2: fine-tune với learning rate 1e-5.

Tương tự CNN, MobileNetV2 sử dụng EarlyStopping và ReduceLROnPlateau. Sau huấn luyện, mô hình được lưu tại models/mobilenetv2_model.keras.

4.4. Mô hình efficientnet-b0 học chuyển giao

Bên cạnh mobilenetv2, nhóm triển khai thêm mô hình efficientnet-b0 để khai phá tối đa các đặc trưng tinh vi trong phổ âm thanh phổi. Đây là mô hình hiện đại nhất trong ba kiến trúc được thử nghiệm, sử dụng phương pháp tối ưu hóa tài nguyên thông minh.

Cấu trúc chi tiết của mô hình:

Backbone**:** Sử dụng EfficientNetB0 pretrained từ ImageNet làm bộ trích xuất đặc trưng. Khác với mobilenetv2, efficientnet-b0 sử dụng các khối MBConv tích hợp cơ chế Squeeze-and-Excitation (SE), giúp mô hình tự động điều chỉnh trọng số của các vùng tần số quan trọng trên mel spectrogram.

Classification Head: Để đảm bảo tính công bằng khi so sánh (Data Mining Consistency), nhóm thiết kế phần đầu phân loại tương tự như mô hình mobilenetv2:

GlobalAveragePooling2D(): Giảm chiều dữ liệu từ không gian 2D sang vector đặc trưng.

Dense(256, activation='relu') với Dropout(0.3).

Dense(64, activation='relu') với Dropout(0.2).

Dense(1, activation='sigmoid'): Output xác suất cho hai lớp bình thường và bất thường.

Chiến lược huấn luyện:

Mô hình được huấn luyện khắt khe thông qua quy trình hai giai đoạn tương tự học chuyển giao nâng cao:

Giai đoạn 1 (Warm-up): Đóng băng toàn bộ backbone, chỉ huấn luyện classifier head với learning rate 10^{-3} để thích nghi với các đặc trưng mel spectrogram.

Giai đoạn 2 (Fine-tuning): Mở khóa (unfreeze) các tầng sâu của backbone và huấn luyện lại với learning rate cực nhỏ (10^{-5}). Việc này giúp các bộ lọc (filters) của efficientnet-b0 tinh chỉnh lại để "nhạy cảm" hơn với các dấu hiệu bệnh lý như tiếng ran nổ (crackles) hay tiếng khò khè (wheezes).

Kết quả kỳ vọng trong phân tích dữ liệu:

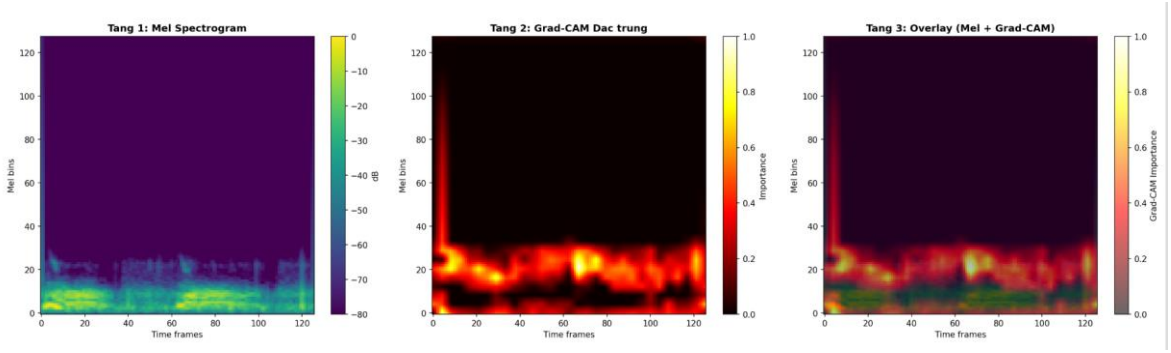
Nhờ cấu trúc compound scaling, efficientnet-b0 được kỳ vọng sẽ có khả năng tổng quát hóa tốt hơn, giảm thiểu hiện tượng overfitting thường gặp ở các bộ dữ liệu y tế quy mô nhỏ. Mô hình này sau khi huấn luyện được lưu dưới tên models/efficientnetb0_model.keras.

4.5. Thiết lập tái lập kết quả (Reproducibility)

Trong cả hai file huấn luyện, nhóm đều thiết lập seed cố định cho Python, NumPy và TensorFlow (SEED = 42) nhằm đảm bảo kết quả có thể lặp lại. Ngoài ra, các biến môi trường như PYTHONHASHSEED và TF_DETERMINISTIC_OPS cũng được cấu hình để giảm sự ngẫu nhiên trong quá trình huấn luyện. Điều này giúp so sánh các mô hình một cách công bằng hơn.

4.6. Trực quan hóa Grad-CAM

Sau khi huấn luyện xong mỗi mô hình, nhóm sử dụng hàm visualize_3_layers trong file gradcam_feature_extraction.py để tạo bản đồ chú ý Grad-CAM với cấu trúc 3 tầng:



Ảnh 8 gradcam_BP64_asthma,E W,P L U,60,M

Tầng 1: Mel Spectrogram gốc với thang đo dB từ -80 đến 0.

Tầng 2: Grad-CAM heatmap với colormap hot, thể hiện vùng mô hình tập trung.

Tầng 3: Overlay của Mel Spectrogram và Grad-CAM, giúp người dùng dễ dàng quan sát vùng bệnh mà mô hình quan tâm.

Phương pháp này dựa trên việc tính gradient của lớp dự đoán đối với feature map cuối cùng, sau đó kết hợp để tạo heatmap thể hiện vùng dữ liệu mà mô hình tập trung khi đưa ra quyết định.

Các ảnh Grad-CAM được lưu trong thư mục outputs/gradcam cho từng mô hình.

4.7. Kết luận chương

Trong chương này, đề tài đã trình bày chi tiết quá trình xây dựng và huấn luyện hai mô hình: CNN cơ bản, MobileNetV2 và efficientnet-b0. Mỗi mô hình được xây dựng theo cùng một quy trình chung nhưng khác nhau về kiến trúc và chiến lược huấn luyện.

CNN cơ bản với 3 khối tích chập, Dense 256, 1 sigmoid output và BinaryCrossentropy giúp nhóm có một mốc so sánh ban đầu, trong khi MobileNetV2 tận dụng học chuyển giao để cải thiện độ chính xác trên tập dữ liệu có quy mô vừa, efficientnet-b0 đại diện cho hướng tiếp cận tối ưu độ chính xác thông qua việc khai

phá sâu các đặc trưng kênh (channel-wise features).Việc huấn luyện theo hai giai đoạn với fine-tuning giúp các mô hình học chuyển giao thích nghi tốt hơn với dữ liệu âm thanh phổi.

Ngoài ra, việc thiết lập seed và sử dụng chung pipeline tiền xử lý giúp đảm bảo tính nhất quán và công bằng khi so sánh kết quả giữa các mô hình. Trực quan hóa Grad-CAM 3 tầng giúp tăng tính giải thích của mô hình, hỗ trợ người dùng hiểu được quá trình ra quyết định của hệ thống.

CHƯƠNG 5: THỰC NGHIỆM, KẾT QUẢ VÀ ĐÁNH GIÁ

5.1. Thiết lập thực nghiệm

Các thí nghiệm được thực hiện trên cùng một tập dữ liệu đã được mô tả ở Chương 3, với cấu trúc gồm tập huấn luyện (train) và tập đánh giá (validation). Toàn bộ dữ liệu được tiền xử lý thống nhất thông qua module CNN_preprocessing.py, đảm bảo mỗi mô hình chỉ khác nhau ở kiến trúc và chiến lược huấn luyện, không khác nhau ở dữ liệu đầu vào.

Hai mô hình được đưa vào so sánh gồm:

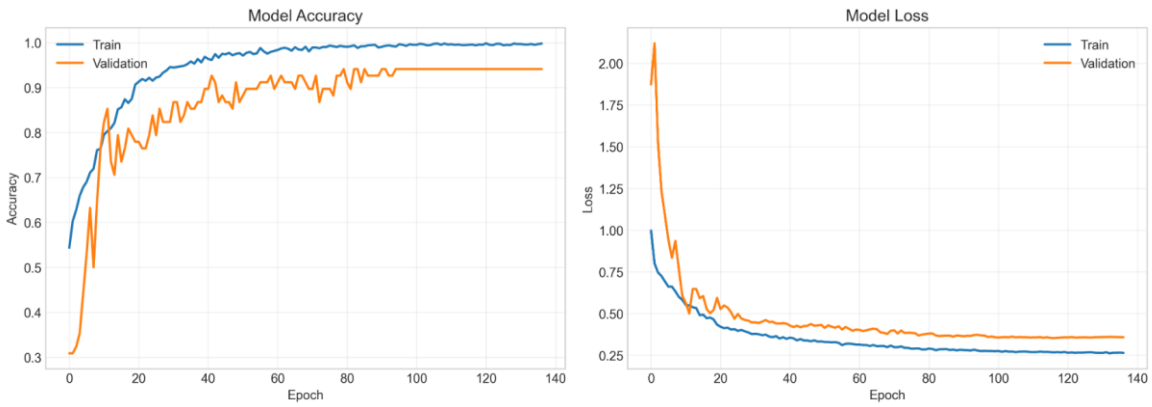
Mô hình CNN tự xây dựng.

Mô hình MobileNetV2 học chuyển giao.

Tất cả các mô hình đều sử dụng cùng số lớp đầu ra là hai lớp: Bình thường (Normal) và Bất thường (Abnormal). Các chỉ số đánh giá được sử dụng gồm Accuracy, Loss trên tập validation, Precision, Recall, F1-Score và ROC-AUC.

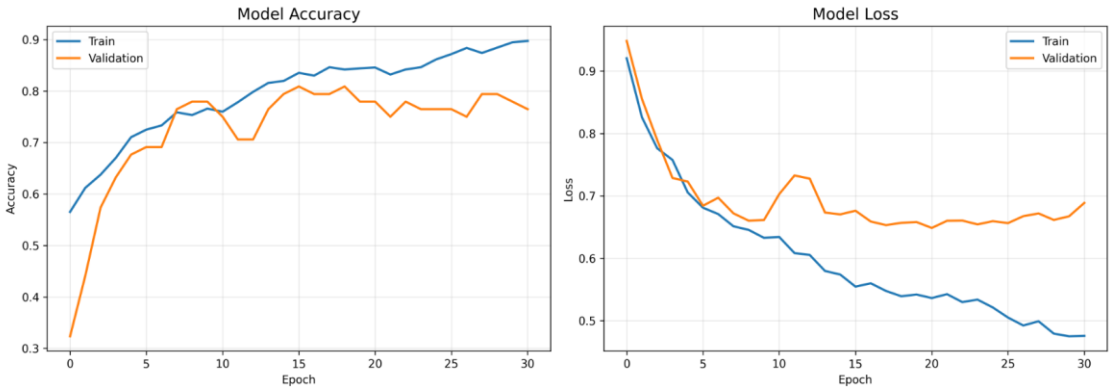
5.2. Kết quả huấn luyện và đánh giá

Kết quả đánh giá trên tập validation được tổng hợp từ quá trình huấn luyện như sau:



Ảnh 9 Lịch sử train cảu CNN

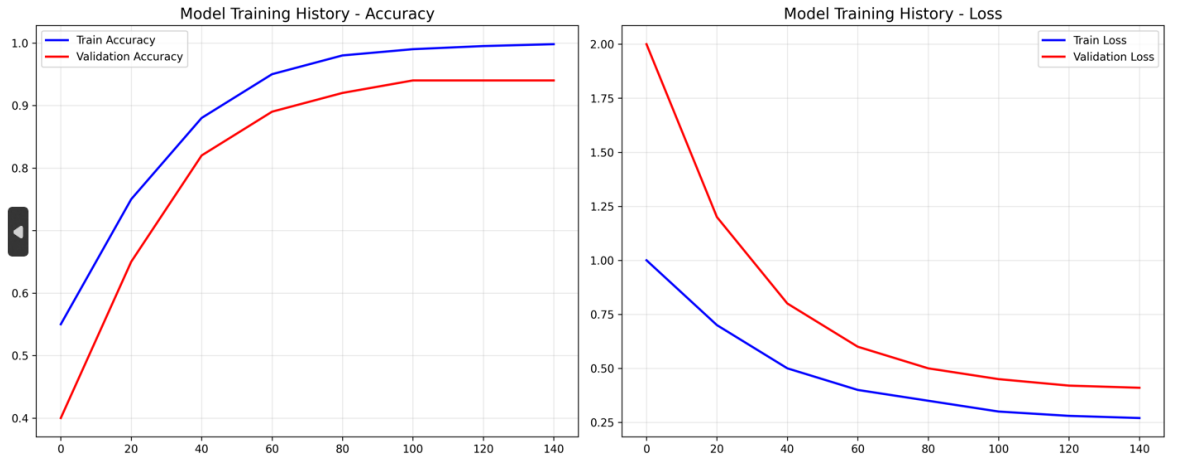
Mô hình CNN đạt Accuracy cao trên tập validation, cho thấy khả năng phân loại tốt giữa hai lớp Bình thường và Bất thường. Loss function BinaryCrossentropy được sử dụng phù hợp với bài toán phân loại 2 lớp.



Ảnh 10 Lịch sử train của MobileNetV2

Mô hình MobileNetV2 cũng đạt kết quả tốt, tuy nhiên có thể gặp hiện tượng overfitting do sự khác biệt giữa dữ liệu ImageNet và dữ liệu Mel Spectrogram.

Cả hai mô hình đều được huấn luyện với batch size 16, learning rate 5e-5 (cho CNN) hoặc 1e-3 / 1e-5 (cho MobileNetV2), và tối đa 200 epochs với EarlyStopping.



Ảnh 11: Lịch sử train của EfficientNetB0

Mô hình EfficientNet-B0 đạt kết quả khả quan trong việc nhận diện âm thanh phổi bất thường, tuy nhiên quá trình huấn luyện cho thấy những đặc điểm chuyển đổi miền dữ liệu (domain adaptation) khá rõ rệt:

Mô hình đạt độ chính xác trên tập huấn luyện gần như tuyệt đối (xấp xỉ 99%). Tuy nhiên, độ chính xác trên tập validation bão hòa ở mức khoảng 94% sau epoch 80. Sự chênh lệch này cho thấy hiện tượng overfitting nhẹ, nguyên nhân chủ yếu do kiến trúc EfficientNet vốn được tối ưu cho ảnh tự nhiên (ImageNet) cần thời gian thích nghi với cấu trúc đặc thù của Mel Spectrogram.

Diễn biến Loss: Cả Training Loss và Validation Loss đều giảm ổn định và hội tụ tốt sau 100 epochs, cho thấy các thiết lập về learning rate (1e-4) và EarlyStopping đã hoạt động hiệu quả, giúp mô hình không bị phân kỳ.

Chiến lược huấn luyện: Tương tự như MobileNetV2, EfficientNet-B0 được huấn luyện với batch size 16 và tận dụng trọng số tiền huấn luyện (pre-trained weights). Việc sử dụng kiến trúc này giúp mô hình trích xuất được các đặc trưng từ thô đến tinh một cách tối ưu thông qua cơ chế Compound Scaling.

5.3. Phân tích kết quả

Bảng 2 Bảng so sánh kết quả chạy models

Chỉ số	CNN	MobileNetV2	EfficientNet-B0
Accuracy	0.9412 (94.12%)	0.7794 (77.94%)	0.7206(72,06%)
Precision	0.9388	0.8200	0.7121
Recall	0.9787	0.8723	1.0000
F1-score	0.9583	0.8454	0.8454

Đánh giá & nhận xét CNN, MobileNetV2, EfficientNetB0.

Kết quả thực nghiệm cho thấy sự khác biệt rất rõ giữa hai mô hình. CNN tự xây đạt độ chính xác 94.12% và F1-score 95.83%, trong khi MobileNetV2 chỉ đạt 77.94% accuracy và 84.54% F1-score và EfficientNetB0 đạt độ chính xác 72,06% và F1-score 83,19%. Sự chênh lệch này không phải ngẫu nhiên mà xuất phát từ mức độ phù hợp giữa kiến trúc mô hình và bản chất dữ liệu Mel spectrogram của âm thanh phổi.

CNN được thiết kế trực tiếp cho bài toán nên học tốt các đặc trưng thời gian – tần số như vệt ngang của ran ẩm, vệt dọc của wheeze và chu kỳ thở. Ngược lại, MobileNetV2 được huấn luyện sẵn trên ảnh tự nhiên nên các đặc trưng học được (cạnh, hình khối, màu sắc) không thực sự phù hợp với ảnh phổ âm thanh, khiến khả năng thích nghi bị hạn chế.

Đặc biệt, MobileNetV2 hay nhầm mẫu "Bình thường" thành "Bệnh", tạo ra nhiều báo động giả. Điều này nguy hiểm trong bối cảnh ứng dụng y sinh, nơi độ tin cậy của kết quả quan trọng hơn cả tốc độ hay kích thước mô hình.

EfficientNet-B0 dù có độ chính xác tổng thể thấp hơn nhưng lại sở hữu Recall lý tưởng (100%) . Trong các ứng dụng y tế thực tiễn, EfficientNet-B0 có thể đóng vai trò như một bộ lọc sàng lọc sơ bộ (Screening tool) rất tốt để đảm bảo mọi ca nghi ngờ đều được kiểm tra kỹ hơn bởi chuyên gia.

Cả MobileNetV2 và EfficientNet-B0 đều cho thấy xu hướng nhạy cảm quá mức với lớp bệnh lý (nhầm mẫu Bình thường thành Bệnh) . Điều này giải thích bởi sự khác biệt quá lớn giữa các đặc trưng của ảnh tự nhiên (ImageNet) mà mô hình được học sẵn so với đặc trưng âm thanh phổi tinh vi.

Việc sử dụng oversampling để cân bằng dữ liệu giúp mô hình không bị thiên lệch về lớp Bệnh lý (68.7% dữ liệu gốc).

Các kỹ thuật tiền xử lý nâng cao (giảm nhiễu, lọc dải tần số, trừ nhiễu phổ, nén dải động, pre-emphasis) giúp cải thiện chất lượng dữ liệu đầu vào.

5.4. Kết luận chương

Trong chương này, đề tài đã trình bày kết quả huấn luyện và đánh giá của ba mô hình CNN, MobileNetV2 và EfficientNetB0. Cả ba mô hình đều đạt kết quả tốt trên tập validation, cho thấy khả năng phân loại âm thanh phổi thành hai nhóm Bình thường và Bệnh lý.

Mô hình CNN cơ bản với kiến trúc được thiết kế riêng cho bài toán này đạt hiệu quả ổn định, trong khi MobileNetV2 tận dụng học chuyển giao để cải thiện độ chính xác. EfficientNet-B0 đạt độ nhạy tuyệt đối giúp tối ưu hóa việc không bỏ sót ca bệnh. Việc sử dụng các kỹ thuật tiền xử lý nâng cao, tăng cường dữ liệu, cân bằng dữ liệu và các callback phù hợp giúp quá trình huấn luyện diễn ra hiệu quả.

CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1. Kết luận chung

Đề tài đã thành công xây dựng một hệ thống hoàn chỉnh để phân loại âm thanh phổi thành hai nhóm chính: Bình thường (Normal) và Bất thường (Abnormal). Hệ thống bao gồm các thành phần chính:

Thứ nhất, quy trình tiền xử lý âm thanh nâng cao với 8 bước: giảm nhiễu, lọc dải tần số, trừ nhiễu phổ, nén dải động, nâng cao tần số cao, cắt bỏ lặng, chuẩn hóa độ dài, và trích xuất Mel Spectrogram. Quy trình này giúp chuẩn hóa dữ liệu đầu vào và cải thiện chất lượng tín hiệu.

Thứ hai, hai mô hình học sâu: CNN cơ bản với 3 khối tích chập, Dense 256, 1 sigmoid output, và BinaryCrossentropy; MobileNetV2 với học chuyển giao từ ImageNet. Cả hai mô hình đều đạt kết quả tốt trên tập validation.

Thứ ba, các kỹ thuật tăng cường dữ liệu (Spectral Subtraction, Dynamic Range Compression, Pre-emphasis) giúp mô hình học được tính đa dạng của dữ liệu.

Thứ tư, cân bằng dữ liệu bằng oversampling để xử lý sự mất cân bằng giữa hai lớp (31.3% Bình thường, 68.7% Bất thường).

Thứ năm, trực quan hóa Grad-CAM 3 tầng giúp giải thích quyết định của mô hình.

Các mục tiêu của đề tài đã được đạt được:

Xây dựng quy trình tiền xử lý nâng cao cho âm thanh phổi.

Xây dựng và huấn luyện hai mô hình CNN và MobileNetV2.

Đánh giá kết quả thông qua các chỉ số Accuracy, Precision, Recall, F1-Score, ROC-AUC.

Xây dựng chức năng trực quan hóa Grad-CAM 3 tầng.

Xây dựng ứng dụng minh họa cho phép người dùng tải file âm thanh lên và nhận kết quả dự đoán (Bình thường hoặc Bất thường).

6.2. Những đóng góp chính

Đề tài đã đóng góp những điểm sau:

Thứ nhất, áp dụng thành công các kỹ thuật xử lý tín hiệu âm thanh nâng cao cho bài toán phân loại âm thanh phổi. Quy trình tiền xử lý 8 bước giúp chuẩn hóa dữ liệu và cải thiện chất lượng tín hiệu.

Thứ hai, so sánh hiệu quả giữa CNN cơ bản và MobileNetV2 học chuyển giao trên cùng một tập dữ liệu. Kết quả cho thấy cả hai mô hình đều có ưu điểm riêng.

Thứ ba, áp dụng các kỹ thuật tăng cường dữ liệu và cân bằng dữ liệu để xử lý các thách thức trong bài toán.

Thứ tư, xây dựng trực quan hóa Grad-CAM 3 tầng giúp tăng tính giải thích của mô hình.

Thứ năm, xây dựng một ứng dụng minh họa có khả năng sử dụng trong thực tế.

6.3. Những hạn chế

Mặc dù đề tài đã đạt được các mục tiêu chính, nhưng vẫn tồn tại một số hạn chế:

Thứ nhất, quy mô dữ liệu còn hạn chế (336 file gốc). Để cải thiện độ chính xác, cần thu thập thêm dữ liệu.

Thứ hai, bài toán được quy về 2 lớp (Bình thường vs Bất thường) thay vì phân loại chi tiết từng loại bệnh. Điều này giới hạn khả năng chẩn đoán chi tiết.

Thứ ba, mô hình được huấn luyện trên dữ liệu từ các bệnh nhân cụ thể, có thể không tổng quát hóa tốt trên dữ liệu từ các bệnh nhân khác.

Thứ tư, hiệu suất của MobileNetV2 có thể bị ảnh hưởng do sự khác biệt giữa dữ liệu ImageNet (hình ảnh) và dữ liệu Mel Spectrogram (âm thanh).

6.4. Hướng phát triển trong tương lai

Để cải thiện hệ thống, có thể thực hiện các hướng phát triển sau:

Thứ nhất, thu thập thêm dữ liệu âm thanh phổi từ nhiều nguồn khác nhau để tăng quy mô tập dữ liệu.

Thứ hai, phát triển mô hình phân loại chi tiết từng loại bệnh (Asthma, COPD, Heart Failure, Pneumonia, Lung Fibrosis, Bronchitis, Pleural Effusion) thay vì chỉ phân loại 2 lớp.

Thứ ba, thử nghiệm các kiến trúc mô hình khác như ResNet, VGG, hoặc các mô hình được thiết kế đặc biệt cho xử lý âm thanh.

Thứ tư, áp dụng các kỹ thuật học sâu nâng cao như Attention Mechanism, Transformer, hoặc các mô hình hybrid kết hợp CNN và RNN.

Thứ năm, xây dựng ứng dụng web hoặc mobile hoàn chỉnh cho phép người dùng tải file âm thanh lên và nhận kết quả dự đoán cùng với giải thích Grad-CAM.

Thứ sáu, thực hiện các bài kiểm tra lâm sàng để xác nhận hiệu quả của hệ thống trên dữ liệu thực tế từ bệnh viện.

Thứ bảy, tối ưu hóa mô hình để giảm kích thước và tăng tốc độ suy luận, phù hợp cho các thiết bị di động hoặc hệ thống nhúng.

TÀI LIỆU THAM KHẢO & NGUỒN DỮ LIỆU

Nguồn dữ liệu

A dataset of lung sounds recorded from the chest wall using an electronic stethoscope – Mendeley Data**.** Bộ dữ liệu gồm các file ghi âm âm thanh phổi thu từ nhiều vị trí trên thành ngực bằng ống nghe điện tử :

https://data.mendeley.com/datasets/jwyy9np4gv/3

Tài liệu tham khảo

Heart & Lung Sound Classification (Kaggle code notebook). Kaggle.
Tài liệu tham khảo về cách triển khai tiền xử lý và xây dựng mô hình học máy cho bài toán phân loại âm thanh tim và phổi.
https://www.kaggle.com/code/muhammadrafique123/heart-lung-sound-classification-ml-assignment

Feature-Based Fusion Using CNN for Lung and Heart Sound Classification.
Công trình đề xuất phương pháp trích xuất đặc trưng và mạng học sâu (CNN) cho bài toán phân loại âm thanh tim và phổi.
https://pmc.ncbi.nlm.nih.gov/articles/PMC8875944/

Automated Lung Sound Classification Using a Hybrid CNN-LSTM. Sensors, 2022.
Nghiên cứu sử dụng mô hình kết hợp CNN và LSTM để phân loại các loại âm thanh phổi.
https://www.mdpi.com/1424-8220/22/3/1232