

國立臺灣大學共同教育中心統計碩士學位學程



碩士論文

Master Program of Statistics

Center of General Education

National Taiwan University

Master Thesis

利用多模態深度學習模型結合胸部 X 光  
和電子健康紀錄以篩檢急性心臟衰竭

Multimodal Deep Learning Model for Screening Acute  
Heart Failure in Emergency Department Using Chest  
X-rays and Electronic Health Records

吳証恩

Jeng-En Wu

指導教授：周呈霽 博士

Advisor: Cheng-Ying Chou, Ph.D.

中華民國 111 年 8 月

August, 2022



# Acknowledgements

碩班的兩年一眨眼就過了，這兩年是我求學過程中覺得最充實、最開心也是收穫最多的時光。雖然過程中所經歷的跟原先預想的相距甚遠，但人生就是因為不確定才精彩，而統計學正是研究不確定性的科學。

首先我要感謝我的爸爸媽媽，謝謝您們的養育之恩，讓我從小大到可以選擇自己想做的事情，而且一路上都很支持我所做的決定。謝謝您們耐心的聽我分享論文內容，也給出了很多有建設性的建議。未來在工作上我依然會繼續努力，不會讓您們擔心，並且會請你們吃飯出去玩哈哈。也感謝我的哥哥姊姊，你們用自己的方式來鼓勵我（應該有吧？），我有感受到。

謝謝周老師願意在碩二上時收留我，跟您做研究的這十個月以來很願意給我資源，也花了很多時間與心力在我們身上。畢業後可能還需要趕一下投期刊的論文，再請老師多多指教。也謝謝楊老師給我發揮能力的機會，也讓我了解自己想要的什麼。

謝謝呂嘉霖、許苡鈴以及丁立恆，跟你們一起修課、聊天講屁話是碩班開心的主要來源，以後以要一起加油喔喔

我也要感謝我的女朋友，在我碩班期間最憂鬱、最無助的時候不斷鼓勵我，常常聽我的抱怨，更帶給我不可取代的喜樂。未來妳讀研究所時我也會好好承擔妳的情緒

最後我也要感謝自己，在遇到困難時你都勇敢地面對，並努力讓自己越來越強。「擇期所愛、愛其所擇」，碩班兩年面臨到的選擇很多，而人生中的選擇更多。我會成為不怕挑戰、不會後悔、不斷進步的人！





## 摘要

本研究旨在使用結合患者胸腔 X 光片和電子健康紀錄的多模態深度學習模型於急診室進行急性心臟衰竭之篩檢。如果病人 N 端前腦利鈉肽的血液濃度高於 300ng/L，則被定義為陽性。對於每張胸腔 X 光片，肺心遮罩生成器最初通過肺心分割模型識別肺和心臟區域。然後通過預定義的演算法手動提取三個估計心臟大小的比例值和 306 個放射組學特徵。最終，多模態深度學習模型融合來自電子健康紀錄和胸腔 X 光片的資訊並輸出最終預測。研究群體是從公開資料集 Medical Information Mart for Intensive Care (MIMIC) IV 中所提取。研究群體包括 1,432 名患者和 1,833 對胸腔 X 光片和電子健康紀錄。其中，71% 的樣本為陽性。53% 的患者為男性，47% 為女性。研究群體的年齡分佈最小為 20 歲最大為 91 歲，年齡平均為 65 歲。此回溯性實驗在此資料集上顯示，當多模態深度學習模型整合來自每個單模態模型的預測結果時，模型預測表性最高可達 AUROC 值 0.89。

**關鍵字：**多模態、深度學習、急性心臟衰竭、胸腔 X 光、電子健康紀錄





# Abstract

This study aims to screen suspected acute heart failure (AHF) in the emergency department using a multimodal deep learning model combining patients' chest X-rays (CXRs) and electronic health records (EHRs). The binary label for AHF is defined as positive if a patient's value of N terminal pro B type natriuretic peptide (NT-proBNP) is higher than 300 ng/L. For each CXR, the lung-heart mask generator initially identified the lung and heart region by the lung-heart segmentation model. Then three heart-size ratios and 306 radiomic features were extracted manually by predefined formulas. Eventually, the information from EHRs and CXRs was fused to output the final prediction. The study population was extracted from the Medical Information Mart for Intensive Care (MIMIC) IV open-source dataset. The study population includes 1,432 patients and 1,833 pairs of CXR and EHR. 71% of the samples are positive. 53% of the patient are male, and 47% are female. The age of the study population range from 20 to 91, with a mean of 65. The retrospective experiments illustrated that the proposed method achieved the highest

AUROC of 0.89 when fusing all predictions from every single-modality model.

**Keywords:** acute heart failure, chest X-ray, deep learning, electronic health record, multimodality





# Contents

|  | <b>Page</b> |
|--|-------------|
| <b>Acknowledgements</b>  | <b>i</b>    |
| <b>摘要</b>  | <b>iii</b>  |
| <b>Abstract</b>  | <b>v</b>    |
| <b>Contents</b>  | <b>vii</b>  |
| <b>List of Figures</b>   | <b>xi</b>   |
| <b>List of Tables</b>  | <b>xiii</b> |
| <b>List of Algorithms</b>  | <b>xiv</b>  |
| <b>Chapter 1 Introduction</b>  | <b>1</b>    |
| 1.1 Background . . . . .   | 1           |
| 1.1.1 Choosing NT-proBNP as an acute heart failure indicator . . . . . | 3           |
| 1.2 Contributions . . . . .  | 4           |
| <b>Chapter 2 Literature Review</b>                                     | <b>7</b>    |
| 2.1 Predicting Thoracic Diseases Using DL and CXR . . . . .            | 7           |
| 2.2 Predicting HF Using DL and CXR . . . . .                           | 8           |
| 2.3 Multimodal DL Combining Medical Imaging and EHR . . . . .          | 9           |
| <b>Chapter 3 Materials and Methods</b>                                 | <b>13</b>   |
| 3.1 Data Description . . . . .   | 13          |



|                  |   |           |
|------------------|---|-----------|
| 3.1.1            | MIMIC-IV . . . . .  | 13        |
| 3.1.2            | MIMIC-CXR . . . . .   | 14        |
| 3.1.3            | JSRT . . . . .  | 15        |
| 3.1.4            | Manually labeled lung-heart masks for images in MIMIC-CXR . . . . . | 15        |
| 3.2              | Data Pre-processing . . . . .                                       | 17        |
| 3.2.1            | EHR from MIMIC-IV . . . . .   | 17        |
| 3.2.2            | CXR from MIMIC-CXR . . . . .  | 18        |
| 3.2.3            | CXR from JSRT . . . . .   | 19        |
| 3.3              | Frontal-lateral Classifier . . . . .                                | 20        |
| 3.4              | Lung-heart Mask Generator . . . . .                                 | 21        |
| 3.4.1            | Segmentation models . . . . .                                       | 22        |
| 3.4.2            | Mask post-processing . . . . .                                      | 23        |
| 3.5              | Feature Extraction from Lung-Heart Masks . . . . .                  | 26        |
| 3.5.1            | Heart-size ratios . . . . .   | 26        |
| 3.5.2            | Radiomic features . . . . .   | 27        |
| 3.5.3            | Thoracic region of interest . . . . .                               | 28        |
| 3.6              | NT-proBNP Prediction Models . . . . .                               | 28        |
| 3.6.1            | Numerical models . . . . .  | 30        |
| 3.6.2            | Image models . . . . .  | 30        |
| 3.6.3            | Multimodal models . . . . .   | 31        |
| 3.6.4            | Fusion strategy . . . . .   | 31        |
| <b>Chapter 4</b> | <b>Results and Discussions</b>                                      | <b>37</b> |
| 4.1              | Demographics of Study Population . . . . .                          | 37        |



|                  |  |           |
|------------------|--|-----------|
| 4.2              | PA, AP, and Frontal Views . . . . .              | 37        |
| 4.3              | Evaluation of Mask Post-processing . . . . .     | 39        |
| 4.4              | Performance Based on Input Combination . . . . . | 40        |
| 4.5              | Performance Based on Fusion Strategy . . . . .   | 42        |
| 4.6              | Analysis of Image Data . . . . .                 | 43        |
| 4.7              | Analysis of Numerical Data . . . . .             | 44        |
| 4.8              | Analysis of Heart-size Ratios . . . . .          | 48        |
| 4.9              | Analysis of Radiomic Features . . . . .          | 49        |
| <b>Chapter 5</b> | <b>Limitations</b>                               | <b>55</b> |
| 5.1              | Label Uncertainty . . . . .                      | 55        |
| 5.2              | Segmentation Models . . . . .                    | 56        |
| 5.3              | Mask Post-processing Algorithms . . . . .        | 58        |
| 5.4              | Generalization . . . . .                         | 59        |
| <b>Chapter 6</b> | <b>Conclusion</b>                                | <b>61</b> |
|                  | <b>References</b>                                | <b>63</b> |
|                  | <b>Appendix A — Figures</b>                      | <b>69</b> |
| A.1              | Flowchart of Inclusion and Exclusion . . . . .   | 69        |
|                  | <b>Appendix B — Tables</b>                       | <b>71</b> |
| B.1              | Radiomic Features . . . . .                      | 71        |



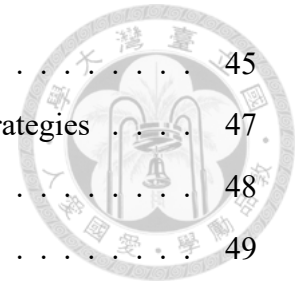




# List of Figures

|      |   |    |
|------|---|----|
| 1.1  | Diagnosis of acute heart failure in an urgent care setting. . . . .       | 2  |
| 3.1  | Examples of images in MIMIC-CXR . . . . .                                 | 14 |
| 3.2  | Examples of images in JSRT . . . . .                                      | 16 |
| 3.3  | Examples of manually labeled images . . . . .                             | 17 |
| 3.4  | Data extraction, data selection, and data merging of MIMIC-IV . . . . .   | 18 |
| 3.5  | Scatter plot of image dimensions . . . . .                                | 19 |
| 3.6  | Examples of images enhanced by CLAHE . . . . .                            | 20 |
| 3.7  | Image processing for JSRT . . . . .                                       | 21 |
| 3.8  | Examples of lung-heart masks . . . . .                                    | 23 |
| 3.9  | Examples of false-positive lung-heart masks . . . . .                     | 24 |
| 3.10 | False negative lung-heart mask . . . . .                                  | 26 |
| 3.11 | Demonstration of heart-size ratios . . . . .                              | 27 |
| 3.12 | Examples of ROI . . . . .   | 29 |
| 3.13 | Image processing pipeline . . . . .                                       | 33 |
| 3.14 | Model architecture – early fusion . . . . .                               | 34 |
| 3.15 | Model architecture – joint fusion . . . . .                               | 34 |
| 3.16 | Model architecture – late fusion . . . . .                                | 35 |
| 4.1  | Examples of PA-view and AP-view CXRs . . . . .                            | 40 |
| 4.2  | Box plots of AUROC of different CXR views . . . . .                       | 41 |
| 4.3  | Mask post-processing . . . . .  | 42 |
| 4.4  | Box plots of AUROC of different mask post-processing algorithms . . . . . | 43 |
| 4.5  | Box plots of AUROC of different input combinations . . . . .              | 44 |

|      |  |    |
|------|--|----|
| 4.6  | Box plots of AUROC of different fusion strategies . . . . .                  | 45 |
| 4.7  | Box plots of different modality combinations and fusion strategies . . . . . | 47 |
| 4.8  | Examples of heatmaps . . . . .   | 48 |
| 4.9  | Feature importance (gain) . . . . .  | 49 |
| 4.10 | SHAP bar plot . . . . .  | 50 |
| 4.11 | SHAP summary plot . . . . .  | 51 |
| 4.12 | Distributions of heart-size ratios . . . . .                                 | 52 |
| 4.13 | Examples of radiomic features . . . . .                                      | 53 |
| 5.1  | Demonstration of label uncertainty . . . . .                                 | 56 |
| A.1  | Flowchart of inclusion and exclusion . . . . .                               | 69 |





# List of Tables

|     |  |    |
|-----|--|----|
| 2.1 | Summary table of related works . . . . .                       | 11 |
| 4.1 | Demographic characteristics of the study population . . . . .  | 38 |
| 4.2 | Demographic characteristics of populations . . . . .           | 39 |
| 4.3 | Model performances of different views of CXR . . . . .         | 39 |
| 4.4 | AUROC of models using different mask post-processing . . . . . | 40 |
| 4.5 | Model performances . . . . .                                   | 46 |





# List of Algorithms

|   |  |    |
|---|--|----|
| 1 | Mask post-processing (I): delete fragments . . . . . | 24 |
| 2 | Mask post-processing (II): align and flip . . . . .  | 25 |







# Chapter 1 Introduction

## 1.1 Background

Acute heart failure (AHF) is defined as a rapid change in the signs and symptoms of heart failure (HF) that result in the need for urgent treatment [1]. Diagnoses of AHF account for approximately one million emergency department (ED) visits in the United States [2]. Despite its high prevalence in the ED setting, diagnosing AHF in ED patients with undifferentiated dyspnea can be challenging, especially in patients with advanced age and comorbid disease [3]. Currently, no single examination can on its own reliably diagnose or rule out AHF [3]. On the other hand, several studies have shown that combining routine history, clinical tests, medical images (for example, chest radiography), and measurement of plasma natriuretic peptide levels improves diagnostic accuracy. The improvement makes the biomarkers recommended in international guidelines for diagnosing and managing AHF. As shown in Fig. 1.1, the current diagnosis workflow of AHF requires several processes. Once patients enter ED with suspected AHF, they will take an initial workup such as medical history, physical examination, electrocardiogram (ECG), or chest X-ray (CXR). If the patient's condition remains uncertain, a B-type natriuretic peptide (BNP) or N terminal pro B type natriuretic peptide (NT-proBNP) test will be taken. Patients with  $BNP < 100$  ng/L or  $NT\text{-}proBNP < 300$  ng/L are AHF unlikely and should

consider other diagnoses. Patients with abnormal levels of BNP or NT-proBNP should take further examinations like echocardiography to confirm the disease.

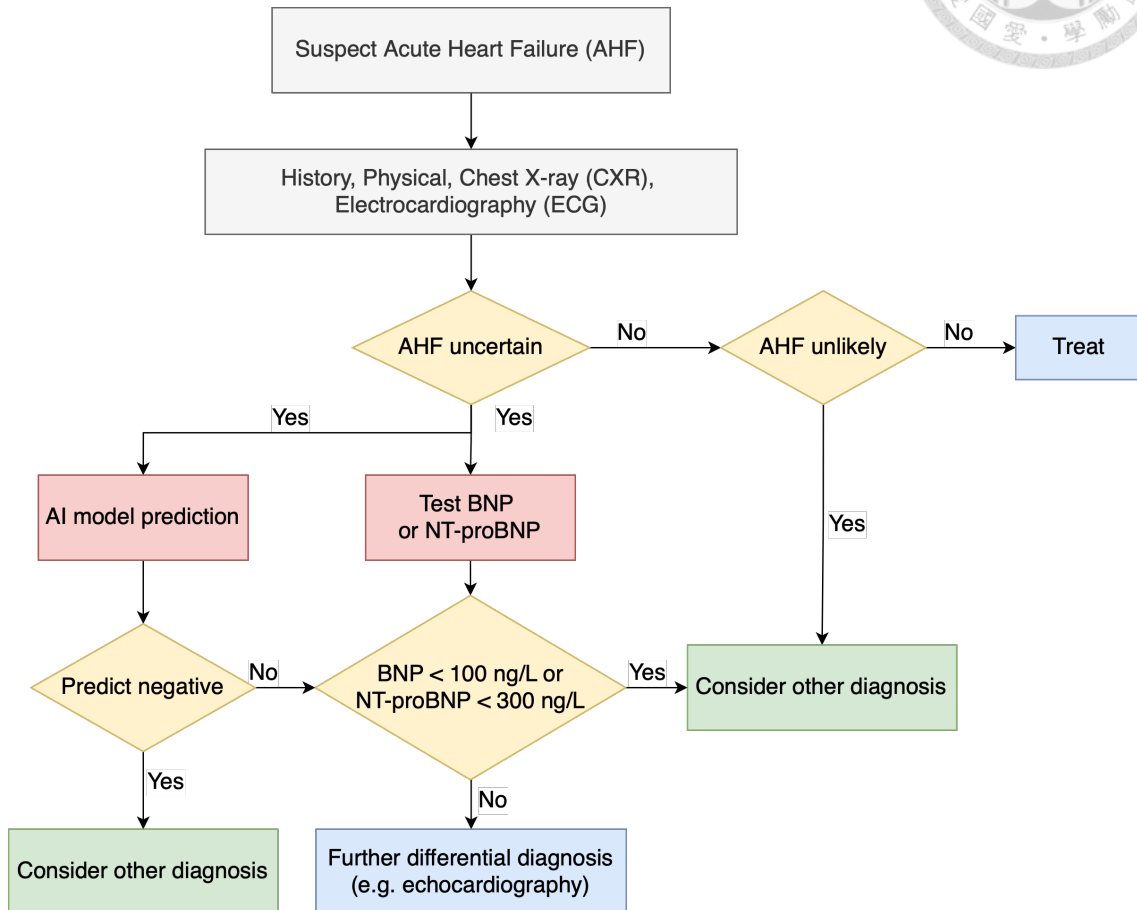
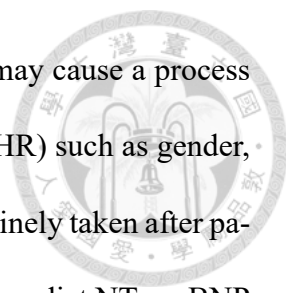


Figure 1.1: Diagnosis of acute heart failure in an urgent care setting.

At the recommended thresholds of 100 ng/L for BNP and 300 ng/L for NT-proBNP, the biomarkers have sensitivities of 0.95 (95% confidence interval 0.93~0.96) and 0.99 (0.97~1.00) and negative predictive values of 0.94 (0.90~0.96) and 0.98 (0.89~1.0), respectively, for a diagnosis of AHF [4]. In addition, according to the 2022 AHA/ACC/HFSA Guideline [5], BNP or NT-proBNP is recommended as a valuable biomarker to rule out AHF.

Despite the remarkable effectiveness of the NT-proBNP test, the average time between taking samples and obtaining results is around 1~2 hours. In acute care settings,

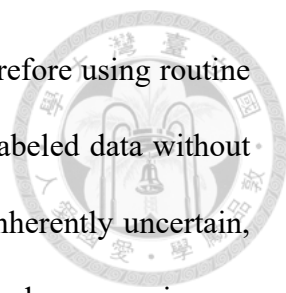


taking an NT-proBNP test is expensive and time-consuming, which may cause a process bottleneck in ED. By contrast, CXR and electronic health records (EHR) such as gender, age, temperature, heart rate, etc., are inexpensive, immediate, and routinely taken after patients enter to ED. In this sense, a deep learning (DL) model that can predict NT-proBNP levels using CXR and EHR can release the burden due to its high-speed performance. Therefore, this study aims to develop a multimodal deep learning model to predict whether the patients' NT-proBNP  $> 300$  ng/L or not using their CXR and EHR in ED.

### **1.1.1 Choosing NT-proBNP as an acute heart failure indicator**

Several studies have used the DL technique and CXR to predict HF but with different ground truths. Seah *et al.* [6] defined HF as BNP  $> 100$  ng/L. Their work used frontal radiographs to develop a generative DL model and visualized neural network learning of chest radiograph features in congestive HF using generative visual rationales. The result illustrated the possibility of generating medical images given various levels of BNP. Matsumoto *et al.* [7] utilized CXR findings "cardiomegaly or congestion" as the ground truth of HF. The ground truths were verified and relabeled by two cardiologists. Their classification model extracted image features from CXR using pre-trained VGG16 and predicted the final results with a trainable multi-layer perceptron. Jabbour *et al.* [8] predicted HF using frontal CXR and patients' EHR, including NT-proBNP. The ground truth of HF in their setting was defined by the International Classification of Diseases 10th (ICD-10).

Studies above show the success of predicting HF by different indicators while using clinical measurements such as NT-proBNP as predictive targets have extra advantages. Training supervised deep learning models in hospitals typically requires a large amount of labeled data manually annotated by physicians. This process may be time and

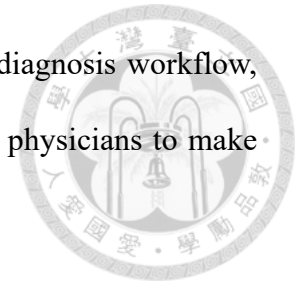


labor-consuming and bring extra loading to the clinical system. Therefore using routine biomarkers as training labels provides a sustainable way to collect labeled data without additional manual annotation. Moreover, the diagnosis of AHF is inherently uncertain, and the results may vary among physicians based on their backgrounds or experiences. Human biases may exist using CXR findings from reports or ICD codes as targets in this case. NT-proBNP is an examination executed in a laboratory, and the test results are more consistent even from different hospitals. In this study, NT-proBNP was used as the prediction target and categorized patients into AHF likely and AHF unlikely by a cut-off at 300 ng/L.

## 1.2 Contributions

To the best of my knowledge, this is the first study that predicts the level of NT-proBNP using CXR and EHR in the acute care setting, showing the possibility of implementing the DL model in real-world clinical situations. I also proposed a novel post-processing algorithm, which approximates the lung masks even when the contours of the lungs are undetectable. Utilizing lung-heart masks, both domain-based (heart-size ratios, radiomic features) and data-driven features were extracted from CXR to boost the performance. Using the human-recognizable features, I interpreted the relationship between image features and the risk of AHF, which numerically verifies the clinical understanding of AHF. Intensive comparisons also were conducted to demonstrate the contributions of different input combinations and the effectiveness of fusion strategies. The comparison results can be considered as a benchmark for future research. The best model achieves an AUROC of 0.8861 for predicting NT-proBNP > 300 ng/L for ED patients using PA view CXR and EHR. The target scenario is the most common use scenario for NT-proBNP,

and our proposed AI model can be embedded in the current AHF diagnosis workflow, as shown in Fig. 1.1. As a rule-out marker, the AI model can help physicians to make decisions in urgent situations.







## Chapter 2 Literature Review

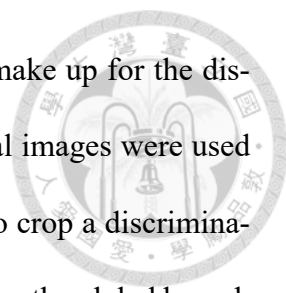
### 2.1 Predicting Thoracic Diseases Using DL and CXR

CheXNet [9], created by Rajpurkar *et al.*, is a 121-layer convolutional neural network trained on ChestX-ray14, which contains over 100,000 frontal-view X-ray images with 14 diseases. CheXNet correctly diagnosed all 14 illnesses in ChestX-ray14, with an overall average AUROC of 0.841.

Deep convolutional neural networks are used in the segmentation-based deep fusion network (SDFN) [10] to automatically detect thoracic illnesses shown in the CXR pictures. The higher-resolution data of specific local lung areas and domain knowledge were combined to propose a unique approach by Liu *et al.* The lung region generator specifically detected and cropped the local lung areas. The features of the whole CXR pictures and the cropped lung area images were then extracted using two CNN-based classification models. Finally, the feature fusion module for disease classification fuses the collected features. On the ChestX-ray14 Dataset, evaluated by the NIH benchmark split, SDFN obtained an average AUROC of 0.815.

A three-branch attention-guided convolution neural network (AG-CNN) [11] was proposed by Guan *et al.* In order to reduce noise and enhance alignment, AG-CNN learns





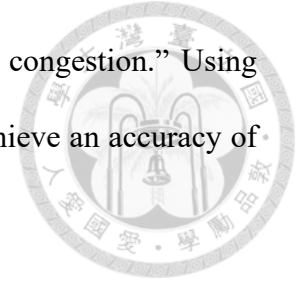
from disease-specific areas. It also incorporates a global branch to make up for the discriminative information that the local branch lost. Particularly, global images were used to train the global CNN branch initially. Then a mask was inferred to crop a discriminative zone from the global picture using the attention heat map created by the global branch as guidance. A local CNN branch was trained using the local images. In order to fine-tune the fusion branch, the last pooling layers of both the global and local branches were concatenated. AG-CNN was evaluated using the ChestX-ray14 dataset and produced an average AUROC of 0.871 when DenseNet-121 was applied as the backbone.

## 2.2 Predicting HF Using DL and CXR

Seah et al. [6] investigated generative visual rationales (GVRs) as a technique for visualizing neural network learning of chest radiograph characteristics in congestive HF. A total of 103,489 frontal chest radiographs from 46,712 individuals were separated into two groups: labeled (with BNP as a marker of congestive HF) and unlabeled. To estimate BNP, a generative model was trained on the unlabeled data set, and a neural network was trained on the encoded representations of the labeled data set. The model was used to depict how a radiograph with a high estimated BNP might seem in the absence of illness (a "healthy" radiograph). The model produced an AUC of 0.82 at a cutoff BNP of 100 ng/L as a marker of congestive HF.

Matsumoto Matsumoto *et al.* [7] studied the performance of a deep learning algorithm in detecting heart failure using CXR. The researchers used 952 CXR images from a labeled collection provided by the National Institutes of Health. A total of 260 "normal" and 378 "HF" pictures were validated and relabeled by two cardiologists. The ground

truth of HF is characterized as CXR findings of "cardiomegaly or congestion." Using CXR, data augmentation and transfer learning were employed to achieve an accuracy of 82% in detecting heart failure.



Jabbour Jabbour *et al.* [8] showed that machine learning models incorporating CXR and EHR data can distinguish HF effectively. Machine learning models were trained to predict HF using CXR and clinical data from the EHR. Based on a physician chart analysis, 363 (22%) of the 1,618 patients in their research had HF. A model that included CXR and EHR data outperformed the model using only one modality. The models performed similarly or better when compared to a randomly selected physician reviewer. The average AUROC for the combination model, image-only model, and EHR-only model was 0.83, 0.80, and 0.79, respectively.

Table 2.1 summarizes the related works of predicting HF using DL and CXR.

### **2.3 Multimodal DL Combining Medical Imaging and EHR**

Kawahara *et al.* [12] proposed a multitask deep convolutional neural network to diagnose skin lesions and categorize the 7-point melanoma checklist criteria. The network was trained on multimodal data (clinical and dermoscopic images and patient information). The neural network was trained using various multitask loss functions, each of which takes various combinations of the input modalities. This made the model resistant to missing data at the moment of inference. The final model categorized the skin disease diagnosis and the seven-point checklist, created multimodal feature vectors suited for image retrieval, and localized clinically discriminant regions.

Yap *et al.* [13] proposed a strategy for improving the performance of automated skin

lesion identification by combining different imaging modalities with patient metadata. The authors tested their model on a binary classification task for comparison with earlier studies and a five-class classification challenge typical of a real-world clinical setting. The results revealed that the multimodal classifier outperformed a baseline classifier that merely used a single macroscopic picture.

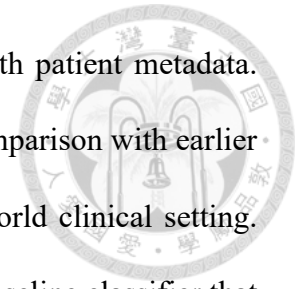


Table 2.1: Summary table of related works

| Publication                 | Task   | Input modality               | Definition of Positive HF  | Model Performance |
|-----------------------------|--|------------------------------|--|-------------------|
| Seah <i>et al.</i> [6]      | Develop a generative DL model and visualized neural network learning of CXR features in congestive HF using generative visual rationales | Frontal CXR                  | BNP > 100 ng/L   | AUROC = 0.82      |
| Matsumoto <i>et al.</i> [7] | Predict HF using a pre-trained DL model  | Frontal CXR                  | CXR findings “cardiomegaly or congestion” labeled by cardiologists | Accuracy = 0.82   |
| Jabbour <i>et al.</i> [8]   | Predict HF using multimodal DL model   | Frontal CXR and EHR          | ICD-10 code  | AUROC = 0.83      |
| Ours                        | Predict AHF in ED using multimodal DL model  | PA-view CXR, EHR, HSR and RF | NT-proBNP > 300 ng/L   | AUROC = 0.89      |







## Chapter 3 Materials and Methods

### 3.1 Data Description

Three open-source datasets, Medical Information Mart for Intensive Care (MIMIC)-IV [14, 15], MIMIC-CXR [16], and the Japanese Society of Radiological Technology (JSRT) dataset [17], were used in this study.

#### 3.1.1 MIMIC-IV

MIMIC-IV serves as the main dataset in our study. MIMIC-IV consists of six modules, Core, Hosp, ICU, ED, CXR, and Note module, which contains all the information of 382,278 patients admitted to an academic medical center in Boston, MA, USA, between 2008 and 2019. The NT-proBNP values for all examinations are stored in `labevents.csv` in the Hosp module. The itemid of NT-proBNP is 50963, which can be found in `d_labitems.csv` in the Hosp module. The patient's age and gender are stored in `patients.csv` in the Core module. The six vital signs (temperature, heart rate, respiration rate, o2 saturation, systolic blood pressure, and diastolic blood pressure) are stored in `vitalsign.csv` and `triage.csv` in the ED module. All age and `subject_id` in MIMIC-IV had been de-identified.



### 3.1.2 MIMIC-CXR

MIMIC-CXR is the core image dataset for our study. MIMIC-CXR contains 377,110 images in JPG format and 227,835 imaging studies for 64,588 patients. Each imaging study includes one or more images, a frontal view or a lateral view. Besides images, metadata including subject\_id, study\_id, dicom\_id, CXR views, study date, study time, etc, are also provided for all CXRs in MIMIC-CXR. dicom\_id, study date, and study time were used to connect the information in MIMIC-IV and MIMIC-CXR datasets. The demo images in MIMIC-CXR are shown in Fig. 3.1

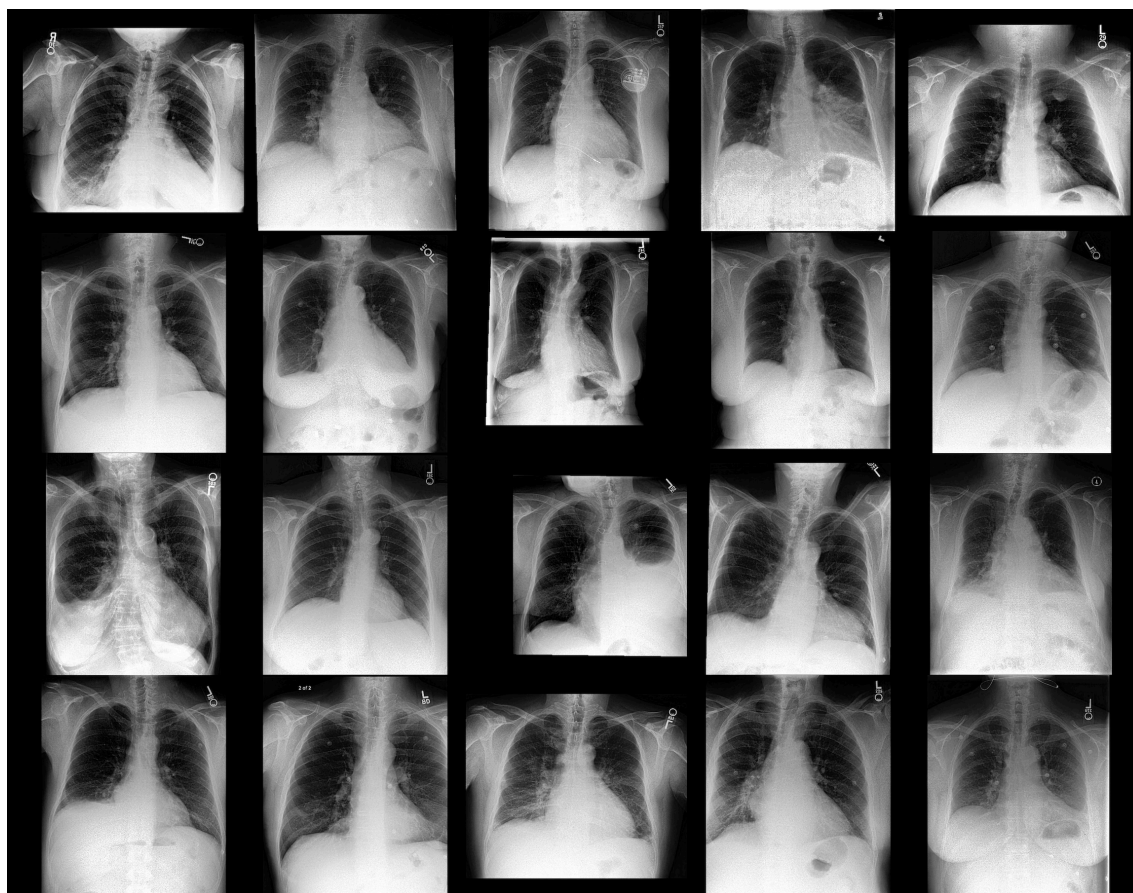


Figure 3.1: Examples of images in MIMIC-CXR



### 3.1.3 JSRT

JSRT dataset was used to develop two segmentation models for image processing purposes in this study. Before this study, the JSRT database has been used by many researchers in the world for various research purposes such as image processing, image compression, evaluation of image display, computer-aided diagnosis (CAD), picture archiving and communication system (PACS), and for training and testing. The JSRT dataset includes 247 frontal views CXR (154 nodule and 93 non-nodule images) with the size of 2048×2048 in 12-bit, 4096 grayscale. Each image in JSRT has its corresponding heart, left lung, right lung, left clavicle, and right clavicle masks. A lung-heart and clavicle segmentation model is trained using the JSRT dataset. The training procedures of the two segmentation models will be described in Sec. 3.4.1. The example images in the JSRT dataset and their corresponding masks are shown in Fig. 3.2. The images in odd columns are the original images from the JSRT dataset, and the images in even columns are the corresponding lung-heart masks. Every two images are a pair in each row, from left to right.

### 3.1.4 Manually labeled lung-heart masks for images in MIMIC-CXR

Visually, the images in the JSRT dataset are different from the ones in MIMIC-CXR. Although several operations can be applied to make the images more homogeneous from the MIMIC-IV and JSRT datasets, there are some non-neglectable differences between them. For example, there are no support devices in JSRT images but are in MIMIC-CXR ones. A few images in MIMIC-CXR have undetectable contours of the lung and heart, but the images in JSRT are relatively clean. 200 PA view CXR from MIMIC-CXR were



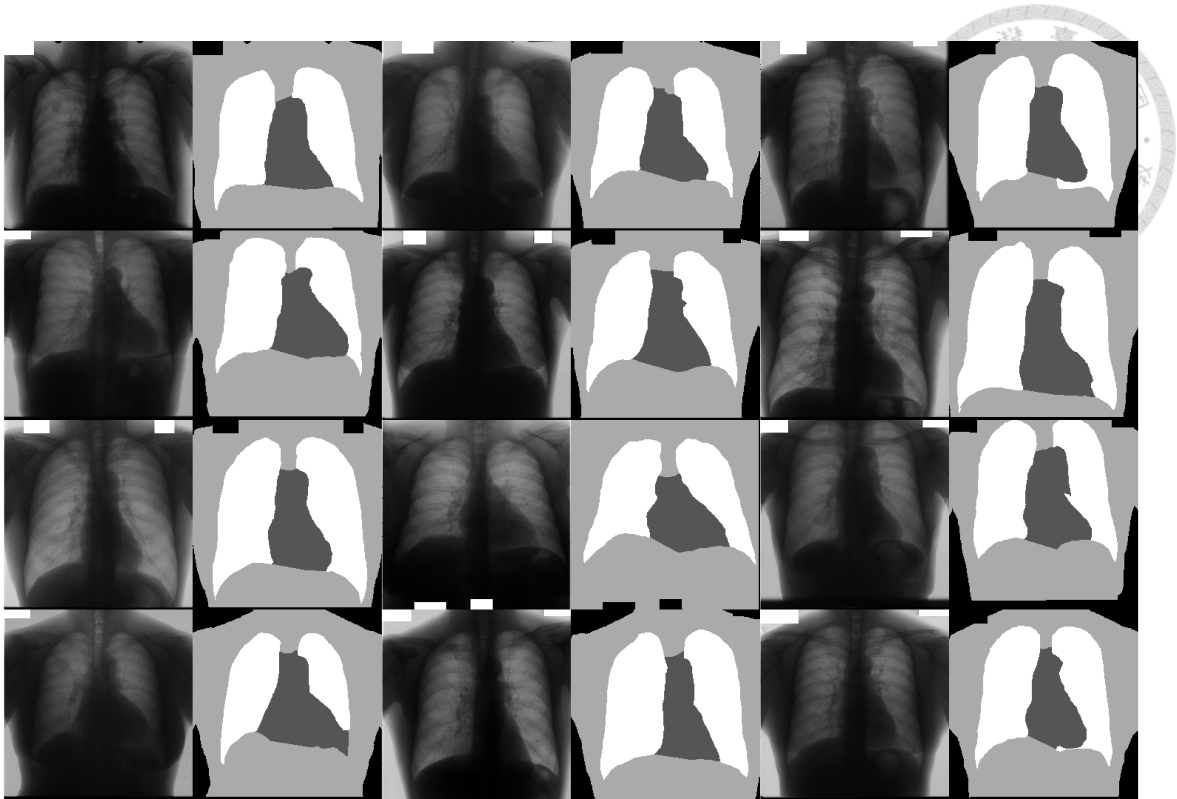


Figure 3.2: Examples of images in JSRT

manually labeled to overcome the divergence. The 200 PA view CXR were randomly selected. Six of them were discarded. Three of the six images are lateral views due to the labeling flaw of MIMIC-CXR, and the other three images have undetectable heart contours. The lung-heart masks were annotated following the same criterion in JSRT by four graduate students in the project using the open-source software CVAT [18]. The final mask annotations were revised and verified by a certified cardiologist in National Taiwan University Hospital Hsin-Chu Branch. The examples of manually labeled lung-heart masks and their corresponding images are shown in Fig. 3.3. The images in odd columns are the original images, and the images in even columns are the manually labeled lung-heart masks. Every two images are a pair in each row, from left to right. Obtaining more pairs of CXR and lung-heart mask makes training data distribution more similar to the distribution of MIMIC-CXR when training the lung-heart segmentation model.

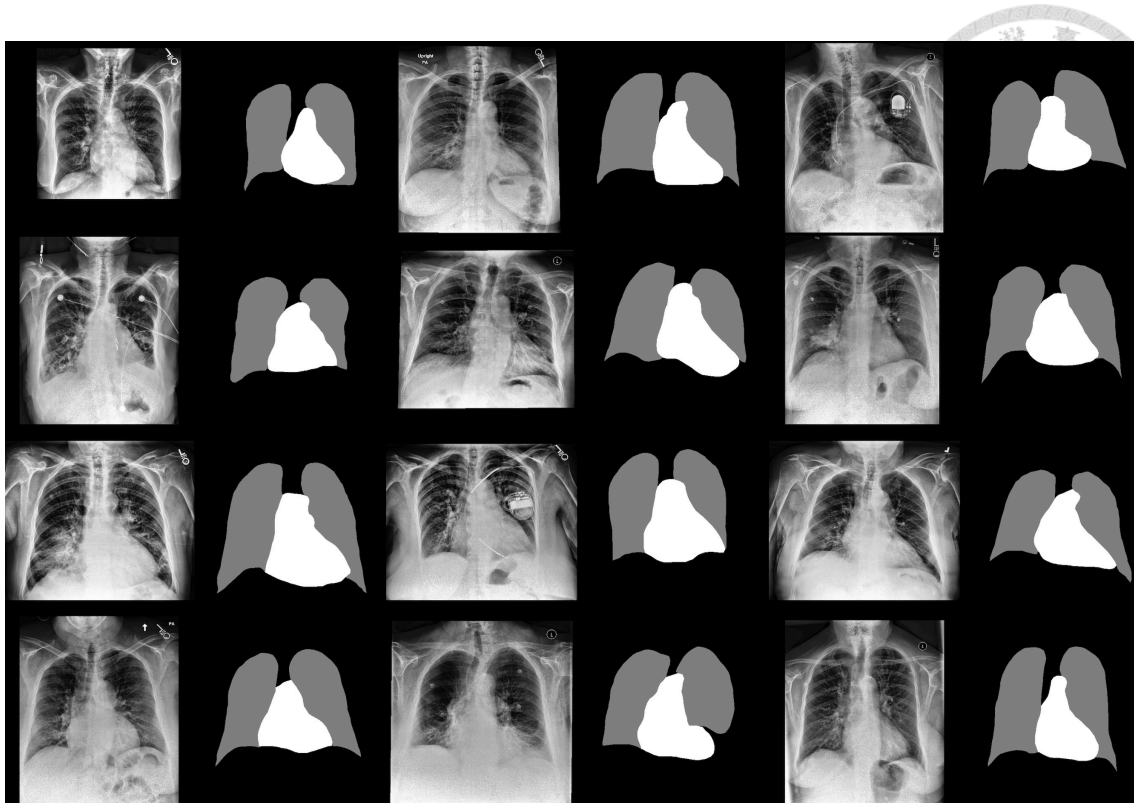


Figure 3.3: Examples of manually labeled images

## 3.2 Data Pre-processing

### 3.2.1 EHR from MIMIC-IV

First, patients who entered ED and took CXR were selected to obtain the target patient ID. Next, according to the unique patient ID, I extracted their gender and age from the patients.csv in the Core module. The patient's vital signs (temperature, heart rate, respiratory rate, O<sub>2</sub> saturation, systolic blood pressure, diastolic blood pressure) were extracted and integrated from vitalsign.csv and triage.csv in the ED module. All NT-proBNP values were extracted from labevents.csv in the HOSP module using itemid 50963. The itemid of NT-proBNP was found in d\_items.csv in the HOSP module. It is worth noting that only PA-view CXR is included in this study since heart size is a crucial factor of heart failure, and heart size is amplified in AP-view CXR. I paired the NT-proBNP value with

the closest CXR for each patient to combine the information. The time interval between two examinations is limited to at most three days. Likewise, NT-proBNP and vital signs are merged with the same time tolerance. Missing values in each patient's vital signs data frame are filled using the recent value based on the sample time. The data selection, extraction, and merging processes are demonstrated in Fig. 3.4.

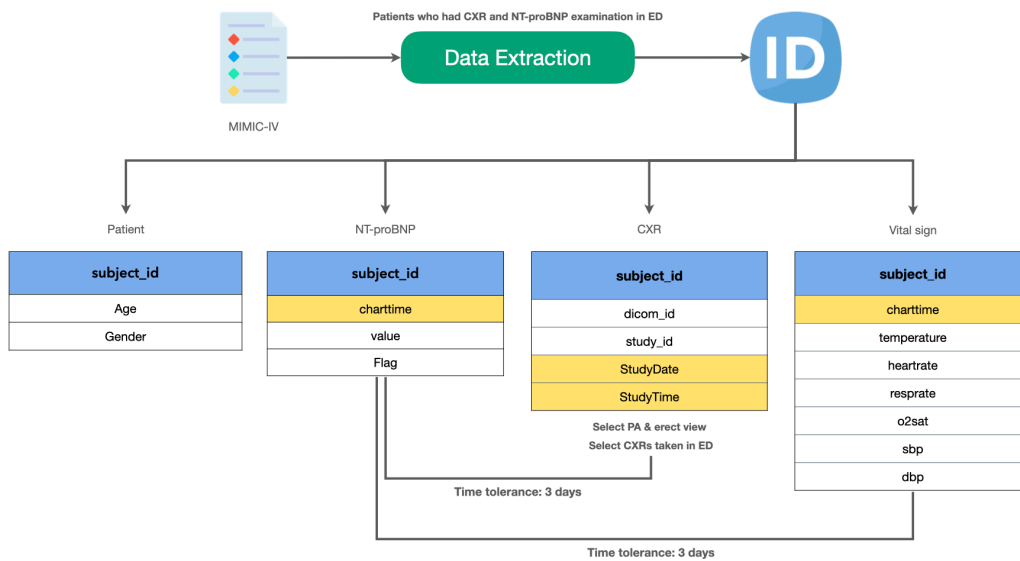


Figure 3.4: Data extraction, data selection, and data merging of MIMIC-IV

### 3.2.2 CXR from MIMIC-CXR

The chest X-ray images were downloaded from the MIMIC-CXR dataset according to the DICOM ID in the study population. The height and width of the images range from 1628 ~ 3056 and 1561 ~ 3056, respectively. As shown in Fig. 3.5, the image dimensions vary across the dataset, thus directly resizing all images to a specified size will distort some images. To maintain the aspect ratio of the original images, each image was first padded to square along the short side and then resized to 512 × 512. Contrast limited adaptive histogram equalization (CLAHE) was applied [19] with two sets of parameters, {clip limit

= 2, tile grid size = (8, 8)} and {clip limit = 4, tile grid size = (20, 20)}, to enhance the details in the second and the third channel, respectively. The effect of CLAHE is shown in Fig. 3.6. The images in the first row are the original CXRs, and the images in the second row and the last row were processed by CLAHE with parameter sets, {clip limit = 2, tile grid size = (8, 8)} and {clip limit = 4, tile grid size = (20, 20)}, respectively.

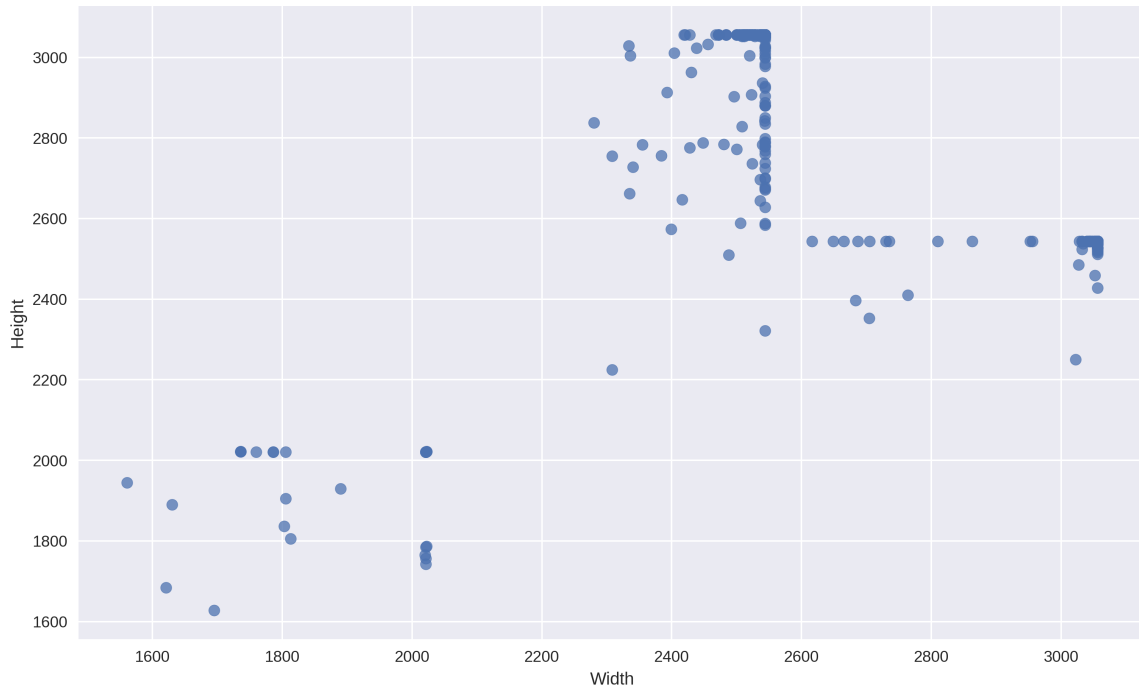


Figure 3.5: Scatter plot of image dimensions

### 3.2.3 CXR from JSRT

The images and masks in JSRT were directly resized to 512 x 512 since they were by default square. As shown in Fig. 3.7, the images are inverted types of MIMIC-CXR, which means the air is white, and the other tissues are dark. To match the pixel distribution of the two datasets, all pixels in JSRT images were inverted, and a histogram matching algorithm was applied to match the image histogram of JSRT to that of the reference image. The reference image was an averaged image of 100 random CXR from MIMIC-CXR. I also used CLAHE with the same parameters as above to enhance the details of the second and



Figure 3.6: Examples of images enhanced by CLAHE

third channels.

### 3.3 Frontal-lateral Classifier

This study merely focuses on PA-view CXRs. Thus only images labeled with PA-view were selected during the data selection stage. However, some lateral-view CXRs were mislabeled as frontal-view in metadata.csv of MIMIC-CXR. Therefore, a frontal-lateral classifier was trained to automatically filter out the miscategorized lateral-view CXRs. The frontal-lateral classifier extracts image information by a pre-trained DenseNet121 [20], and the output dimension of the last linear layer was changed from 1,000 to 1. A total of 1,000 frontal (500 PA-view and 500 AP-view) and 1,000 lateral CXRs were randomly sampled from the MIMIC-CXR dataset. All 2,000 CXRs were randomly split into training

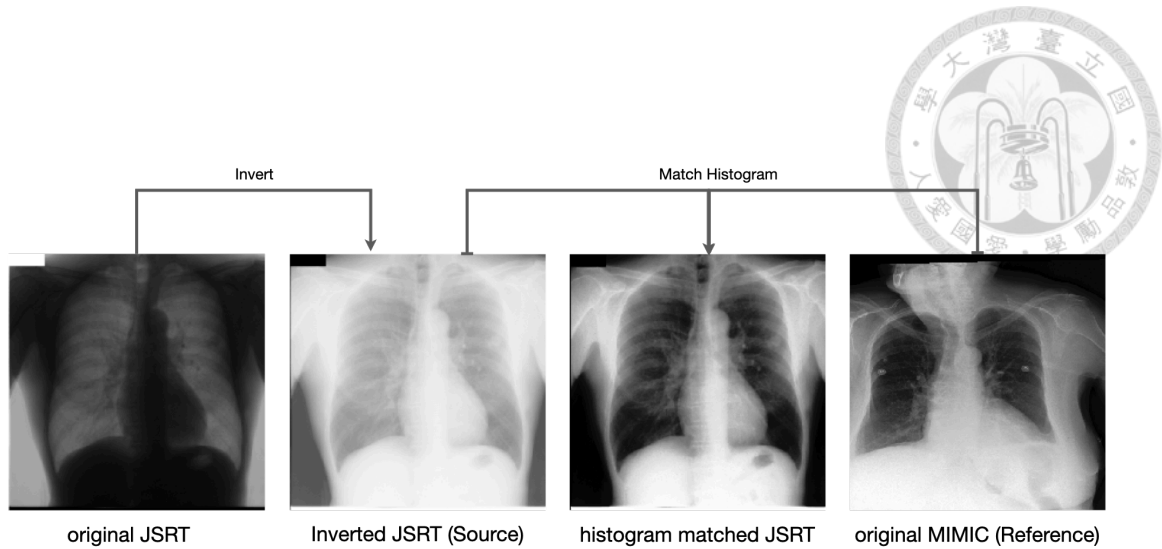


Figure 3.7: Image processing for JSRT

and testing sets of 1,600 and 400 CXRs. Each output from the FC layer was normalized to  $[0, 1]$  using the sigmoid activation function, defined by Eq. (3.1)

$$\hat{p} = \frac{1}{1 + e^{-s}}, \quad (3.1)$$

where  $\hat{p}$  is the predicted probability and  $s$  is the output from the FC layer.

### 3.4 Lung-heart Mask Generator

The lung-heart mask generator comprises a lung-heart segmentation model and a series of post-processing algorithms. For each CXR, the lung-heart segmentation model outputs the initial lung-heart mask. And two post-processing algorithms were applied to ensure the anatomical characteristics of the lung-heart mask.



### 3.4.1 Segmentation models

A multi-class semantic segmentation model was trained to generate lung-heart masks. U-Net [21] is usually the benchmark model for semantic segmentation tasks in medical imaging. U-Net consists of an encoder and a decoder. The encoder reduces feature maps and learns image representations, and the decoder recovers the spatial information and outputs a mask with the same dimension as the input image. U-Net also introduces skip connections from the encoder to the decoder, which allows the network to obtain sharp mask boundaries. Despite the success of U-Net, it falls short of capturing contextual information across the whole image. DeepLabv3+ [22] combines the encoder-decoder structure and the spatial pyramid pooling module and is inherently able to capture sharper object boundaries and learn multi-scale contextual information. In our lung-heart segmentation task, precise boundaries lead to reliable features later on, and contextual information is necessary due to the anatomical characteristics of CXR. A pre-trained encoder enables the network to extract better features. Therefore, I chose DeepLabv3+ as the network and ResNet50 [23] pre-trained on ImageNet as the backbone. Every location of the output maps was normalized along channels by the softmax function defined as:

$$\text{softmax}(p_i) = \frac{e^{p_i}}{\sum_j e^{p_j}}. \quad (3.2)$$

As defined by Eq. (3.3), the dice loss function was minimized to obtain optimal segmentation performances.

$$L_{dice}(\hat{P}, Y) = 1 - \frac{2|\hat{P} \cap Y|}{|\hat{P}| + |Y|}, \quad (3.3)$$

where  $\hat{P}$  is the segmentation confidence map and  $Y$  is the ground truth.

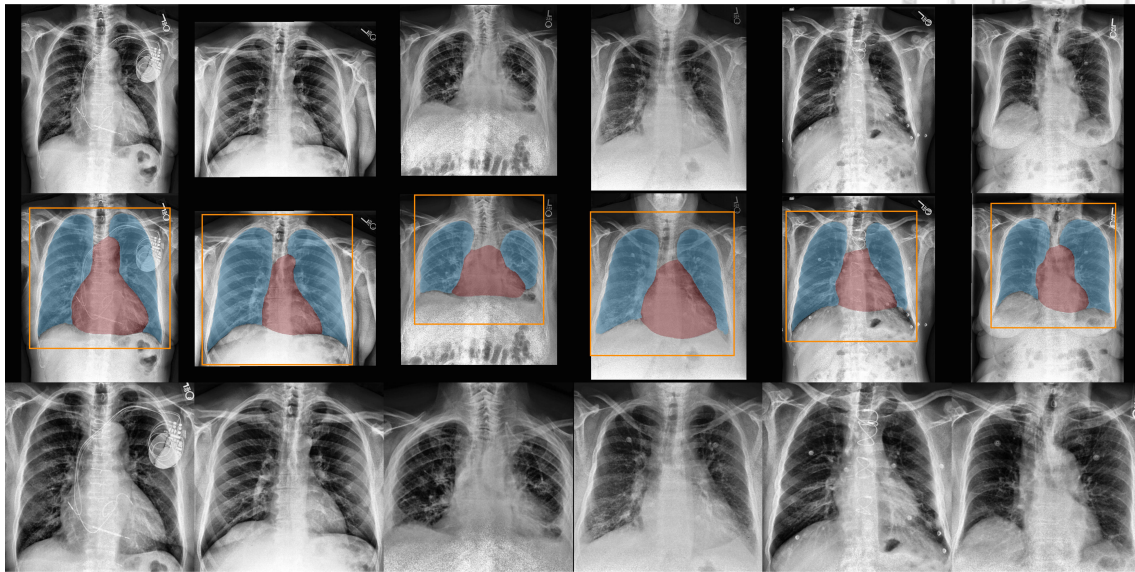


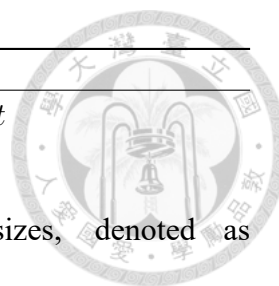
Figure 3.8: Examples of lung-heart masks

The examples of images and the predicted lung-heart masks are shown in Fig. 3.8. The background images are the original CXRs. The blue and red masks are the predicted lung and heart masks, respectively.

### 3.4.2 Mask post-processing

I found two types of problems among the output lung-heart masks. One problem was the false-positive regions, and the other was undetectable lung or heart contours. Without a doubt, a normal human has one heart and two lungs. Utilizing the anatomical characteristics, only the largest connected component in the heart mask will be kept, and at most two largest connected components in the lung mask will be saved. The second large region in the lung mask also is deleted if its size is smaller than one-third of the size of the largest part. The pixels in deleted regions are replaced by the most frequent value on the dilated border of the region.





---

**Algorithm 1** Mask post-processing (I): delete fragments

---

**Input:** Target mask  $M$ , number of components to keep  $k$ , threshold  $t$

**Output:** Post-processed mask  $M^*$

- 1: Find all connected components  $c_1, c_2, \dots, c_n$  in  $M$
  - 2: Sort all connected components descendingly by their sizes, denoted as  $c_{(1)}, c_{(2)}, \dots, c_{(n)}$ .
  - 3: **for**  $i = 2$  to  $n$  **do**
  - 4:   **if**  $(i > k)$  or  $(c_{(i)} < t * c_{(1)})$  **then**
  - 5:     replace all values in  $c_{(i)}$  with the most frequent value on its dilate border.
  - 6:   **end if**
  - 7: **end for**
  - 8: **return**  $M^*$
- 

Fig. 3.9 demonstrates the results of Alg. 1. The top row displays the original images. The masks in the middle row exist false positive regions. There are fragments of heart masks around or in the right lung regions in the first two images and fragments of lung masks in the other four images. The air in the intestines is more likely to be identified as the lung mask. The bottom row shows that the false-positive regions can be correctly eliminated by the algorithm.

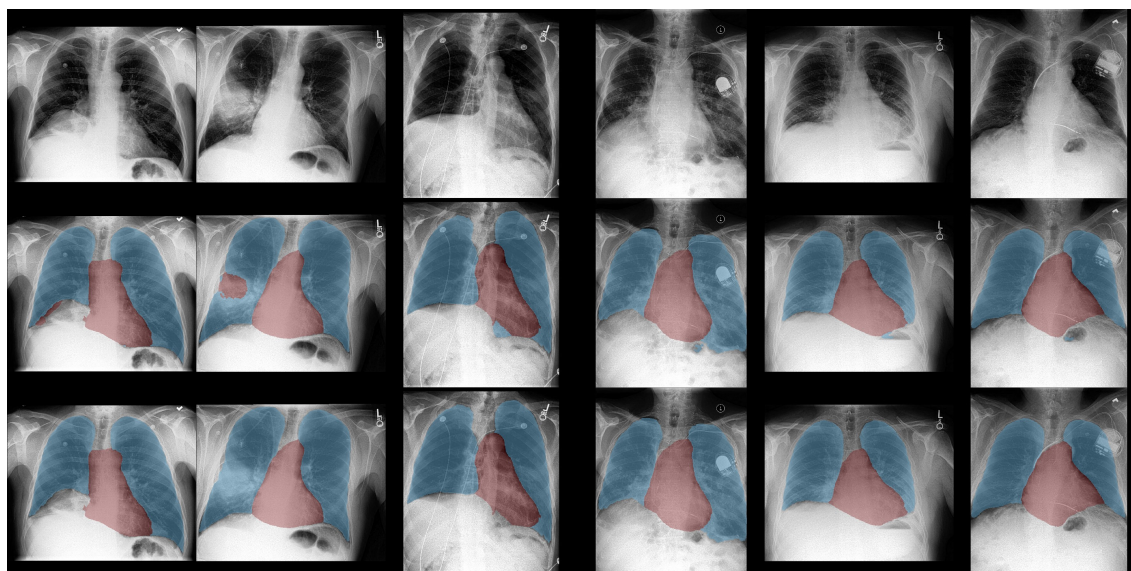


Figure 3.9: Examples of false-positive lung-heart masks

After false-positive regions were deleted, I found that the algorithm also deleted the lung with a large area of whiteness. Since lung-heart masks were used to generate multi-

ple features, it would cause tremendous bias if undetectable lungs or hearts were simply ignored. An algorithm was developed to recover the lost lung based on the property of human symmetric clavicles and lungs. A clavicle segmentation model was trained following the same procedure when training the lung-heart segmentation model. After localizing the left and right clavicles, I computed each clavicle's center and the line segment's slope connecting the two centers. Then the image was rotated and shifted to make the clavicle vertically symmetrical to the center of the image. Last, the lung mask was horizontally flipped to approximate two lungs.

---

**Algorithm 2** Mask post-processing (II): align and flip

---

**Input:** Lung-heart mask  $M_{LH}$ , clavicle mask  $M_{cla}$

**Output:** Post-processed lung-heart mask  $M_{LH}^*$

1: Find the centers of the left and right clavicle  $(x_{lc}, y_{lc})$  and  $(x_{rc}, y_{rc})$ , respectively.

2:  $\theta \leftarrow \arctan\left(\frac{y_{lc}-y_{rc}}{x_{lc}-x_{rc}}\right)$

3:  $x_c \leftarrow \frac{x_{lc}+x_{rc}}{2}$

4:  $M_{LH}^* \leftarrow$  apply affine transformation on  $M_{LH}$  with affine matrix

$$A = \begin{bmatrix} \cos \theta & \sin \theta & x_c - x \cos \theta - y \sin \theta \\ -\sin \theta & \cos \theta & 0 \end{bmatrix}$$

5: **return**  $M_{LH}^*$

---

The images in the first row of Fig. 3.10 show the original CXRs with undetectable lung or heart borders. The blue, red, and white masks indicate the lung, heart, and clavicle regions, respectively. The three red dots from left to right are the center of the right clavicle, the center of two centers of the left and right clavicles, and the center of the left clavicle, respectively. The blue dashed line in each image is the vertical center line of that image. All images and masks are affine transformed in the middle row by matrix  $A$  mentioned in Alg. 2. The green masks are the flipped lung masks, and the orange rectangles are the bounding boxes of the lung-heart regions, which will be introduced in Sec. 3.5.3. The last row shows the cropped images using the orange bounding boxes.

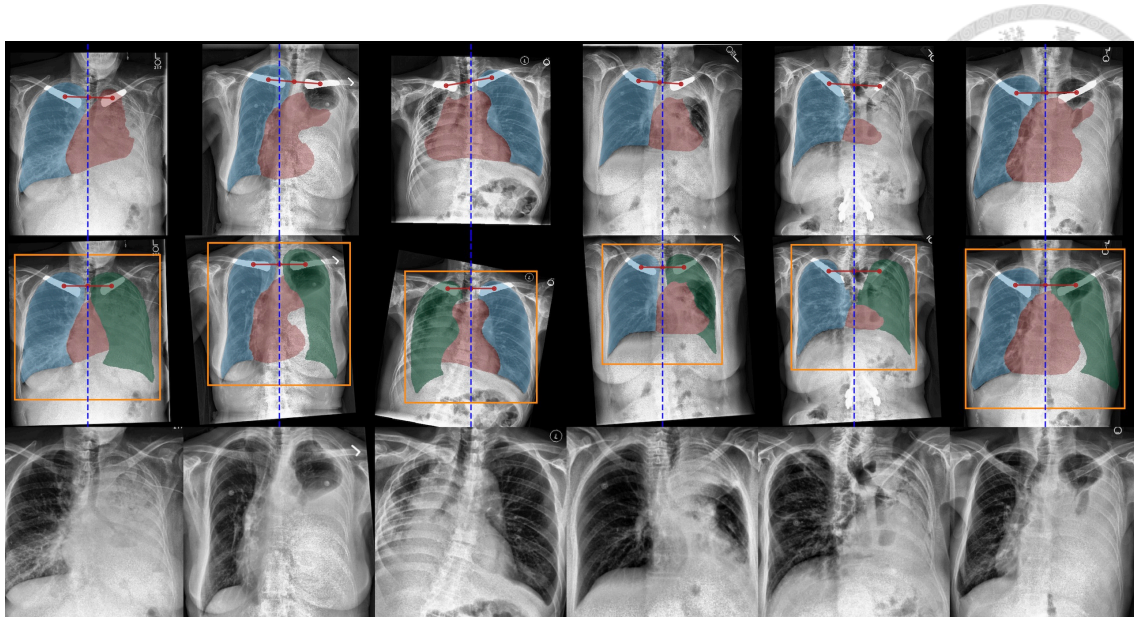


Figure 3.10: False negative lung-heart mask

## 3.5 Feature Extraction from Lung-Heart Masks

### 3.5.1 Heart-size ratios

Patients with AHF typically come together with enlarged heart size or cardiomegaly. Heart size is most commonly measured by the cardiothoracic ratio (CTR), the ratio of maximum horizontal cardiac diameter to maximum horizontal thoracic diameter. Cardiomegaly is defined as  $CTR > 0.50$ , and the normal range of CTR is from 0.42 to 0.50. Although CTR is simple and widely used, it is an alternative to estimating real heart size. Since heart sizes are important features to predict AHF, instead of using CTR as the only feature to measure heart size, we further derived the height and area ratios from the lung-heart masks to make the measurement more reliable. The three heart size ratios (HSR) are illustrated in Fig. 3.11 and defined in Eq. 3.4. The analysis of the three HSRs will be described in Sec. 4.8.

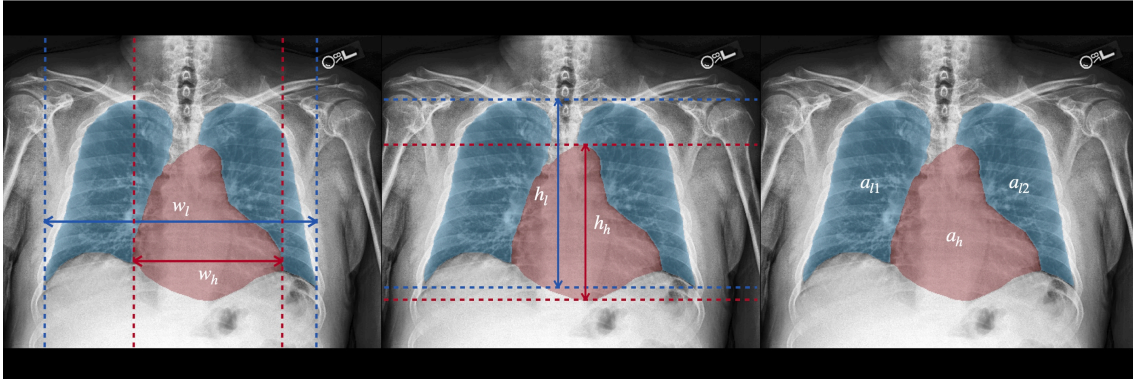


Figure 3.11: Demonstration of heart-size ratios

$$\begin{aligned}
 \text{width ratio} &= \frac{w_h}{w_l} \\
 \text{height ratio} &= \frac{h_h}{h_l} \\
 \text{area ratio} &= \frac{a_h}{a_h + a_{l1} + a_{l2}}
 \end{aligned} \tag{3.4}$$

### 3.5.2 Radiomic features

Chest X-ray findings in AHF patients include alveolar edema, Kerley B lines, dilated upper lobe vessels, and pleural effusion. These findings visually result in abnormal patterns or opacities in chest regions. Radiomic features (RF) were used to quantify the shape and textual differences in CXR. Radiomic features are features extracted from medical images by data-characterization algorithms. Radiomic features provide information about image intensity, size, and shape, and the relationship between image pixels. A total of 306 (102 features of the heart region, 102 features of the lung region, and 102 features of the union region of the lung and heart regions) are extracted using the Python open source library, Pyradiomics [24]. The analysis of the radiomic features is described in Sec. 4.9.



### 3.5.3 Thoracic region of interest

In the image processing pipeline, all images were padded to square and downsampled to  $512 \times 512$  due to the variety of image dimensions and hardware restrictions. However, these image operations inevitably added redundant information on the margins and caused information loss in details. The subtle differences in image details may affect model performances. Inspired by Liu *et al.* [10], I utilized lung-heart masks to construct bounding boxes for the thoracic regions of interest (ROI). A bounding box is the smallest square containing the lung-heart masks. Each bounding box was then expanded to match the size of the padded CXR. In addition, each bounding box was further expanded 3% larger based on the size of the original image. Lastly, the ROI was cropped using the expanded bounding box and downsampled to  $512 \times 512$ . Downsampling from high-resolution images loses less information than upsampling from low-resolution images if the final image size is the same. The ROIs excluded noises outside the chest and helped the model focus on the crucial region. The examples of ROI are shown in Fig. 3.12. The top row shows the original CXRs. The orange rectangles in the middle row are the bounding boxes of the thoracic regions. And the bottom row demonstrates the ROI images.

## 3.6 NT-proBNP Prediction Models

The proposed multimodal NT-proBNP prediction models consist of two types of single-modality models, numerical models, and image models. Each single-modality model includes an encoder and a classifier. The Encoders extract, transform and project features into a more complex vector space, and then the classifiers learn to predict the true binary label  $y^*$ , where

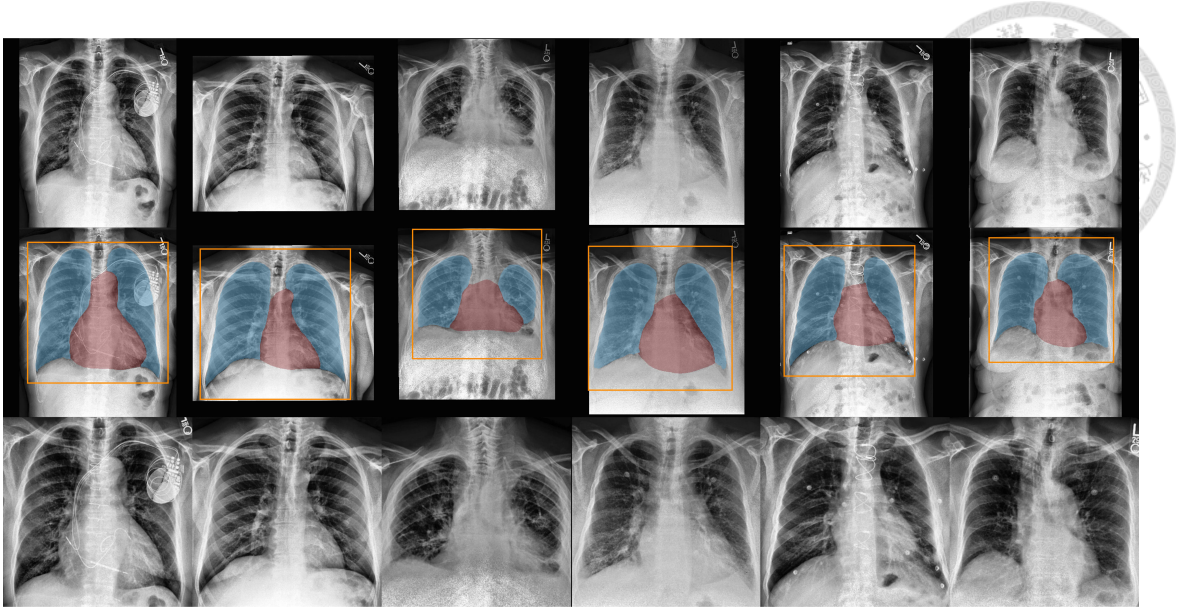


Figure 3.12: Examples of ROI

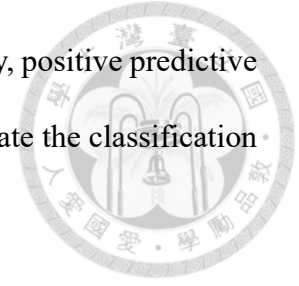
$$y^* = \begin{cases} 1, & \text{if NT-proBNP} \geq 300 \text{ ng/L} \\ 0, & \text{otherwise.} \end{cases}$$

All models were trained to minimize the binary cross-entropy (BCE) loss function.

$$L_{BCE}(\hat{\mathbf{p}}, \mathbf{y}^*) = -\frac{1}{N} \sum_{i=1}^N [y_i^* \log(\hat{p}_i) + (1 - y_i^*) \log(1 - \hat{p}_i)], \quad (3.5)$$

using Adam optimizer [25], where  $N$  is the batch size,  $\hat{p}_i$  is the predicted probability and  $y_i^*$  is the ground truth. To combat overfitting, regularization techniques such as weight decay and dropout were applied to all models. When training the models, the dataset was split into a training set and a preserved testing set at the patient level with a proportion of 8:2. The training dataset was further evenly split into five folds. In each iteration, the model was trained using four folds of data, validated by the last fold, and so on. Random sampling with weights was implemented to address data imbalance. Positive and negative data sampling weights are the inverse ratio of the two classes. The area under

the receiver operating characteristic (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) was used to evaluate the classification performance of the models.



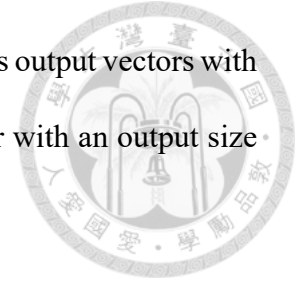
### 3.6.1 Numerical models

A total of 317 numerical features were used in this study.  $x_{EHR}$  is the eight EHR every patient initially has.  $x_{HSR}$  and  $x_{RF}$  respectively represent the three HSRs and 306 RFs extracted during the image processing pipeline for each CXR. Three numerical categories can be combined to seven input combinations ( $x_{EHR}$ ,  $x_{HSR}$ ,  $x_{RF}$ , ( $x_{EHR}$ ,  $x_{HSR}$ ), ( $x_{EHR}$ ,  $x_{RF}$ ), ( $x_{HSR}$ ,  $x_{RF}$ ) and ( $x_{EHR}$ ,  $x_{HSR}$ ,  $x_{RF}$ )). The vectors in the input combination set were concatenated for each input combination. The concatenated vectors are then passed into the encoder, a fully-connected layer (FC) with 1,024 hidden sizes. The hidden size was set to 1,024 to match the output size of the image encoder, which will be described in the next subsection. The FC with an input size of 1,024 and an output size of 1 is the classifier for the numerical models. ReLU was added after every hidden layer as an activation function.

### 3.6.2 Image models

Two image models were developed to predict the target using CXR ( $x_{CXR}$ ) and ROI ( $x_{ROI}$ ), respectively. Each encoder of the image model is a DenseNet121 with pre-trained weights on ImageNet. The original classifier from DenseNet121 was discarded. Since DenseNet121 is a powerful model with high flexibility, all model parameters were fixed except the ones from the last dense layer to prevent severe overfitting. I also added a global

average pooling layer following the image encoder. The image models output vectors with 1,024 dimensions. The classifier of the image models is an FC layer with an output size of 1.



### 3.6.3 Multimodal models

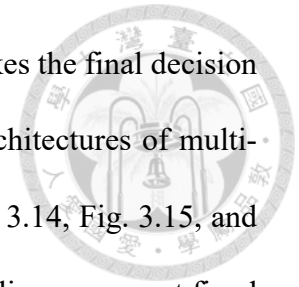
Multimodal models fuse information from different modalities to gain a better ability to learn the underlying function. Six input combinations were designed to demonstrate the significance of different input modalities. Taking  $x_{CXR}$  image model as the benchmark.  $(x_{EHR}, x_{CXR})$  shows the benefit of simply adding patient background information.  $(x_{HSR}, x_{RF}, x_{CXR})$  shows the benefit of the manually extracted features. Inspired by [10],  $(x_{CXR}, x_{ROI})$  verifies the benefit of combining global and local images. Inspired by [13],  $(x_{EHR}, x_{CXR}, x_{ROI})$  verifies the benefit of combining patients' information global images and local images.  $(x_{HSR}, x_{RF}, x_{CXR}, x_{ROI})$  shows the benefit of combining all the features extracted from CXR.  $(x_{EHR}, x_{HSR}, x_{RF}, x_{CXR}, x_{ROI})$  shows the benefit of combining all features.

### 3.6.4 Fusion strategy

After well training single-modality models, early fusion, joint fusion, and late fusion strategies were used to develop multi-modality models. According to the definitions introduced in [26], early fusion is a method to fuse (concatenate or pool) raw features or extracted features and only backpropagate the gradient to the final classifier while training. Joint fusion is similar to early fusion, but all parameters, including ones of feature encoders and classifiers, are trainable during the learning process. Late fusion is a pro-



cess that leverages predictions from single-modality models and makes the final decision using an aggregation function such as averaging or voting. The architectures of multi-modal models with early, joint, and late fusion are illustrated in Fig. 3.14, Fig. 3.15, and Fig. 3.16, respectively. The rectangles with solid and dashed border lines represent fixed and trainable components during the training process.



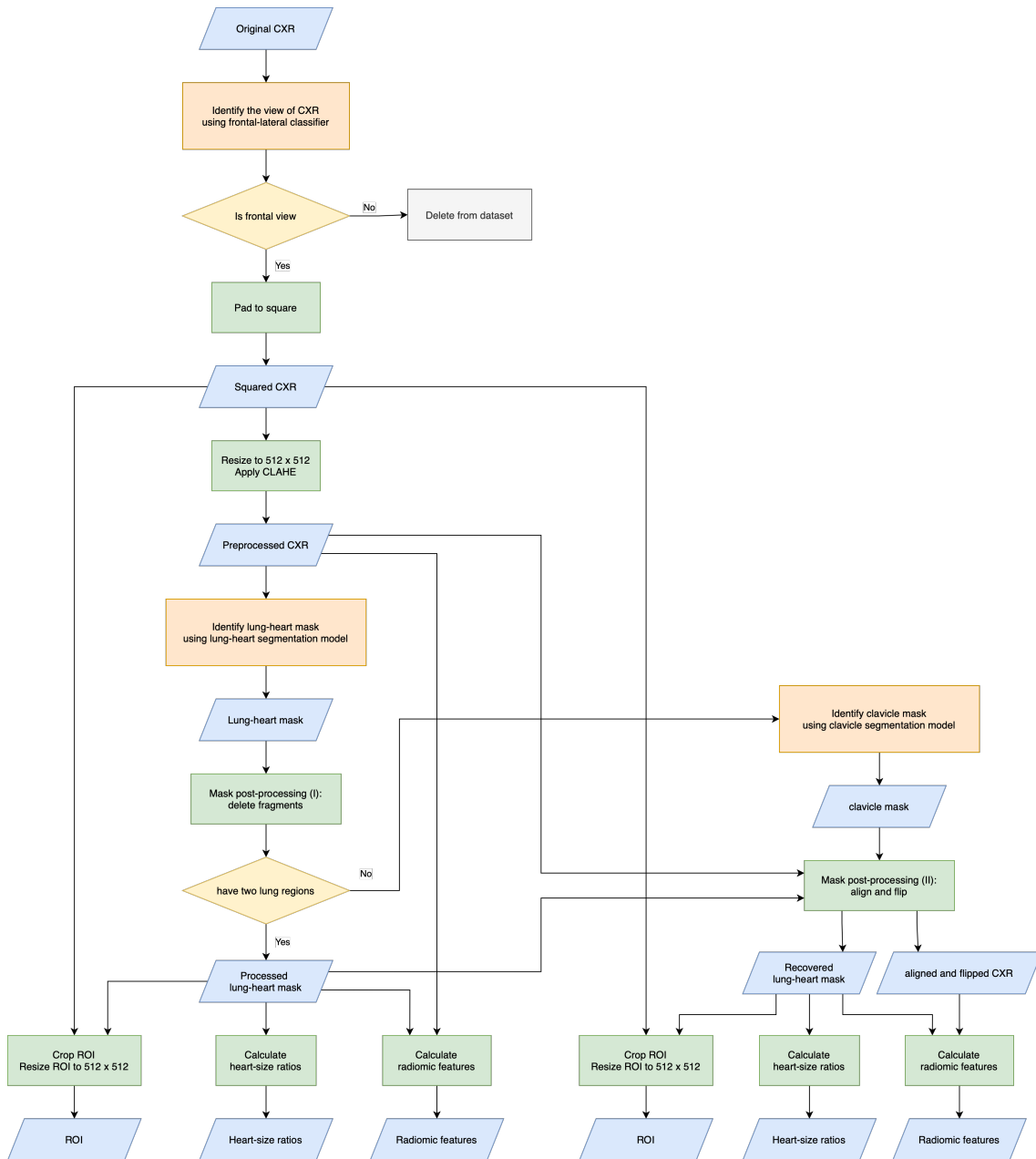


Figure 3.13: Image processing pipeline

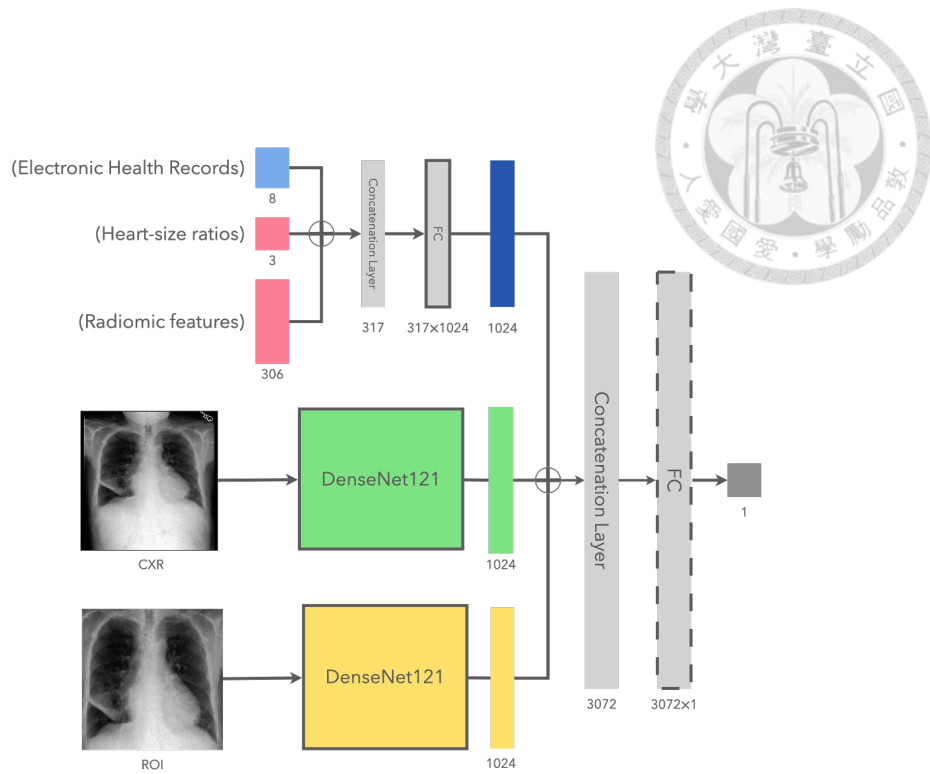


Figure 3.14: Model architecture – early fusion

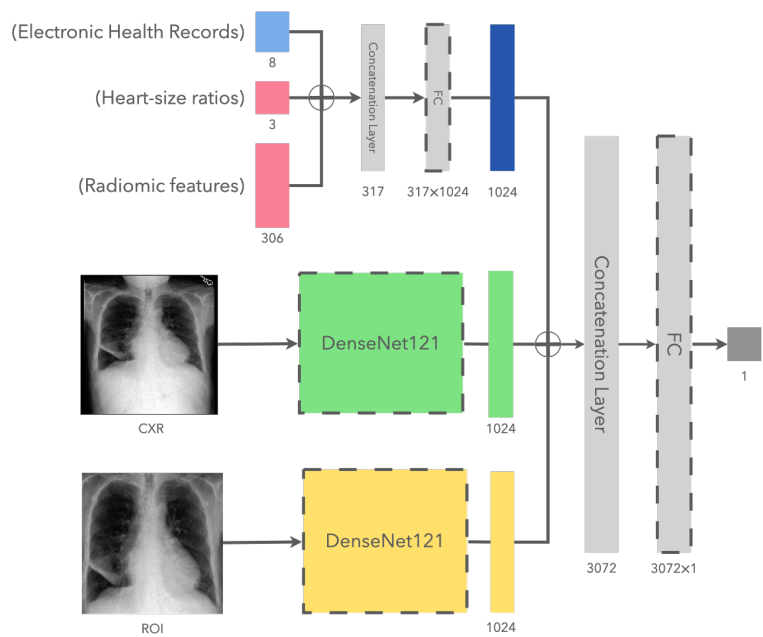


Figure 3.15: Model architecture – joint fusion

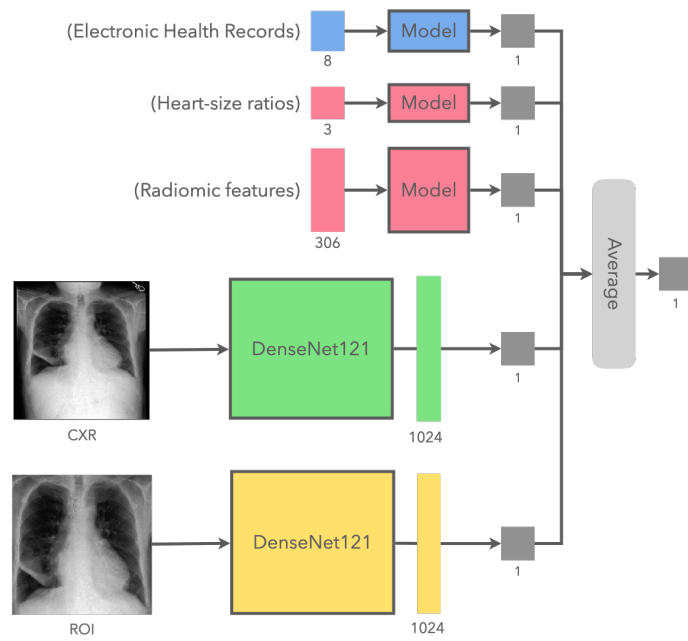


Figure 3.16: Model architecture – late fusion





## Chapter 4 Results and Discussions

### 4.1 Demographics of Study Population

A total of 1,833 pairs of samples from 1,432 patients were included. Overall, 71% of the sample are NTproBNP > 300 ng/L. The proportion of men (53%) is higher than that of women (47%), and most patients are elderly in the study population. The demographic characteristics of the study population are reasonable since high-risk patients commonly take chest radiography and NT-proBNP examinations in a short time interval. The mean and median time differences between the NT-proBNP and CXR examinations are 4.82 and 2.7 hours, respectively.

### 4.2 PA, AP, and Frontal Views

In this study, only PA-view CXRs were selected for the study population because PA-view CXRs are the most common view of CXR and do not amplify the size of the heart. Table 4.2 summarizes the demographic characteristics of populations with PA, AP, and frontal (PA + AP) views. According to clinical knowledge, patients with more severe conditions could only take AP-view CXRs since merely standing is challenging. Based on the data, patients in the AP population have higher NT-proBNP values, heart rates, and

Table 4.1: Demographic characteristics of the study population

| Characteristic                                  | Training        | Testing         | All             |
|---|-----------------|-----------------|-----------------|
| No. of CXR                                      | 1464            | 369             | 1833            |
| No. of CXR with NT-proBNP<300 ng/L              | 417 (28.48%)    | 116 (31.44%)    | 533 (29.08%)    |
| Median NT-proBNP (ng/L)*                        | 1147 (245-4147) | 1164 (219-4054) | 1158 (240-4130) |
| No. of patients                                 | 1145            | 287             | 1432            |
| No. of female patients                          | 541 (47.25%)    | 131 (45.64%)    | 672 (46.93%)    |
| No. of male patients                            | 604 (52.75%)    | 156 (54.36%)    | 760 (53.07%)    |
| Age range                                       | 20-91           | 26-91           | 20-91           |
| Mean age <sup>†</sup>                           | 65 (15)         | 66 (14)         | 65 (15)         |
| Mean temperature (°F) <sup>†</sup>              | 98.16 (0.79)    | 98.03 (0.79)    | 98.14 (0.79)    |
| Mean heart rate (bpm) <sup>†</sup>              | 82.28 (18.21)   | 79.69 (18.16)   | 81.76 (18.23)   |
| Mean respiratory rate (bpm) <sup>†</sup>        | 18.86 (3.47)    | 18.52 (3.48)    | 18.80 (3.48)    |
| Mean O <sub>2</sub> saturation (%) <sup>†</sup> | 97.43 (2.48)    | 97.58 (2.30)    | 97.46 (2.44)    |
| Mean SBP (mmHg) <sup>†</sup>                    | 134.51 (24.94)  | 136.93 (25.29)  | 135.00 (25.02)  |
| Mean DBP (mmHg) <sup>†</sup>                    | 72.77 (15.29)   | 73.72 (15.78)   | 72.96 (15.39)   |

\* Data in parentheses are interquartile ranges

<sup>†</sup> Data in parentheses are standard deviations

respiratory rates, which reflect their bad health conditions.

I trained three models using the CXR images of three populations (PA, AP, and frontal view), and the results are shown in Table 4.3. The model trained by all frontal CXRs performs the best, and the model trained by AP-view CXR has the worst performance. As shown in Fig. 4.1, AP-view CXRs have larger diversity and therefore are more difficult to learn using a small dataset. On the other hand, because the frontal-view population includes more data, the model trained by all frontal CXRs performs the best. The results show that the number of data plays a great role in model development, and the proposed multimodal model could achieve a higher AUROC if the lung-heart segmentation model can well identify the lung-heart masks in AP-view CXRs and trained with all available frontal CXRs. Fig. 4.2 displays the AUROC of CXR models trained by populations of different views.

Table 4.2: Demographic characteristics of populations

| Characteristic                       | PA              | AP              | Frontal         |
|--------------------------------------|-----------------|-----------------|-----------------|
| No. of CXR                           | 1833            | 2287            | 4120            |
| No. of CXR with NT-proBNP < 300 ng/L | 533 (29.08%)    | 324 (14.17%)    | 857 (20.80%)    |
| Median NT-proBNP (ng/L)*             | 1158 (240-4130) | 2427 (679-7237) | 1789 (404-5686) |
| No. of patients                      | 1432            | 1789            | 3021            |
| No. of female patients               | 672 (46.93%)    | 970 (54.22%)    | 1551 (51.34%)   |
| No. of male patients                 | 760 (53.07%)    | 819 (45.78%)    | 1470 (48.66%)   |
| Age range                            | 20-91           | 18-91           | 18-91           |
| Mean age†                            | 65 (15)         | 72 (14)         | 69 (15)         |
| Mean temperature (°F)†               | 98.14 (0.79)    | 98.25 (1.05)    | 98.20 (0.95)    |
| Mean heart rate (bpm)†               | 81.76 (18.23)   | 86.81 (21.86)   | 84.54 (20.46)   |
| Mean respiratory rate (bpm)†         | 18.80 (3.48)    | 20.62 (4.92)    | 19.80 (4.43)    |
| Mean O2 saturation (%)†              | 97.46 (2.44)    | 97.02 (3.12)    | 97.22 (2.84)    |
| Mean SBP (mmHg)†                     | 135.00 (25.02)  | 128.73 (27.03)  | 131.54 (26.33)  |
| Mean DBP (mmHg)†                     | 72.96 (15.39)   | 69.21 (17.09)   | 70.89 (16.45)   |

\* Data in parentheses are interquartile ranges

† Data in parentheses are standard deviations

Table 4.3: Model performances of different views of CXR

|     | PA view         | AP view         | Frontal view           |
|-----|-----------------|-----------------|------------------------|
| CXR | 0.8248 (0.0150) | 0.7931 (0.0221) | <b>0.8426 (0.0070)</b> |

Data in parentheses are standard deviations.

### 4.3 Evaluation of Mask Post-processing

In this section, I will demonstrate how mask post-processing algorithms affect final results. Since only HSR, RF, and ROI are affected by the mask post-processing algorithms, only the results of the three are reported. From left to right in Fig. 4.3 shows the original mask generated by the lung-heart segmentation model, the mask deleted fragments by Alg. 1, and the mask applied both Alg. 1 and Alg. 2. Table 4.4 shows that the models constantly perform best when both post-processing algorithms were implemented and perform worst when only post-processing (I) was applied. The post-processing algorithms maintain the anatomical characteristics of the lung-heart mask and output more appropriate masks.



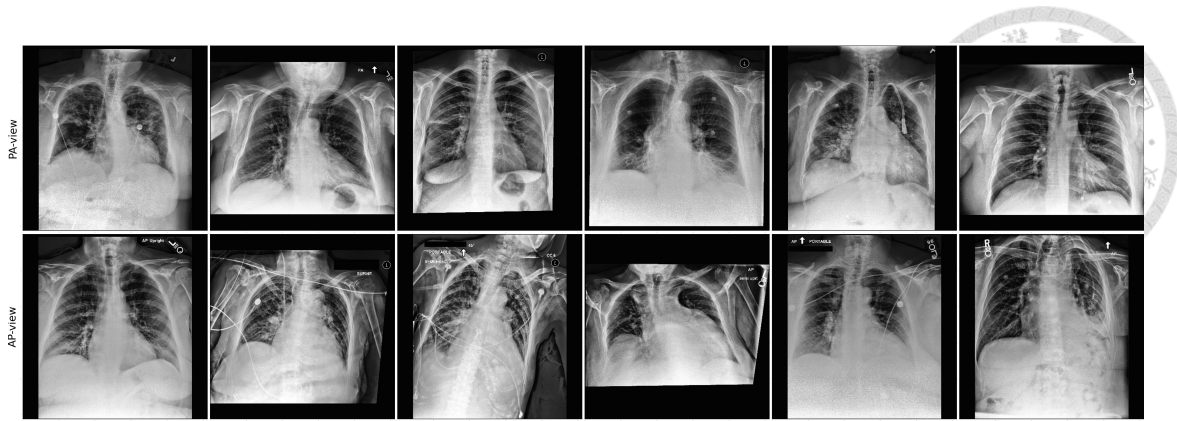


Figure 4.1: Examples of PA-view and AP-view CXRs

Based on the results, I conclude that the proposed mask post-processing algorithms can improve model performance and should be applied together. Fig. 4.4 displays the AUROC of HSR, RF, and ROI with different post-processing algorithms.

Table 4.4: AUROC of models using different mask post-processing

|     | No Post-processing | Post-processing (I) | Post-processing (I) & (II) |
|-----|--------------------|---------------------|----------------------------|
| HSR | 0.7512 (0.0055)    | 0.7468 (0.0023)     | <b>0.7591 (0.0151)</b>     |
| RF  | 0.8626 (0.0051)    | 0.8603 (0.0027)     | <b>0.8645 (0.0060)</b>     |
| ROI | 0.8186 (0.0118)    | 0.8170 (0.0129)     | <b>0.8494 (0.0063)</b>     |

Note: Data in parentheses are standard deviations.

#### 4.4 Performance Based on Input Combination

Table 4.5 summarizes the five-fold cross-validation results. The numerical model input EHRs achieves an average AUROC of 0.7146, meaning that the basic vital signs could help predict the result. The manually extracted features, HSRs and RFs, provide useful information, which can build up a model with an average AUROC of 0.7582 and 0.8634, respectively. The numerical model combining EHRs, HSRs, and RFs achieved great performance with an average AUROC of 0.8650. The average AUROC increased from 0.7146 to 0.7582 after adding three HSRs and jumped to 0.8650 when 306 radiomic

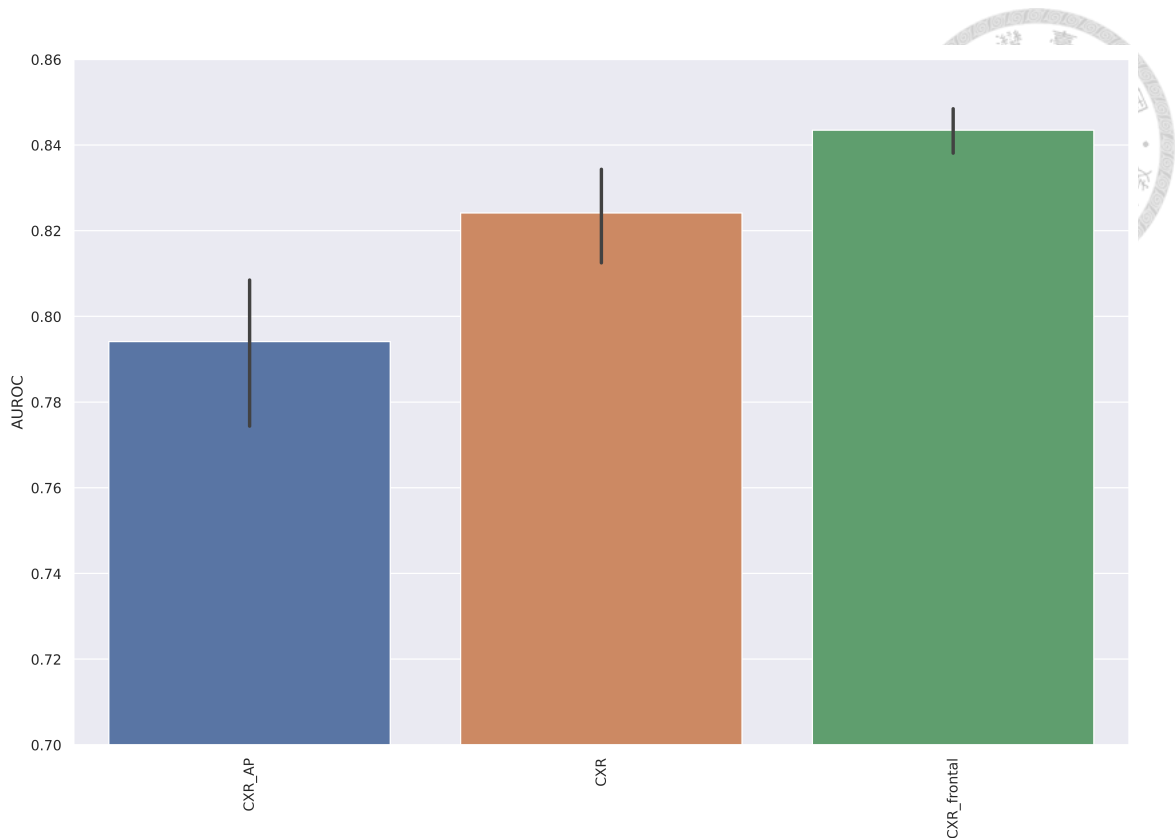


Figure 4.2: Box plots of AUROC of different CXR views

features joined in.

The AUROC of CXR and ROI image models are 0.8241 and 0.8490, respectively. Since the ROI images exclude redundant parts and focus on the crucial region of the original CXRs, the ROI model is expected to perform better than the CXR model.

Surprisingly, the RF numerical model outperforms the CXR and ROI image models. In our opinion, training a convolutional deep learning network requires a large dataset, and the number of data in our study population may not satisfy the demand for deep learning models. On the contrary, RFs are pre-defined complex features designed by domain experts and, therefore, can produce specific information in medical imaging tasks without learning the data.

The multimodal models outperform every modality model with the highest mean

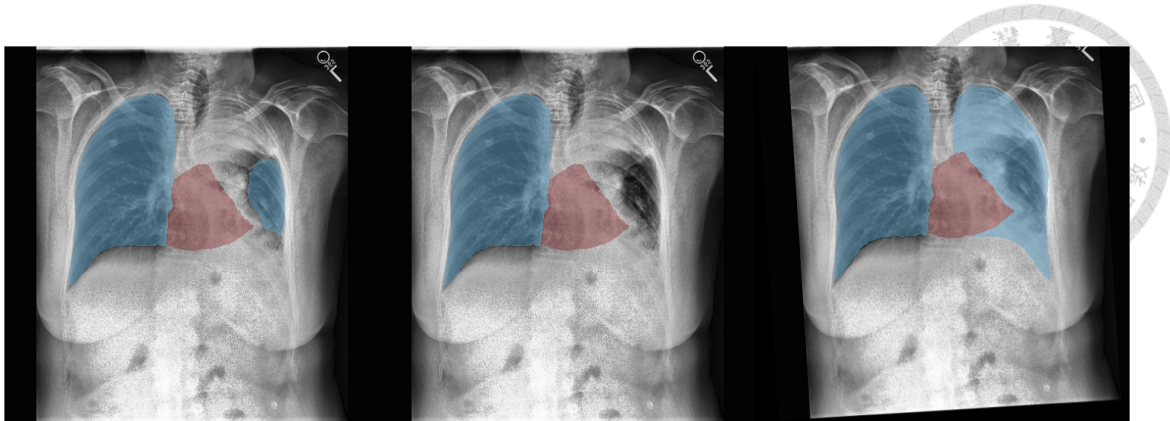


Figure 4.3: Mask post-processing

AUROC of 0.8861. The trend in Fig. 4.5 from left to right visualizes the incremental improvements when more information is provided to the model.

## 4.5 Performance Based on Fusion Strategy

Fig. 4.6 visualizes the performance of multimodal models stratified by three different fusion strategies. Overall, models trained with late fusion have better prediction performance. The late fusion strategy takes average on all predictions from each single-modality model. This strategy combines the outcomes without further training, which is more likely to prevent overfitting and produce more robust results. On the other hand, the feature extractor or the final classifiers are continuously trained when using early and joint fusion. These fusion strategies enable multimodal models to better capture task-specific features. However, the extra flexibility also causes overfitting or makes the models overly pay attention to a single modality, resulting in inferior performance. Despite the detailed discussion about fusion strategies, the difference between the strategies is minor, and the result may differ from dataset to dataset. Therefore, researchers and practitioners should explore different fusion strategies and search for the best method for their problem.

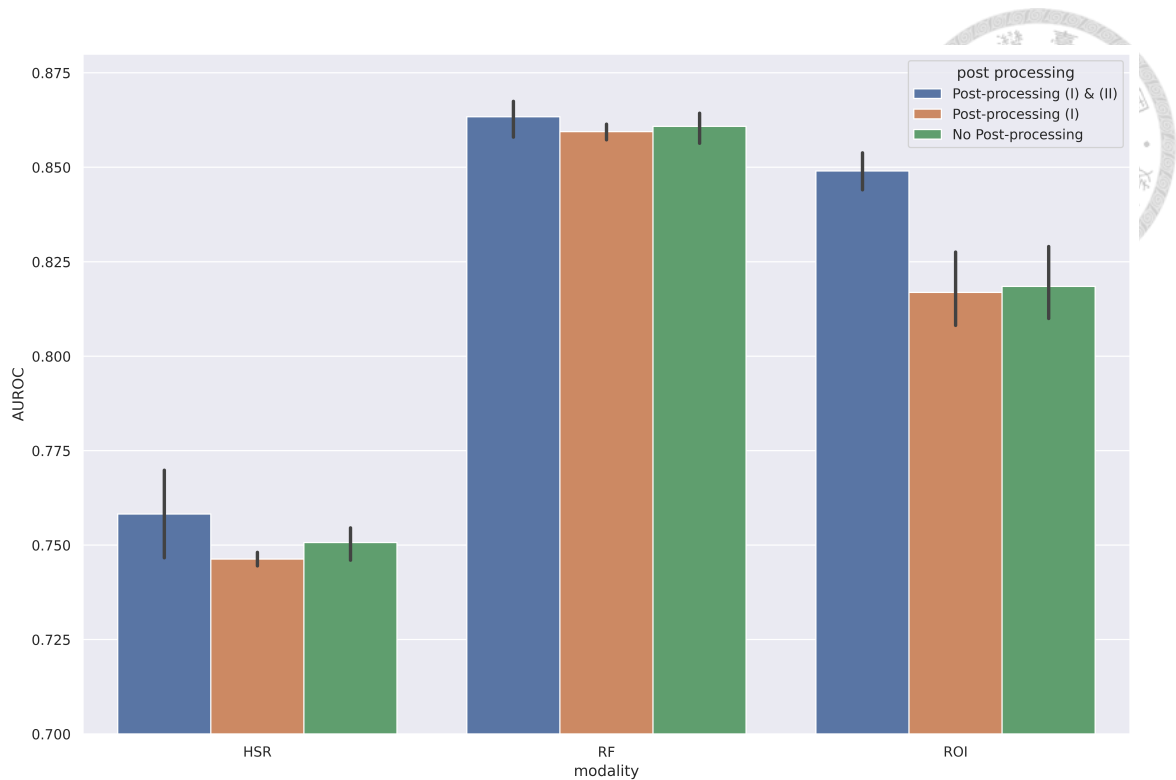


Figure 4.4: Box plots of AUROC of different mask post-processing algorithms

## 4.6 Analysis of Image Data

CAM [27] is a technique to visualize what CNN models see. It is a weighted linear combination of the learned parameters and the corresponding feature maps. By simply upsampling the weighted sum of feature maps to the input image size, the image regions most relevant to the target classes can be identified. The heatmap of the global and the local branch is displayed in Fig. 4.8. The models put their attention on the heart region. This result is consistent with the clinical knowledge since NT-proBNP is a biomarker indicating the condition of the heart. Additionally, it is reasonable that the models also focus on pacemakers because patients equipped with pacemakers are usually in more challenging situations.

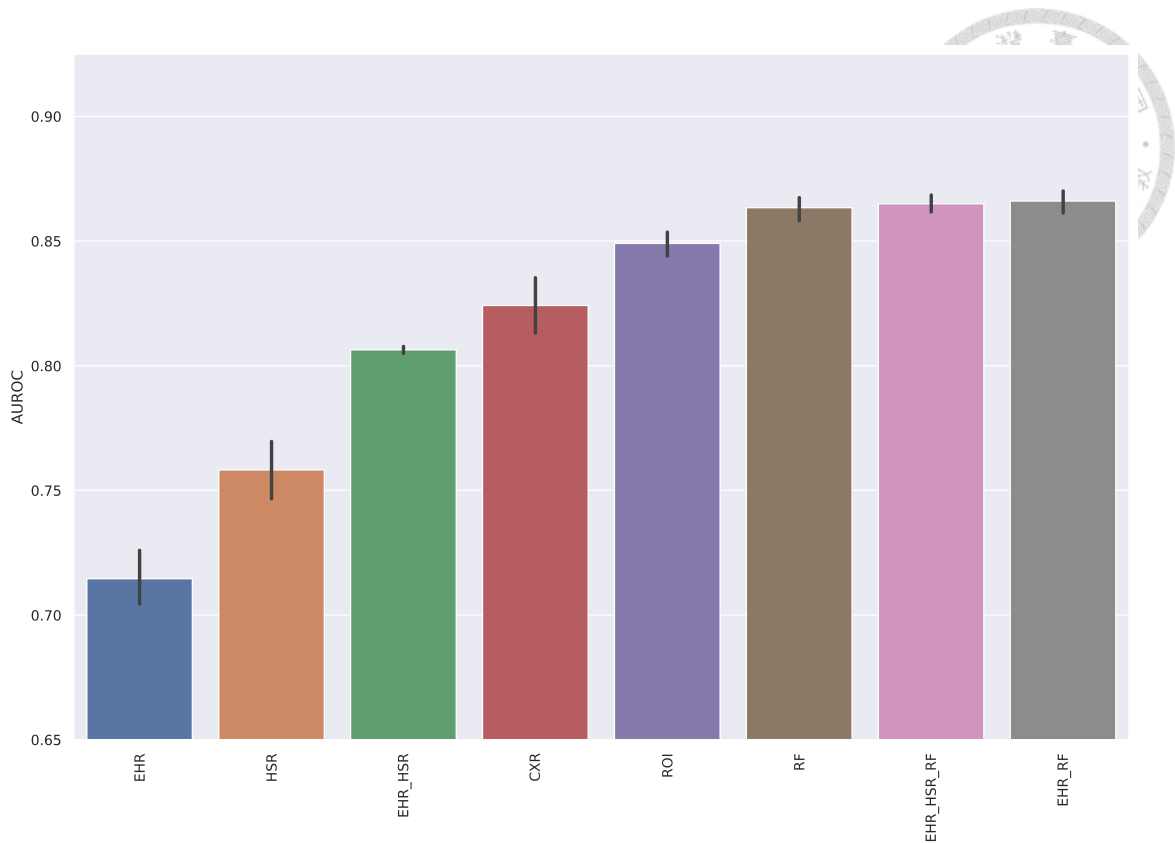


Figure 4.5: Box plots of AUROC of different input combinations

## 4.7 Analysis of Numerical Data

A total of 317 numerical features were used in this study, including eight EHR features, three HSRs, and 306 RFs. Compared to deep learning features, human-recognizable numerical features provide more interpretability and help humans understand how machine learning models make decisions. I utilized the XGBoost [28] algorithm to explain the relationship between 317 numerical features and the level of NTproBNP. XGBoost, eXtreme Gradient Boosting, is a famous tree-based algorithm that extends to gradient boosted decision trees (GBM) and is specially designed to improve speed and performance. Like other tree-based models, feature importance can be calculated after training an XGBoost model. Feature importance gives us a quantified measure of a certain feature's importance during the training procedure. The definition of feature importance varies;

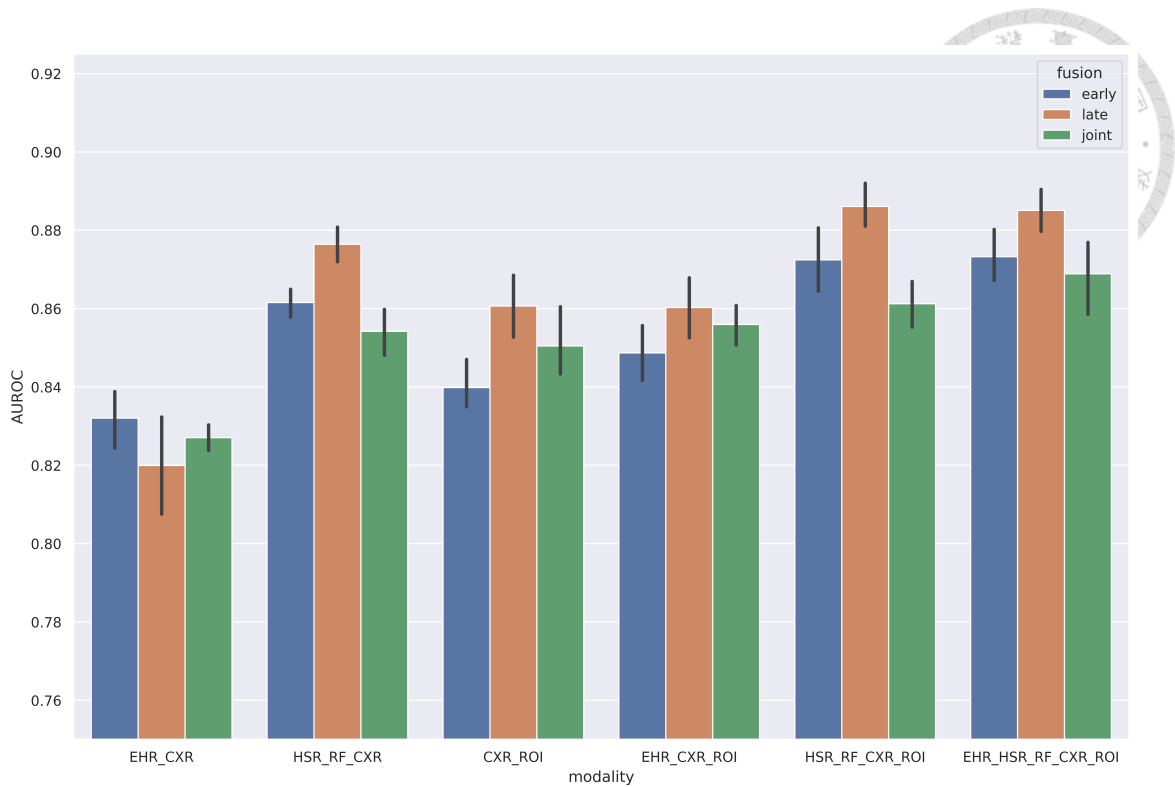


Figure 4.6: Box plots of AUROC of different fusion strategies

gain is the most commonly used and intuitive. The Gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. A higher value of gain, when compared to another feature, implies it is more important for generating a prediction. Because the randomness of data splitting and model training differ in the results of feature importance, I trained XGBoost for 100 iterations and average the feature importance to obtain more robust results. For each iteration, the dataset was randomly split into training and testing datasets with a proportion of 8:2 by patient level. Fig. 4.9 shows the top 20 important variables including `area_ratio`, `ngtdm_Coarseness_heart`, `firstorder_10Percentile_heart`, `width_ratio`, `age`, etc. It is worth noting that both patient information (`age`) and image information (`area_ratio`, `width_ratio`) are at the top of the rank, implying that multimodal prediction is beneficial to developing a better model. Although the feature importance from XGBoost reveals the black box of the model itself, the biggest shortage is that it only provides the magnitude of

Table 4.5: Model performances

| Modality           | AUC                    | Sensitivity     | Specificity     | PPV             | NPV             | Accuracy        |
|--------------------|------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Numerical Models   |                        |                 |                 |                 |                 |                 |
| EHR                | 0.7146 (0.0144)        | 0.6875 (0.0138) | 0.6577 (0.0382) | 0.8356 (0.0171) | 0.4541 (0.0223) | 0.6790 (0.0183) |
| HSR                | 0.7582 (0.0151)        | 0.6228 (0.1224) | 0.7500 (0.0739) | 0.8652 (0.0172) | 0.4501 (0.0586) | 0.6589 (0.0681) |
| RF                 | 0.8634 (0.0060)        | 0.7445 (0.0284) | 0.8481 (0.0299) | 0.9257 (0.0121) | 0.5685 (0.0201) | 0.7738 (0.0147) |
| EHR_HSR            | 0.8064 (0.0018)        | 0.7186 (0.0145) | 0.7788 (0.0136) | 0.8916 (0.0047) | 0.5228 (0.0103) | 0.7357 (0.0082) |
| EHR_RF             | <b>0.8660 (0.0057)</b> | 0.7224 (0.0320) | 0.8500 (0.0161) | 0.9241 (0.0074) | 0.5489 (0.0286) | 0.7586 (0.0220) |
| HSR_RF             | 0.8639 (0.0018)        | 0.7407 (0.0189) | 0.8423 (0.0110) | 0.9224 (0.0042) | 0.5627 (0.0169) | 0.7695 (0.0120) |
| EHR_HSR_RF         | 0.8650 (0.0043)        | 0.7475 (0.0180) | 0.8346 (0.0267) | 0.9198 (0.0106) | 0.5669 (0.0110) | 0.7722 (0.0071) |
| Image Models       |                        |                 |                 |                 |                 |                 |
| CXR                | 0.8241 (0.0150)        | 0.8076 (0.1696) | 0.5885 (0.2688) | 0.8468 (0.0669) | 0.6204 (0.1522) | 0.7455 (0.0581) |
| ROI                | <b>0.8490 (0.0063)</b> | 0.7643 (0.1225) | 0.7500 (0.1493) | 0.8932 (0.0472) | 0.5817 (0.0932) | 0.7602 (0.0472) |
| Early Fusion       |                        |                 |                 |                 |                 |                 |
| EHR_CXR            | 0.8320 (0.0088)        | 0.8456 (0.0330) | 0.6308 (0.0675) | 0.8535 (0.0185) | 0.6208 (0.0342) | 0.7847 (0.0132) |
| HSR_RF_CXR         | 0.8615 (0.0048)        | 0.8190 (0.0333) | 0.7385 (0.0570) | 0.8887 (0.0180) | 0.6197 (0.0247) | 0.7962 (0.0083) |
| CXR_ROI            | 0.8398 (0.0081)        | 0.8365 (0.0409) | 0.6308 (0.0622) | 0.8521 (0.0163) | 0.6080 (0.0371) | 0.7782 (0.0176) |
| EHR_CXR_ROI        | 0.8487 (0.0089)        | 0.8304 (0.0232) | 0.6923 (0.0366) | 0.8725 (0.0102) | 0.6189 (0.0224) | 0.7913 (0.0079) |
| HSR_RF_CXR_ROI     | 0.8724 (0.0108)        | 0.8312 (0.0525) | 0.7365 (0.0422) | 0.8890 (0.0111) | 0.6407 (0.0679) | 0.8044 (0.0295) |
| EHR_HSR_RF_CXR_ROI | <b>0.8733 (0.0081)</b> | 0.8274 (0.0517) | 0.7404 (0.0739) | 0.8911 (0.0238) | 0.6354 (0.0507) | 0.8027 (0.0197) |
| Joint Fusion       |                        |                 |                 |                 |                 |                 |
| EHR_CXR            | 0.8270 (0.0042)        | 0.8768 (0.0681) | 0.5404 (0.1391) | 0.8312 (0.0317) | 0.6578 (0.0807) | 0.7815 (0.0130) |
| HSR_RF_CXR         | 0.8542 (0.0080)        | 0.8076 (0.0606) | 0.7173 (0.0939) | 0.8806 (0.0268) | 0.6035 (0.0471) | 0.7820 (0.0190) |
| CXR_ROI            | 0.8504 (0.0117)        | 0.8540 (0.0449) | 0.5865 (0.0905) | 0.8405 (0.0228) | 0.6210 (0.0526) | 0.7782 (0.0175) |
| EHR_CXR_ROI        | 0.8560 (0.0070)        | 0.8920 (0.0586) | 0.5462 (0.1515) | 0.8358 (0.0392) | 0.6814 (0.0666) | 0.7940 (0.0179) |
| HSR_RF_CXR_ROI     | 0.8612 (0.0071)        | 0.8700 (0.0603) | 0.6231 (0.1136) | 0.8564 (0.0333) | 0.6657 (0.0519) | 0.8000 (0.0136) |
| EHR_HSR_RF_CXR_ROI | <b>0.8689 (0.0109)</b> | 0.8669 (0.0490) | 0.6788 (0.1206) | 0.8746 (0.0353) | 0.6752 (0.0500) | 0.8136 (0.0235) |
| Late Fusion        |                        |                 |                 |                 |                 |                 |
| EHR_CXR            | 0.8199 (0.0158)        | 0.8190 (0.1135) | 0.6173 (0.2223) | 0.8532 (0.0576) | 0.6092 (0.0992) | 0.7619 (0.0296) |
| HSR_RF_CXR         | 0.8764 (0.0061)        | 0.8068 (0.1169) | 0.7788 (0.1281) | 0.9076 (0.0384) | 0.6437 (0.1114) | 0.7989 (0.0524) |
| CXR_ROI            | 0.8606 (0.0100)        | 0.8068 (0.1318) | 0.6808 (0.2188) | 0.8770 (0.0662) | 0.6231 (0.1125) | 0.7711 (0.0345) |
| EHR_CXR_ROI        | 0.8603 (0.0100)        | 0.8160 (0.1072) | 0.6923 (0.1874) | 0.8787 (0.0554) | 0.6262 (0.0910) | 0.7809 (0.0266) |
| HSR_RF_CXR_ROI     | <b>0.8861 (0.0069)</b> | 0.8038 (0.0993) | 0.7731 (0.1396) | 0.9053 (0.0444) | 0.6284 (0.0830) | 0.7951 (0.0394) |
| EHR_HSR_RF_CXR_ROI | 0.8851 (0.0068)        | 0.8091 (0.0928) | 0.7846 (0.1304) | 0.9097 (0.0414) | 0.6367 (0.0810) | 0.8022 (0.0374) |

Model performance on the held-out test set using a probability threshold of 0.5.

Data in parentheses are standard deviations.

the feature without directions. Take the top 5 features. For example, a person with a little medical background can understand that people with advanced age or enlarged heart size are more likely to have AHF. However, it is more difficult to identify the direction of how `ngtdm_Coarseness_heart` and `firstorder_10Percentile_heart` affect the prediction results. To overcome this shortage, I implemented SHAP analysis.

SHAP (SHapley Additive exPlanations) [29] is a model-agnostic method based on the optimal Shapley values from game theory to explain individual predictions. SHAP uses a linear sum of feature contributions to explain the difference between the overall outcome and a specific prediction. Besides the interpretation of a single sample, the general feature importance can also be measured by summing all the absolute values of Shapley

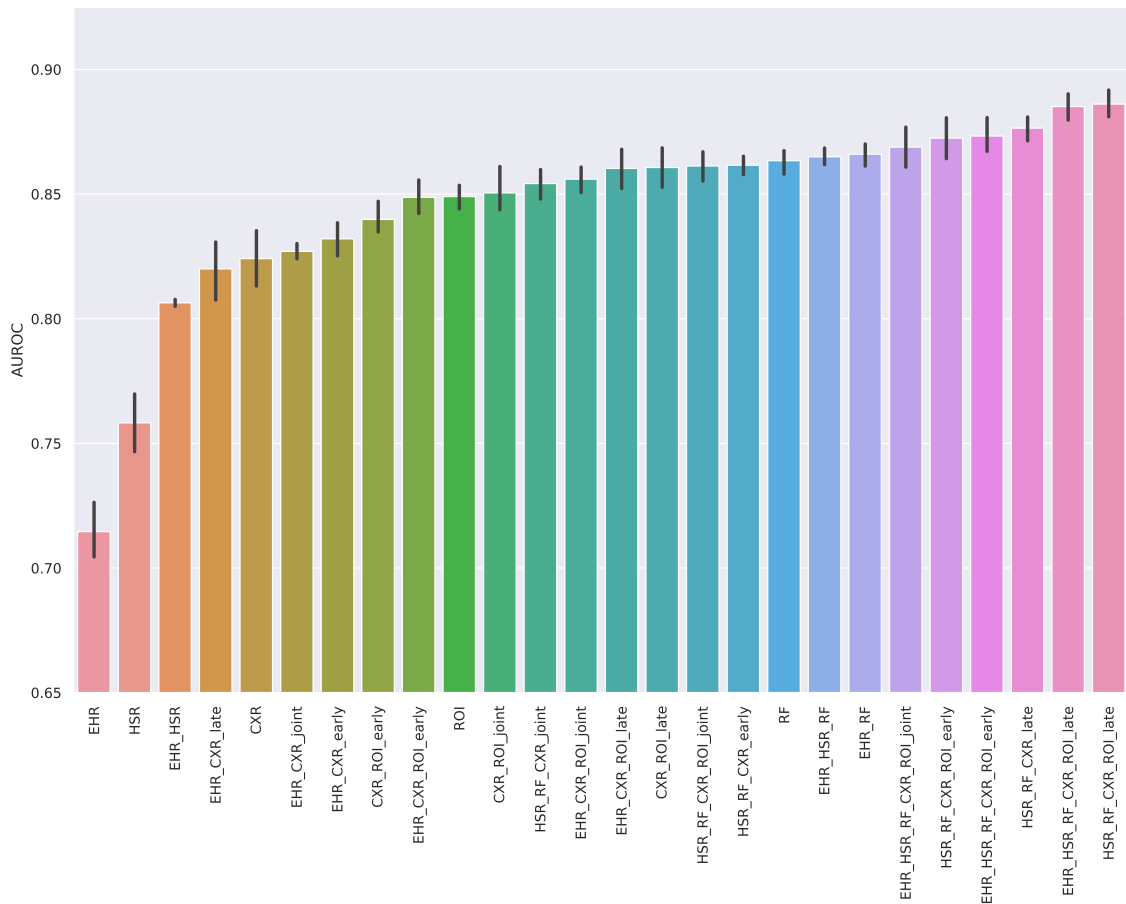


Figure 4.7: Box plots of different modality combinations and fusion strategies

values for each feature and instance. I used SHAP to explain the 317 numerical features. The SHAP bar plot (Fig. 4.10) shows the top 20 features based on the SHAP value. Moreover, the SHAP summary plot provides the magnitude and direction of how a feature affects the result. As shown in Fig. 4.11, the x-axis represents the SHAP value, the y-axis represents feature names descendingly ranked by their sum of absolute SHAP values, and the color shows the relative magnitude within a feature (the redder, the higher). For example, age is the most crucial feature since it's at the top. And the color implies that the older patient has a higher risk of anomaly NT-proBNP. The area and width ratios are the second and third critical features reflecting that large heart size relates to a bad heart condition. Since AHF is a cardio-related disease, it is reasonable that most top-ranked radiomic features are derived from the heart ROI. All the findings from the SHAP analysis are aligned



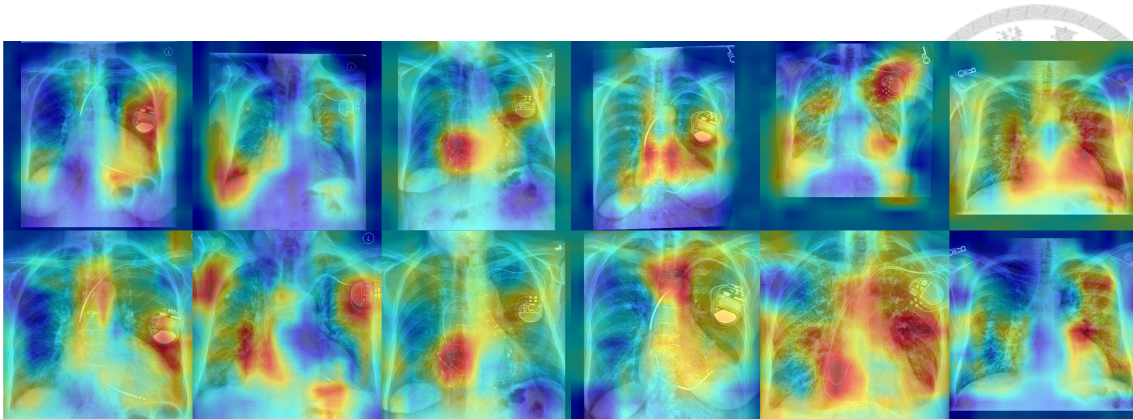


Figure 4.8: Examples of heatmaps

with clinical knowledge. And further analysis of HSR and RF will be demonstrated in Sec. 4.8 and Sec. 4.9, respectively.

## 4.8 Analysis of Heart-size Ratios

Three HSRs (width, height, and area ratio) were extracted in Sec 3.5.1 using patients' CXRs and the corresponding lung-heart masks. The histograms of the three ratios are displayed in Fig. 4.12. The top row shows the overall distributions, and the bottom row shows the distributions stratified by the labels. The width ratio, or the CT ratio, is the most common indicator to estimate heart size in CXR. The mean and median width ratios in our dataset are 0.5616 and 0.5593, respectively. These two numbers are larger than 0.5, meaning the patients in the study population mostly have cardiomegaly. Since patients with AHF tend to have larger heart sizes, this result also aligns with the fact that the number of AHF likely patients is more than AHF unlikely patients in the study population. Visually, the distributions of AHF likely patients and AHF unlikely patients in width ratio and area ratio are more separate than the ones in height ratio. When a feature's distributions stratified by the label are apart, the feature is a good predictor and has more ability to classify the target. As shown in Fig. 4.9 and Fig. 4.10, area ratio and width ratio are the

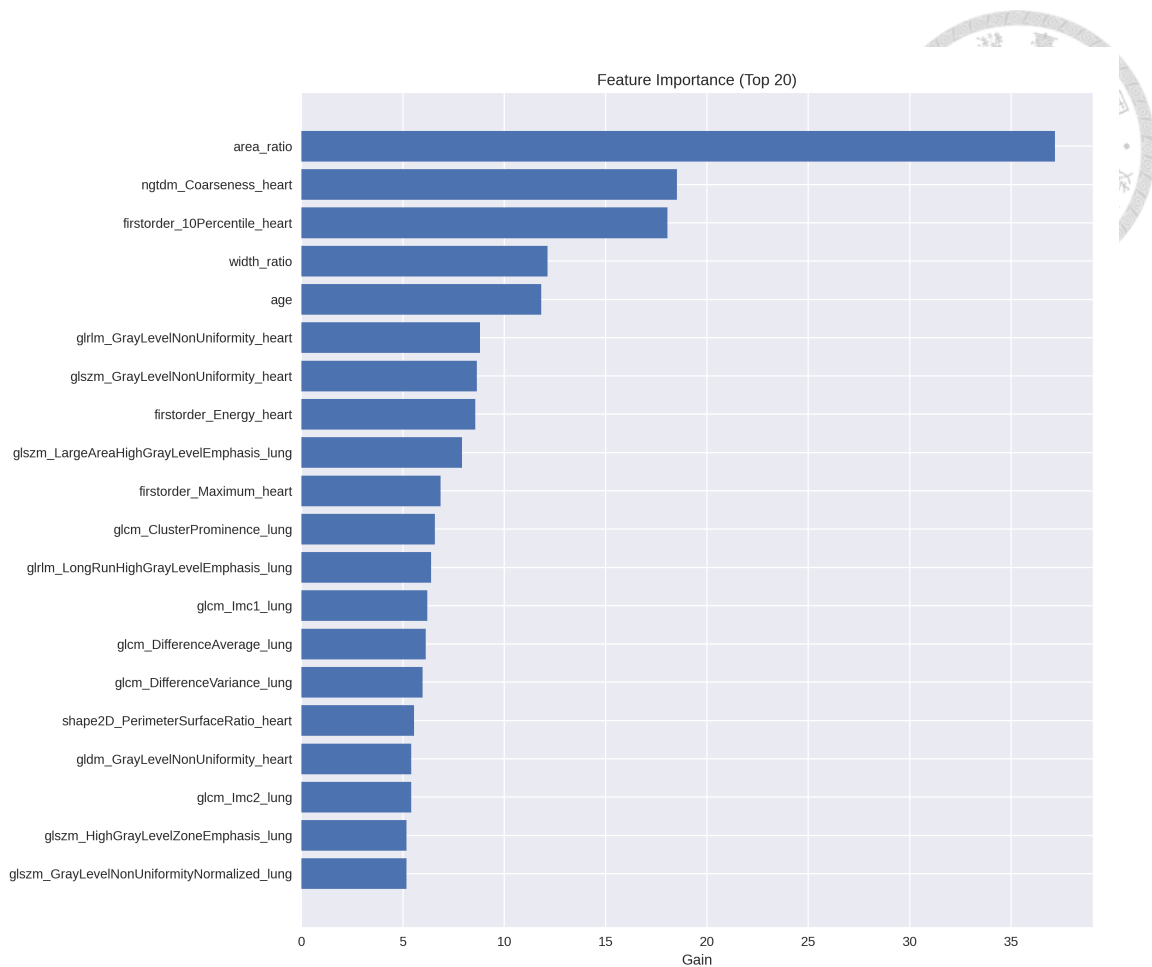


Figure 4.9: Feature importance (gain)

second and third important features among all numerical features, but height ratio is not included in the top 20s. The findings from the distributions of HSRs agree with the results of feature important analysis, showing that width ratio and area ratio are indeed more crucial than height ratio, and the size of the heart does play a big role in diagnosing AHF.

## 4.9 Analysis of Radiomic Features

A total of 306 RFs were extracted in Sec. 3.5.2, 102 features from the heart region, 102 features from the lung region, and 102 features from the union region of the heart and lung. Fig. 4.13 demonstrates how RFs reflect on images. The five most important RFs

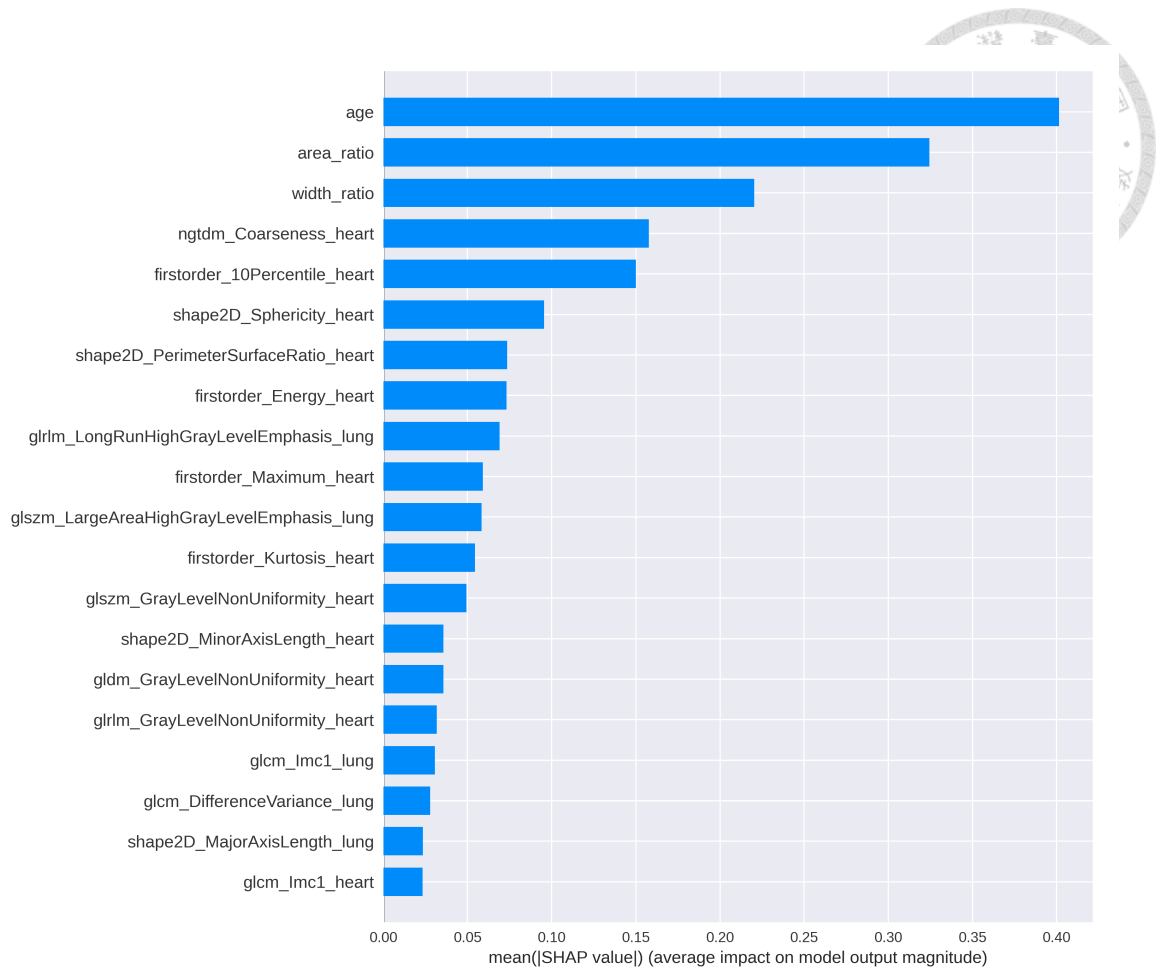


Figure 4.10: SHAP bar plot

(ngtdm\_Coarseness\_heart, glrlm\_LongRunHighGrayLevelEmphasis\_lung, shape2D\_Sphericity\_heart, shape2D\_PerimeterSurfaceRatio\_heart, and firstorder\_Energy\_heart) were selected according to the rank from SHAP analysis in Fig. 4.10, and then for each important RF, I displayed the images with 10%, 25%, 50%, 75%, and 90% percentile of that given RF.

According to PyRadiomics documentation [24], “coarseness is a measure of the average difference between the center voxel and its neighborhood and indicates the spatial rate of change. A higher value indicates a lower spatial change rate and a locally more uniform texture”. “Long run high gray level emphasis measures the joint distribution of long-run lengths with higher gray-level values”. “Sphericity is the ratio of the perimeter

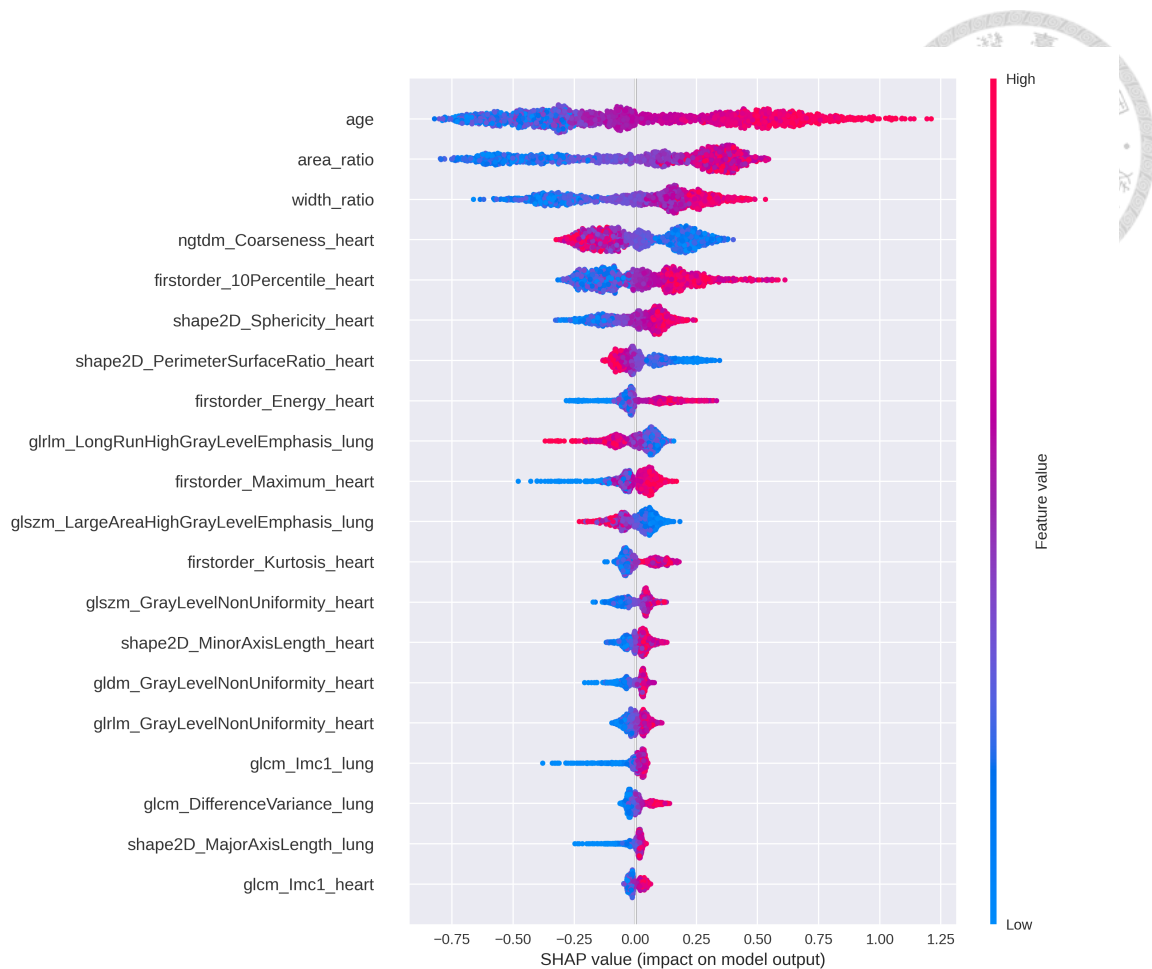


Figure 4.11: SHAP summary plot

of the tumor region to the perimeter of a circle with the same surface area as the tumor region and, therefore, a measure of the roundness of the shape of the tumor region relative to a circle. A value of 1 indicates a perfect circle”. “Perimeter-to-Surface ratio of a lower value indicates a more compact (circle-like) shape”. And “Energy is a measure of the magnitude of voxel values in an image. A larger value implies a greater sum of the squares of these values”.

Patients with AHF usually have higher intracardiac pressure. The extra pressure in the left and right atriums makes the heart look more circular than normal. Therefore, higher sphericity and lower perimeter-to-surface ratio are related to a higher risk of having AHF. Alveolar edema results in a large white area in the lung region, and it is a common

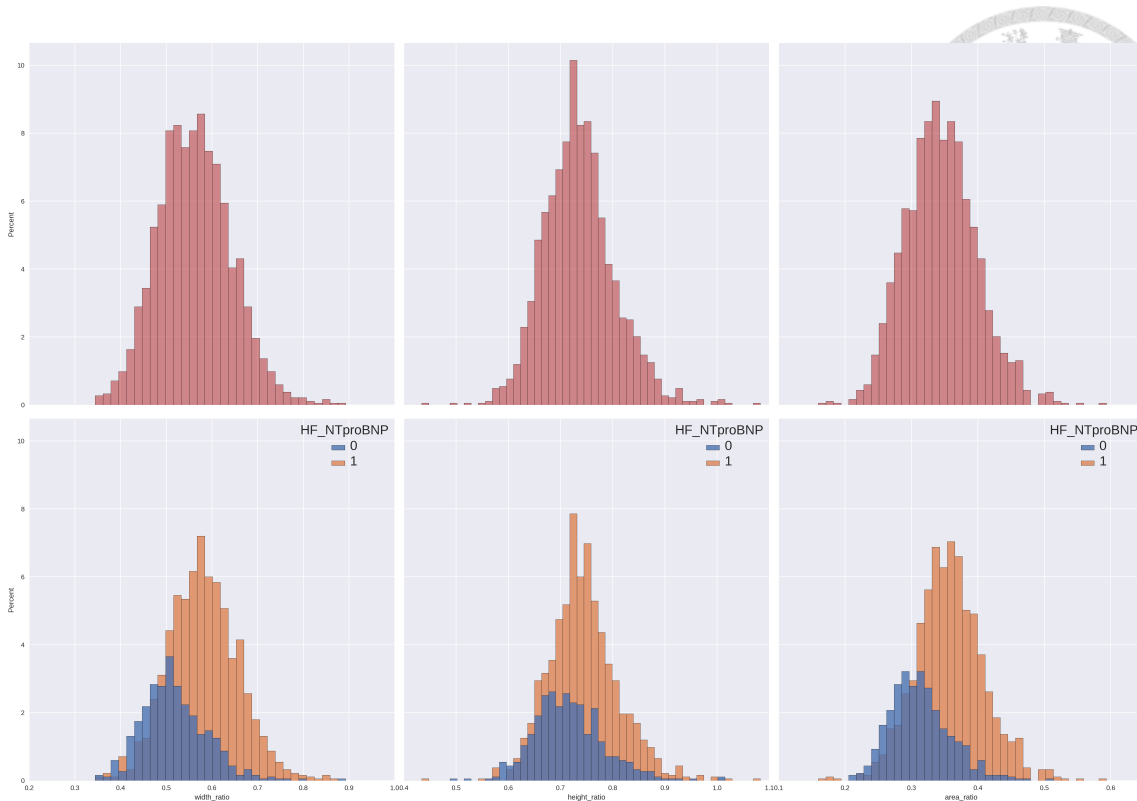


Figure 4.12: Distributions of heart-size ratios

CXR finding among AHF patients. Namely, the CXRs of AHF patients have less concentration of long-run lengths of whiter pixels. The lower the long-run high gray level emphasis, the higher the risk of having AHF reflects this clinical understanding.

The visual change according to RFs is mostly consistent with clinical understanding, and these findings provide a quantitative fashion to explain the behavior of machine learning models and make the model more human-reliable.

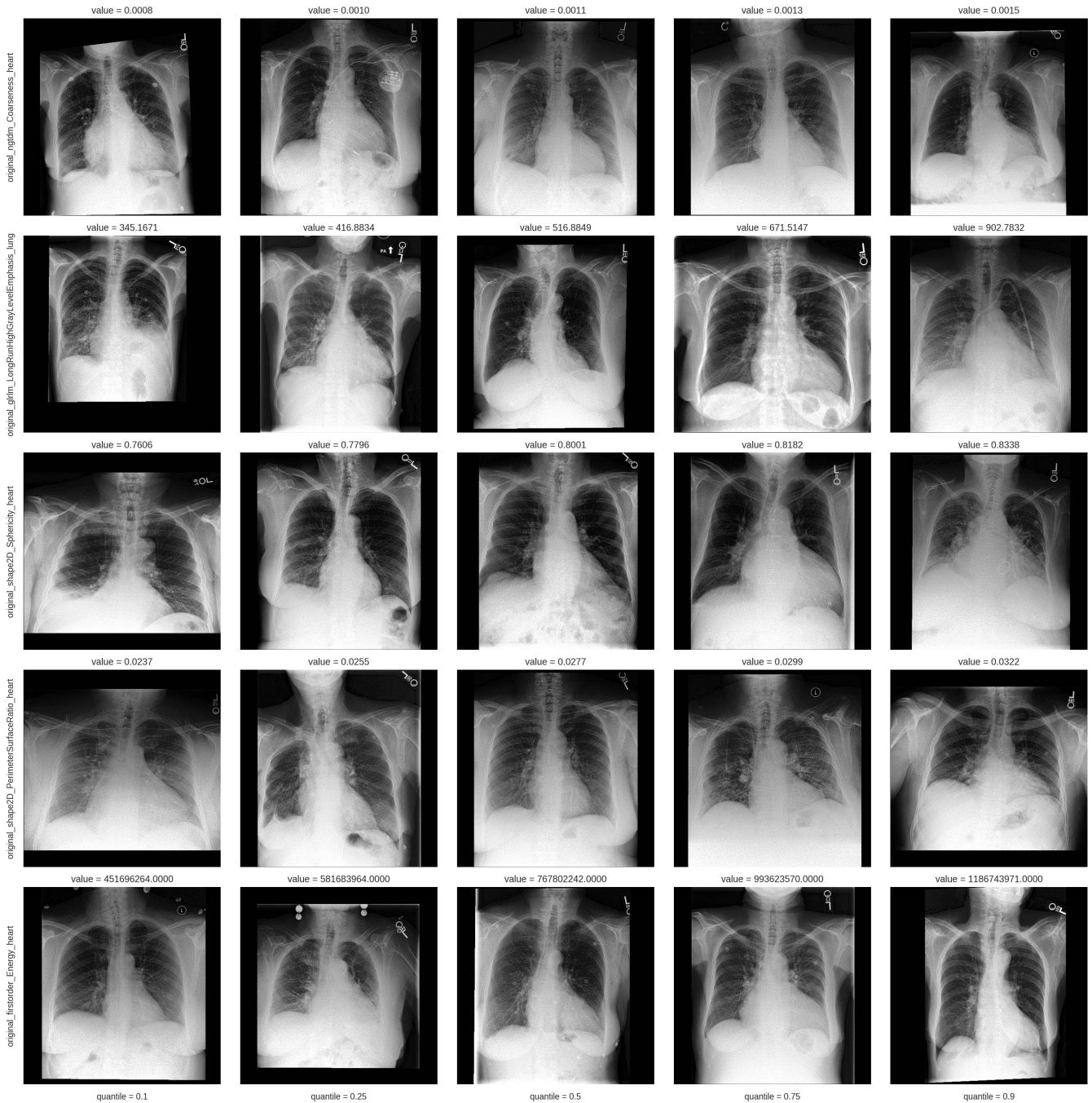


Figure 4.13: Examples of radiomic features. The rows represents the five most important RFs. The columns, from left to right, are 10%, 25%, 50%, 75%, and 90% percentile of the corresponding RF of the row





## Chapter 5 Limitations

This study proposed a multimodal deep learning model to predict the NT-proBNP level in ED using patients' EHR and CXR. Although the proposed model achieved remarkable classification performance, some limitations were found that require careful consideration.

### 5.1 Label Uncertainty

After analyzing failure cases, I noticed that some misclassified cases were peculiar and were almost inevitable to avoid. As shown in Fig. 5.1, both images visually contain strong characteristics implying high NT-proBNP levels. For example, there are abnormally large hearts in CXR, and the patient in the left image wears a pacemaker. However, their NT-proBNP values are below 300 ng/L, 240, and 218 ng/L, respectively. In our opinion, two potential possibilities lead to the conflict outcomes of predictors and ground truths. First, since chest radiography and NT-proBNP data were merged with the smallest time difference, some unlearnable biases may exist if the patient's health condition changed dramatically between the two examinations. Second, the health condition itself is inherently uncertain. The same cause may lead to different results for different people.

Collecting more features about patients' health conditions may be able to address this



type of problem. Assume that there is a missing but important confounder heavily causing the label uncertainty. If one can input that missing feature into the proposed multimodal model, the model may learn the relationship between the missing feature and the true label. However, the eight EHR (age, gender, temperature, heart rate, respiratory rate, O<sub>2</sub> saturation, systolic blood pressure, diastolic blood pressure) and CXR were chosen as the input modality because of their accessibility in ED. Therefore, if obtaining the hypothetical missing feature costs too much, it is unsuitable for our AHF screening scenario.

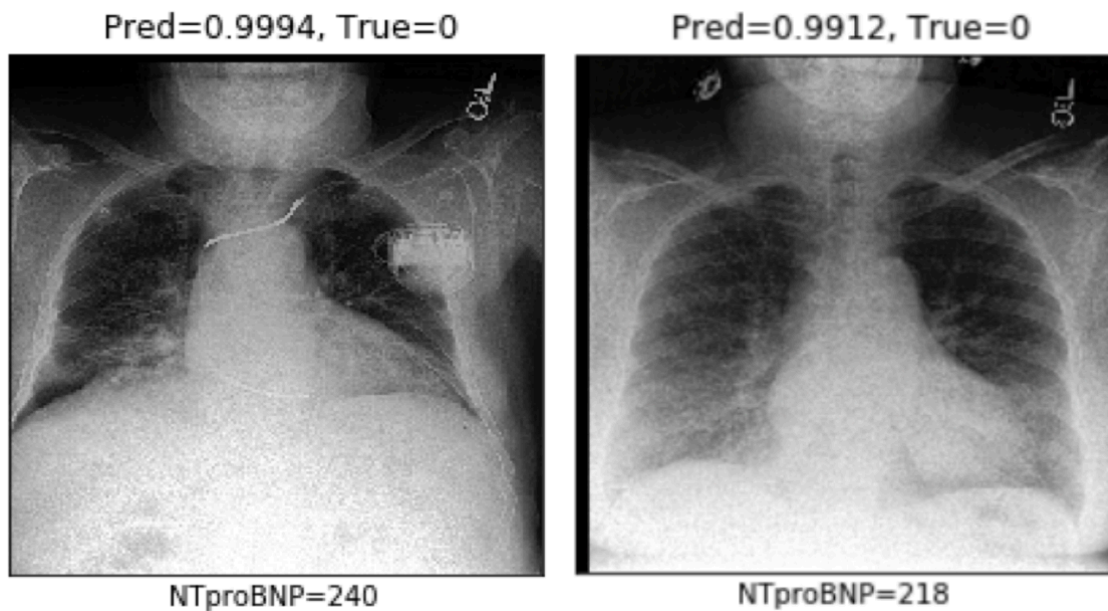
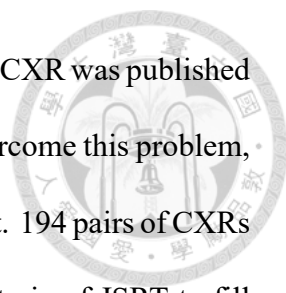


Figure 5.1: Demonstration of label uncertainty

## 5.2 Segmentation Models

For some CXRs, their lung-heart masks are challenging to be accurately segmented. As demonstrated in Fig. 3.12, the bottom corners of the CXR were barely detectable, and the lung region's whiteness tends to be misclassified as the heart region. And since lung-heart masks are one of the most significant components in our modeling pipeline, any subtle difference matters.



There are two main reasons causing the limitation. First, MIMIC-CXR was published with weak labels for classifying 14 common thoracic diseases. To overcome this problem, I utilize 247 pairs of CXRs and lung-heart masks from the JSRT dataset. 194 pairs of CXRs and lung-heart masks were manually labeled following the same criteria of JSRT to fill the gap of the domain shift between the MIMIC-CXR and the JSRT dataset. The lung-heart segmentation model performed surprisingly well, but since manually annotating the semantic segmentation task is both labor and time-consuming, not all the lung-heart masks in the study population could be annotated. Therefore, some edge cases can not be handled well enough. Second, image quality and health conditions themselves are tremendously diverse. Unlike JSRT, MIMIC-CXR collected CXRs from the real clinical fields for years, so the image quality and patients' health conditions have width distributions. For example, Some CXRs have clear borders, and some CXRs have blur edges. Some CXRs are straight, and some CXRs are tilted. Some CXRs are in high resolution, and some CXRs are in low resolution. Due to the variety of CXRs, developing a well generalizable lung-heart segmentation training from the small dataset is hard.

I believe that a high-quality lung-heart segmentation model provides great benefits. The benefits come from their derived features. Combining patients' CXRs and the corresponding lung-heart masks, clinically meaningful features such as three HSRs and 306 RFs were extracted for each pair of CXR and mask. These extracted numerical features are human-recognizable and open the black box of machine learning models. The magnitude and the direction of how a feature affects the prediction can be quantified using XGboost and SHAP analysis. Moreover, the lung-heart masks even localized the thoracic region (ROI) and let the image models focus on the important region instead of the AHF irrelevant area.

A potential way to generate high-quality lung-heart masks in the future is to build up a system that continuously updates the segmentation model and introduces the idea of active learning. When updating the models, the images requiring manual annotation have the most incorrect or visual differences. However, the potential update is not in the scope of this study.

### 5.3 Mask Post-processing Algorithms

In Sec. 3.4.2, I developed two mask post-processing algorithms to maintain the anatomical characteristics of the heart and lungs. The main purposes of the two algorithms are to recover and approximate the undetectable lung-heart mask due to the limitation of the segmentation model. In this section, I will discuss the cases in which the two algorithms may not be accurate enough.

The first mask post-processing algorithm deletes fragments by replacing the original value with the most frequent value in their edges. When a heart fragment occurs between lung masks and background masks, the algorithm may replace the wrong value since the number of wrong pixels may outnumber the correct ones and become the mode of the edge.

The second mask post-processing algorithm was developed assuming that human chests are bilateral symmetry. Utilizing the property, the algorithm identifies left and right clavicles and then rotates, shifts, and mirrors the image according to the centers of the two clavicles. As shown in Fig. 3.10, if the patient's body was tilted when taking CXR, the assumption of bilateral symmetry is violated, resulting in a bad approximation of the lung mask. Another failure case is when the entire heart mask is undetectable or is deleted by

the first algorithm. In this case, the heart mask can not be recovered.

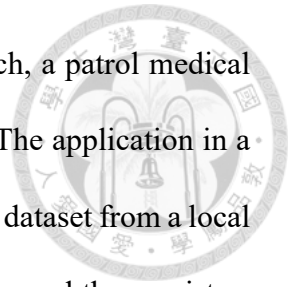


## 5.4 Generalization

In this study, MIMIC-IV serves as the core dataset. The data in MIMIC-IV were collected by the patients admitted to an academic medical center in Boston, MA, USA, between 2008 and 2019. Although MIMIC-IV is currently the largest open-source dataset containing 377,110 CXRs from 64,588 patients, it is still a local uni-centered dataset with little diversity in terms of ethnicity. Thus, our trained multimodal deep learning model may fail to perform as well as the performance in the study population. The performance drop will depend on the divergence between MIMIC-IV and the target dataset. The more the divergence, the worse I expect the model to perform. Our study has demonstrated the benefits of combining EHR and CXR when screening AHF in ED using open-source data. In the future, I plan to cooperate with National Taiwan University Hospital (NTUH) and retrained the model using the data from NTUH. I hope to implement our model into a prospective scenario and maximize the value of medical data.

Currently, our study focus on predicting AHF in ED for patients with EHR and CXR. This scenario is where NT-proBNP is most commonly used. However, in our opinion, the application of predicting NT-proBNP can be more than that. There are two possible situations that can be benefited from the NT-proBNP prediction model. One situation is in the hospital without the ability to execute the NT-proBNP test. As shown in Fig. 1.1, our model can be embedded into the current AHF diagnosis workflow. Even if a patient could not take NT-proBNP testing in a hospital, the patient's probability of having AHF can still be approximated by our model. The other scenario is in a remote area. Since the eight

EHRs in our study and CXR are cheap data sources within easy reach, a patrol medical vehicle with an X-ray device can carry out a community screening. The application in a remote area is inexecutable even though the model is retrained using a dataset from a local hospital because the data from MIMIC-IV were collected in hospitals, and there exists a huge difference between the distribution of patients and the distribution of the general population. Therefore, the data from general health examinations are necessary to fulfill the great picture of the application of NT-proBNP prediction.





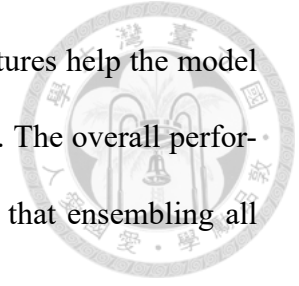
## Chapter 6 Conclusion

In this paper, I proposed a multimodal deep learning model that combined the CXR and EHR to screen AHF in ED. The prediction target is a binary label indicating whether the value of NT-proBNP is greater than 300 ng/L. The model aims to be integrated into the current AHF diagnosis workflow and release the potential bottleneck as no single examination can independently reliably diagnose or rule out AHF in the workflow.

The input modalities are eight EHRs and frontal CXRs, which are the most common data in ED. Once images are input into the modeling pipeline, the lung-heart mask generator initially identified the lung and heart region by the lung-heart segmentation model and remained the anatomical characteristics of the heart and lungs by two mask post-processing algorithms. After obtaining the lung-heart masks, three heart-size ratios and 306 radiomic features were extracted by predefined formulas. These manually extracted features provide interpretability and clinical meanings. Moreover, ROIs were also extracted to eliminate noises in original CXRs and make the image model focus on the important regions. Eventually, all elements above, including 317 numerical features and two images, were fused to output the final prediction.

The ablation study illustrates that EHRs improve model performance from AUROC 0.8241 to 0.8320. Adding HSRs and RFs to the model also boosts prediction performance

from AUC 0.8241 to 0.8615. The results verify that multimodal features help the model learn the underlying relationship between input modalities and AHF. The overall performance of late fusion outperformed early and joint fusion, showing that ensembling all single-modal models outputs a more robust and accurate prediction.



The heatmap analysis shows that the two image models make decisions that mostly depend on heart regions or the pacemakers if exist. The importance of the XGBoost features and SHAP analysis highlighted the most significant features, including age, width and area ratios, and RFs related to heart size and lung texture. These findings are aligned with our clinical background knowledge.

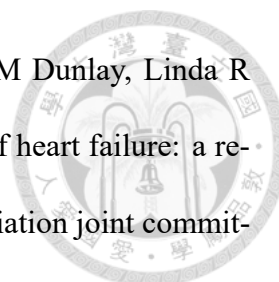
Using the MIMIC-IV and MIMIC-CXR, this study illustrated that the proposed method achieved the highest classification performance when combining all modalities.



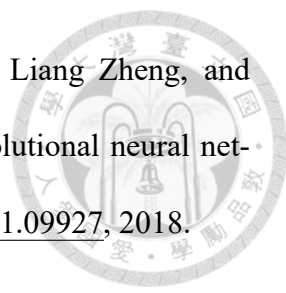
## References

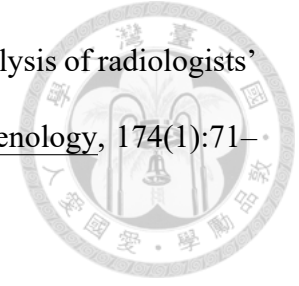
- [1] Mihai Gheorghide, Faiez Zannad, George Sopko, Liviu Klein, Ileana L Piña, Marvin A Konstam, Barry M Massie, Edmond Roland, Shari Targum, Sean P Collins, et al. Acute heart failure syndromes: current state and framework for future research. Circulation, 112(25):3958–3968, 2005.
- [2] Alan B Storrow, Cathy A Jenkins, Wesley H Self, Pauline T Alexander, Tyler W Barrett, Jin H Han, Candace D McNaughton, Benjamin S Heavrin, Mihai Gheorghide, and Sean P Collins. The burden of acute heart failure on us emergency departments. JACC: Heart Failure, 2(3):269–277, 2014.
- [3] Jennifer L Martindale, Abel Wakai, Sean P Collins, Phillip D Levy, Deborah Diercks, Brian C Hiestand, Gregory J Fermann, Ian Desouza, and Richard Sinert. Diagnosing acute heart failure in the emergency department: a systematic review and meta-analysis. Academic Emergency Medicine, 23(3):223–242, 2016.
- [4] Emmert Roberts, Andrew J Ludman, Katharina Dworzynski, Abdallah Al-Mohammad, Martin R Cowie, John JV McMurray, and Jonathan Mant. The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. BMJ, 350, 2015.
- [5] Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun,





- Monica M Colvin, Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, et al. 2022 aha/acc/hfsa guideline for the management of heart failure: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. Journal of the American College of Cardiology, 79(17):e263–e421, 2022.
- [6] Jarrel CY Seah, Jennifer SN Tang, Andy Kitchen, Frank Gaillard, and Andrew F Dixon. Chest radiographs in congestive heart failure: visualizing neural network learning. Radiology, 290(2):514–522, 2019.
- [7] Takuya Matsumoto, Satoshi Kodera, Hiroki Shinohara, Hirotaka Ieki, Toshihiro Yamaguchi, Yasutomi Higashikuni, Arihiro Kiyosue, Kaoru Ito, Jiro Ando, Eiki Takimoto, et al. Diagnosing heart failure from chest x-ray images using deep learning. International Heart Journal, 61(4):781–786, 2020.
- [8] Sarah Jabbour, David Fouhey, Ella Kazerooni, Jenna Wiens, and Michael W Sjong. Combining chest x-rays and electronic health record (ehr) data using machine learning to diagnose acute respiratory failure. Journal of the American Medical Informatics Association, 29(6):1060–1068, 2022.
- [9] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225, 2017.
- [10] Han Liu, Lei Wang, Yandong Nan, Faguang Jin, Qi Wang, and Jiantao Pu. Sdfn: Segmentation-based deep fusion network for thoracic disease classification in chest x-ray images. Computerized Medical Imaging and Graphics, 75:66–73, 2019.

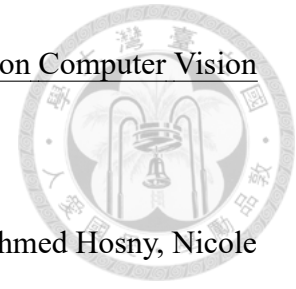
- 
- [11] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927, 2018.
- [12] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics, 23(2):538–546, 2018.
- [13] Jordan Yap, William Yolland, and Philipp Tschandl. Multimodal skin lesion classification using deep learning. Experimental Dermatology, 27(11):1261–1267, 2018.
- [14] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. PhysioNet. Available online at: <https://physionet.org/content/mimiciv/1.0/>(accessed August 23, 2021), 2020.
- [15] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation, 101(23):e215–e220, 2000.
- [16] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042, 2019.
- [17] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with



and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American Journal of Roentgenology, 174(1):71–74, 2000.

- [18] Boris Sekachev, M Nikita, and Z Andrey. Computer vision annotation tool: a universal approach to data annotation (2019). Available at: <https://github.com/openvinotoolkit/cvat>.
- [19] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing, 39(3):355–368, 1987.
- [20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4700–4708, 2017.
- [21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-assisted Intervention, pages 234–241. Springer, 2015.
- [22] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 801–818, 2018.
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning

for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.



- [24] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. Cancer Research, 77(21):e104–e107, 2017.
- [25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- [26] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digital Medicine, 3(1):1–9, 2020.
- [27] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2921–2929, 2016.
- [28] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, 2016.
- [29] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 2017.





## Appendix A — Figures

### A.1 Flowchart of Inclusion and Exclusion

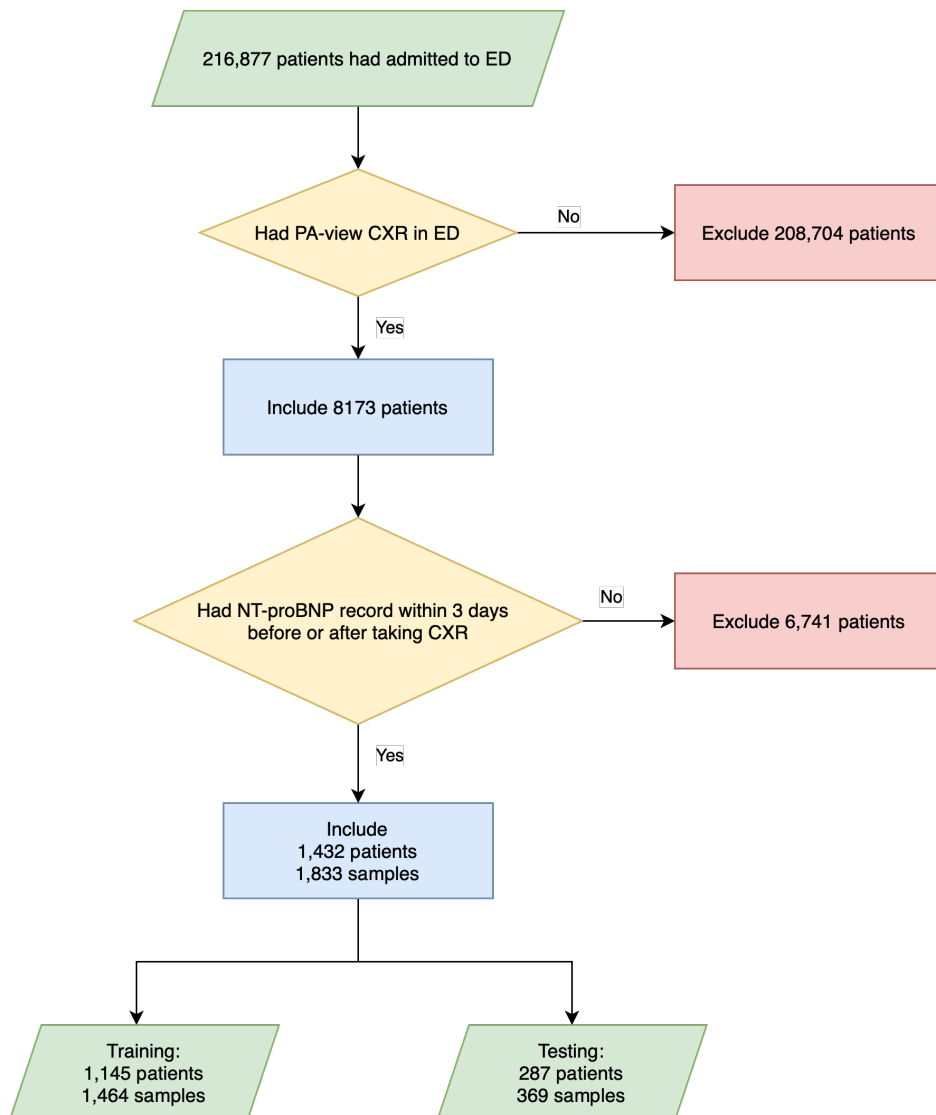


Figure A.1: Flowchart of inclusion and exclusion

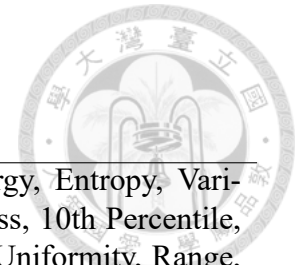




## Appendix B — Tables

### B.1 Radiomic Features





| Feature Classes      | Feature Names   |
|----------------------|---|
| First Order Features | Maximum, Median, Minimum, Mean Energy, Entropy, Variance, Kurtosis, Root Mean Square, Skewness, 10th Percentile, 90th Percentile, Mean Absolute Deviation, Uniformity, Range, Robust Mean Absolute Deviation, Total Energy, Interquartile Range   |
| Shape Features (2D)  | Volume, Elongation, Surface Area, Max 2D Diameter, Mesh Volume, Major Axis Length, Max 2D Diameter Row, Max 2D Diameter Column, Surface Volume Ratio, Sphericity, Minor Axis Length   |
| GLCM Features        | Contrast, Correlation, Autocorrelation, Cluster Tendency, Sum Average, Sum Entropy, Sum Squares, Difference Average, Difference Variance, Difference Entropy, Cluster Prominence, Cluster Shade, Maximum Probability, Inverse Difference Moment, Informational Measure of Correlation 1/2, Inverse Difference Moment Normalized, Inverse Difference Normalized, Inverse Difference, Inverse Variance, Maximal Correlation Coefficient, Joint Average, Joint Energy, Joint Entropy |
| GLSZM Features       | Gray-Level Non-Uniformity, Gray-Level Non-Uniformity Normalized, Gray-Level Non-Uniformity Normalized, High Gray-Level Zone Emphasis, Large Area Emphasis, Large Area High Gray-Level Emphasis, Large Area Low Gray-Level Emphasis, Low Gray-Level Zone Emphasis, Size Zone Non-Uniformity, Size Zone Non-Uniformity Normalized, Small Area Emphasis, Small Area High Gray-Level Emphasis, Small Area Low Gray-Level Emphasis, Zone Entropy, Zone Percentage, Zone Variance       |
| GLRLM Features       | Gray-Level Non-uniformity, Gray-Level Non-uniformity Normalized, Gray-Level Variance, High Gray-Level Run Emphasis, Long Run Emphasis, Long Run High Gray-Level Emphasis, Long Run Low Gray-Level Emphasis, Low Gray-Level Run Emphasis, Run Entropy, Run Length Non-Uniformity, Run Length Non-Uniformity Normalized, Run Percentage, Run Variance, Short Run Emphasis, Short Run High Gray-Level Emphasis, Short Run Low Gray-Level Emphasis                                    |
| NGTDM Features       | Coarseness, Contrast, Busyness, Complexity, Strength  |
| GLDM Features        | Dependence Entropy, Dependence Non-Uniformity, Dependence Non-Uniformity Normalized, Dependence Variance, Gray-Level Non-Uniformity, Gray-Level Variance, High Gray-Level Emphasis, Large Dependence Emphasis, Large Dependence High Gray-Level Emphasis, Large Dependence Low Gray-Level Emphasis, Low Gray-Level Emphasis, Small Dependence Emphasis, Small Dependence High Gray-Level Emphasis, Small Dependence Low Gray-Level Emphasis                                       |