UNIVERSITY OF CALIFORNIA SAN DIEGO

Deep Learning Radiographic Assessment of Pulmonary Edema from Serum Biomarkers

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Computer Science

by

Justin Huynh

Committee in charge:

Professor Albert Hsiao, Chair
Professor Manmohan Chandraker, Co-Chair
Professor Laurel Riek

2022

The Thesis of Justin Huynh is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis, in full, is a reprint of the material as it appears in IEEE Access 2022. Huynh, Justin; Masoudi, Samira; Noorbaksh, Abraham; Mahmoodi, Amin; Kligerman, Seth; Yen, Andrew; Jacobs, Kathleen; Hahn, Lewis; Hasenstab, Kyle; Pazzani, Michael; Hsiao, Albert. *Deep Learning Radiographic Assessment of Pulmonary Edema: Optimizing Clinical Performance, Training with Serum Biomakers*. IEEE Access, 2022. The thesis author was the first author of this paper.

The author acknowledges research grant support from NSF RAPID 2026809 and DARPA N00173-21-Q-0141, in-kind support from Microsoft AI for Health, NVIDIA and Amazon Web Services for this work.

The author would like to especially thank research advisor and thesis chair Dr. Albert Hsiao for his guidance throughout the last two years. The author would like to thank Dr. Samira Masoudi for her close collaboration on every aspect of this work. The author would also like to thank all collaborators whom he had the pleasure of working with including Dr. Kyle Hasenstab, Dr. Abraham Noorbaksh, Dr. Michael Pazzani, Amin Mahmoodi, Dr. Lewis Hahn, Dr. Kathleen Jacobs, Dr. Andrew Yen, Dr. Seth Kligerman, Dr. Evan Masutani, Dr. Tara Retson, Dr. Sophie You, Dr. Shanmukha Srinivas, Rahul Chandruptala, Brendan Crabb, and others. The author would like to thank Professor Manmohan Chandraker for an amazing computer vision course and agreeing to be a thesis co-chair despite having different research areas. The author would like to thank Professor Laurel Riek for being on the thesis committee.

ABSTRACT OF THE THESIS

Deep Learning Radiographic Assessment of Pulmonary Edema from Serum Biomarkers

by

Justin Huynh

Master of Science in Computer Science

University of California San Diego, 2022

Professor Albert Hsiao, Chair
Professor Manmohan Chandraker, Co-Chair

A major obstacle when developing convolutional neural networks (CNNs) for medical imaging is the acquisition of training labels: Most current approaches rely on manual class labels from physicians, which may be challenging to obtain. Clinical biomarkers, often measured alongside medical images and used in diagnostic workup, may provide a rich set of data that can be collected retrospectively and utilized to train diagnostic models. In this work, we focused on assessing the potential of blood serum biomarkers, B-type natriuretic peptide (BNP) and NT-pro B-type natriuretic peptide (BNPP), indicative of acute heart failure (HF) and cardiogenic pulmonary edema to be used as continuously valued labels for training a

radiographic deep learning algorithm. For this purpose, a CNN was trained using 27748 radiographs to automatically infer BNP and BNPP, and achieved strong performance (AUC=0.903, sensitivity=0.926, specificity=0.857, r=0.787). Also, the trained models achieved strong performance (AUC=0.801) for pulmonary edema detection when evaluated with radiologist labels. Since relevant radiographic features visible to the CNN may vary greatly based on image resolution, we also assessed the impact of image resolution on model learning and performance, comparing CNNs trained at five image sizes (64x64 to 1024x1024). Increasing image resolutions had diminishing but positive gains in AUC. Perhaps more importantly, experiments using three activation mapping techniques (saliency, Grad-CAM, XRAI) revealed considerably increased attention in the lungs with larger image sizes. This result emphasizes the need to utilize radiographs near native resolution for optimal CNN performance, which may not be fully captured by summary metrics like AUC.

INTRODUCTION

Recently, convolutional neural networks (CNN) have been proposed to perform automated classification, localization or segmentation of various diseases and anatomical landmarks directly from medical images to facilitate medical image acquisition and interpretation. Most of these approaches rely heavily on supervised training using manual annotations from physicians, which must be created prospectively and may be labor and time intensive to obtain, or in certain disease states, may not be perfectly reliable. To overcome this obstacle, approaches such as transfer learning, semi-supervised learning, self-supervised learning, and natural language processing (NLP) based labels have been suggested. In this work, we propose and explore an additional, alternative data-centric approach: using clinical serum biomarkers as continuously valued training labels.

Clinical biomarkers, specifically blood serum tests, have served a vital role in clinical diagnosis for over 70 years (Krebs 1950). Blood serum biomarkers are often measured alongside medical images and used in diagnostic workup and may provide a rich set of data that can be collected retrospectively and utilized to train diagnostic models. Specifically, we focus on the usage of B-type natriuretic peptide (BNP) and NT-pro B-type natriuretic peptide (NT-proBNP or BNPP), which are indicative of acute heart failure (HF) and cardiogenic pulmonary edema. In some instances, they may be concurrently obtained at the time of acquisition of a chest radiograph.

Pulmonary edema is a condition characterized by excess fluid in the lungs, often caused by congestive heart failure (HF) among other etiologies (Staub 1974; Murray 2011). Diagnostic workup of pulmonary edema may involve a variety of techniques including patient history, clinical exam, chest radiograph, and blood serum tests (Ware and Matthay 2005). Due to its wide

availability and ability to provide alternative diagnoses that may share similar clinical features, chest radiographs are commonly used to diagnose and monitor the progression of pulmonary edema (Hammon et al. 2014; Halperin et al. 1985)

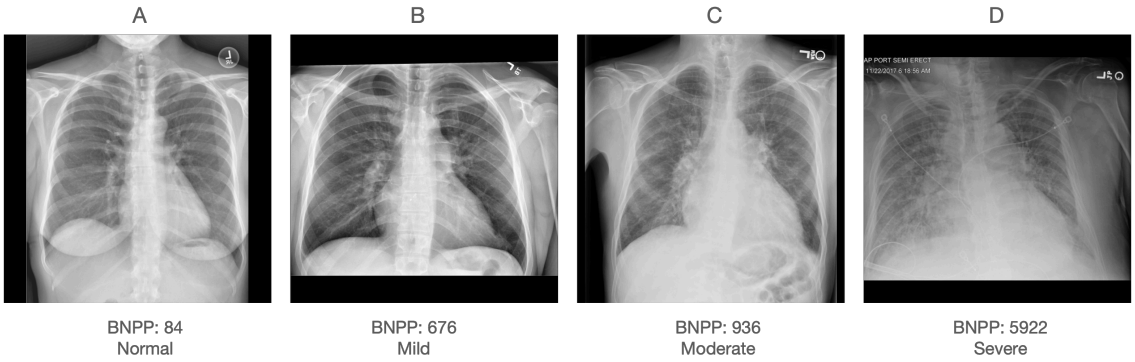

Fig 1. ***Relationship of radiographic appearance to BNPP and radiologist grade of severity of pulmonary edema.*** In (a), patient with normal laboratory measurement shows no signs of pulmonary edema. In (b) a patient with mildly elevated BNPP shows Kerley B lines, peribronchial cuffing and indistinctness of pulmonary vasculature. In (c) patient with moderately elevated BNPP shows the addition of perihilar opacities. In (d) patient with highly elevated BNPP shows frank alveolar opacities.

Radiographic assessment of pulmonary edema is complicated due to differences in features exhibited at different severity levels. While severe cases exhibit alveolar flooding and are relatively easy to identify in the chest radiograph (Fig. 1D), mild cases are characterized by interstitial fluid buildup and rely on much subtler findings such as Kerley B lines and peribronchial cuffing (Fig. 1B) (Milne et al. 1985; Gluecker et al. 1999; Aberle et al. 1988). Individual patients with mild edema can quickly progress into moderate and severe edema or vice versa (Barile 2020). Conditions such as heart failure can cause progressive worsening (Assaad et al. 2018; Gropper, Wiener-Kronish, and Hashimoto 1994). Therapies such as fluid diuresis or dialysis can reverse pulmonary edema (Krämer, Schweda, and Riegger 1999). Thus, accurate assessment of pulmonary edema is crucial for guiding and monitoring response to treatment.

Recently, several groups (Lakhani and Sundaram 2017; Wang, Lin, and Wong 2020; Hwang et al. 2019) have reported the application of deep convolutional neural networks (CNNs)

to classify chest radiographs for various pathologies, including pneumonia, pulmonary edema, pneumothorax, and many others. While these early works show the promise of CNNs for radiographic interpretation, most lack the specificity and granularity in diagnosis at a level that is typically required for diagnostic utility. For example, prior work does not draw the distinction between mild, moderate, and severe pulmonary edema, aspects which help determine the necessity for intervention or change in therapy.

There are several obstacles that impede the development of CNNs capable of quantifying the severity of pulmonary edema from chest radiographs. One obstacle is the lack of reliable labeled training data. Collecting a sufficiently large set of images that are annotated by radiologists is labor and time intensive, and specifically for pulmonary edema, may have limited agreement even amongst expert readers(Duggan et al. 2021). To address this problem, we explored an additional and objective source of ground truth for training CNNs for disease detection: B-type natriuretic peptide (BNP) and NT-pro B-type natriuretic peptide (BNPP). BNP and BNPP are continuously valued cardiac biomarkers measured from blood serum and are part of the diagnostic workup of suspected cardiogenic pulmonary edema (Ware and Matthay 2005).

Elevated values of BNP and BNPP are indicative of atrial stretch, observed in acute heart failure and pulmonary edema (Ray et al. 2005; Huang et al. 2016). A CNN that could infer BNP and BNPP directly from a chest radiograph could perceive variations that correlate with pulmonary edema and heart failure. Such a model could be used as an assistive diagnostic tool to help clinicians further analyze a chest radiograph.

We further observed that in the published literature, many CNN algorithms have been trained and evaluated on low-resolution down sampled images commonly provided in public databases(Pan, Cadrin-Chênevert, and Cheng 2019; Jaeger et al. 2014; Seah et al. 2019). Many

of the characteristics of pulmonary edema lie near the threshold of resolution of chest radiographs, including interstitial Kerley B lines and peribronchial cuffing. Thus, we investigated the ability of CNNs to infer BNP and BNPP when trained at a variety of input image sizes (64x64 – 1024x1024).

Finally, to assess the visual fields used by CNNs to achieve their performance, we described two methods for assessing spatial attention, leveraging a separate anatomic segmentation CNN, which we called “area attention” and “blur sensitivity”. These methods were used to determine the ratio of lung attention and model sensitivity to image ablation in the lungs used by each CNN model. We proposed these methods to measure the use of essential regions of the radiograph in assessment of pulmonary edema.

METHODS

A. DATASET

With institutional review board (IRB) approval and waiver of informed consent, we constructed a dataset of 27748 frontal chest radiographs with BNP or BNPP laboratory values from 16401 patients curated from the electric medical record and picture archiving and communication system (PACS) of our institution. We included all radiographs and laboratory measurements from Nov 4th, 2017 to Dec 1st, 2020 for patients who underwent either measurement of BNP or BNPP within 24 hours of a radiograph.

TABLE I
CHEST RADIOGRAPH AND SERUM LABORATORY DATA USED FOR TRAINING AND EVALUATION.

Dataset	**Train (80%)**	**Validation (10%)**	**Test (10%)**
NT-proBNP			
Patients (n = 15409)	12327	1541	1541
Radiographs (n = 26667)	21374	2602	2691
NT-proBNP > 400 (n = 22021)	17631 (82.5%)	2168 (83.3%)	2222 (82.5%)
Measured BNPP (Mean)	4997	4825	4227
Measured BNPP (SD)	11443	11369	9914
BNP			
Patients (n = 1325)	1044	141	140
Radiograph (n = 1423)	1124	148	151
BNP > 100 (n = 640)	512 (45.6%)	61 (41.2%)	67 (44.4%)
Measured BNP (Mean)	542	695	672
Measured BNP (SD)	944	1619	1205

Radiographs varied in image dimension depending on source x-ray device, ranging from 1400 – 4700 in height and width. A total of 26667 of these radiographs from 15409 patients had a corresponding BNPP measurement and the 1423 remaining radiographs from 1325 patients had a corresponding BNP measurement (Table 1). There was little overlap in these populations – only 342

radiographs had both BNP and BNPP measurements. Mean absolute difference between time of radiograph and BNPP laboratory sample collection was 2 hours, 22 minutes. Mean absolute difference between time of radiograph and BNP laboratory sample collection was 8 hours, 44 minutes. Mean and standard deviation of measured BNP values for the population was 556 ± 993 pg/mL. Mean and standard deviation of measured BNPP values was 4902 ± 11294 pg/mL.

Data was then divided by patients, not by radiographic images, into training (80%), validation (10%) and test (10%) cohorts. There was no significant difference in BNPP or BNP value distributions between training, validation, and testing sets ($p < 0.1$, *Kolmogorov Smirnov Test*).

B. CNN TRAINING

1. TWO STAGE TRAINING

Because the BNP dataset was significantly smaller than the BNPP dataset (n=1423 vs 26667 respectively), a two-stage pipeline was used to train a bifurcated CNN to jointly predict BNP and BNPP, shown in Fig. 2. All CNNs were trained using Adam optimizer with a fixed learning rate of 1e-5 for 50 epochs, and batch size 16. In the first stage of training, a ResNet152v2 model(He et al. 2016), pretrained on the ImageNet dataset(Deng et al. 2009), was trained to infer BNPP from a chest radiograph. A custom loss function based on mean absolute error (MAE) was used. Given a dataset of n input radiographs, we defined the loss over the dataset as:

$$MAE^{BNPP} = \frac{1}{n}\sum_{k\in\{1,\ldots,n\}} AE_k^{BNPP}$$

where AE_k^{BNPP} is defined as:

$$AE_k^{BNPP} = |\ln(1 + y_k^{BNPP}) - \ln(1 + \hat{y}_k^{BNPP})|$$

where y_k^{BNPP} is the lab measured BNPP value and $\hat{y}_k^{BNPP}$ is the inferred BNPP value for the k[th] input radiograph in the dataset. The BNPP values range from (0-70,000 pg/mL) and are

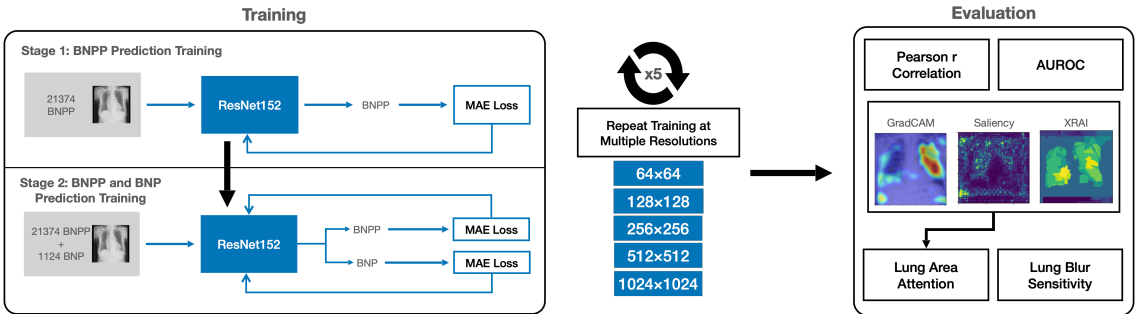

Fig 2. ***Flow chart of CNN training and evaluation***. ResNet152v2 CNNs were trained in two stages, first to predict BNPP from the radiograph, then to predict both BNPP and BNP from the radiograph. Multiple CNNs were trained, each at different input resolutions. Each of these were evaluated for performance with Pearson's R and AUROC. Finally, three attention mapping techniques (Saliency, Grad-CAM, and XRAI) and two metrics were used to quantify level of CNN attention in the lungs used to perform the inference.

exponentially distributed, with a small number of values significantly higher than the mean. To account for this and prevent overfitting to outliers using MAE loss, we used log transformation of the measured and inferred BNPP values when calculating AE_k^{BNPP}.

In the second stage of training, an additional fully connected layer was added to the end of the ResNet152v2 to additionally predict BNP from a chest radiograph. Weights acquired from the first stage of training were frozen, except for the fully connected layers at the end of the model: one for BNPP and one for BNP. Both BNP and BNPP datasets were used to train the stage 2 model to jointly predict BNP and BNPP from a chest radiograph. A scheduler was used to balance the number of BNP labeled radiographs and BNPP labeled radiographs in each minibatch of training examples. This ensures that for each epoch, the entire BNP training set was used (n=1124), while an equal number of BNPP labeled images were randomly sampled without replacement from the BNPP dataset. We further modified our custom MAE loss function from stage 1 (eq 1) to train both tasks simultaneously. When computing the loss, we ignored missing BNP or BNPP measurements:

$$MAE = \frac{1}{n} \sum_{k \in \{1,\ldots,n\}} \alpha_k AE_k^{BNPP} + \beta_k AE_k^{BNP}$$

where

$$\alpha_k = \begin{cases} 1, & y^{BNPP}\ available \\ 0, & otherwise \end{cases}$$

$$\beta_k = \begin{cases} 1, & y^{BNP}\ available \\ 0, & otherwise \end{cases}$$

2. TRAINING AT MULTIPLE RESOLUTIONS

To explore the effect of image resolution on model performance, we trained multiple CNNs for each of five input resolutions with image sizes of 64x64, 128x128, 256x256, 512x512, and 1024x1024. No additional architectural modifications were made to the models to adjust for input resolution. Images were cropped at their larger dimension to equal height and width and downscaled to the desired resolution with bilinear interpolation from python OpenCV 4.5.1.48 library.

A single Nvidia V100 GPU was used to train lower resolution models (64x64 – 512x512) and 8 NVIDIA V100 GPUs from an NVIDIA DGX cluster running in an NGC container on the Singularity runtime environment were used to train the 1024x1024 CNN. Synchronous distributed training was performed using TensorFlow 2.1.0 with *MirroredStrategy*.

C. CNN EVALUATION

Each CNN was evaluated on the same test set, comparing Pearson *r* and area under the receiver operating characteristic curve (AUROC or AUC ROC). ROC curves were computed after binary thresholding of BNP and BNPP measurements, according to previously established screening thresholds for acute heart failure detection (greater than 400 for BNPP, greater than

100 for BNP) (Kim and Januzzi 2011). The Youden's j index was computed for each CNN and used to calculate sensitivity, specificity, and confusion matrices for BNP and BNPP inference.

To further evaluate CNN performance, we recruited five subspecialty radiologists who independently graded the presence or absence of at least mild pulmonary edema for 250 radiographs from 250 patients, randomly drawn from the test set. Each radiologist was randomly assigned 50 radiographs without overlap to serve as ground truth.

D. CNN ACTIVATION MAPPING

To assess the effect of resolution on CNN activation, we applied three activation mapping techniques (Saliency(Simonyan, Vedaldi, and Zisserman 2014), grad-CAM(Selvaraju et al. 2017), and XRAI(Kapishnikov et al. 2019)) to each trained CNN. Activation maps were generated for each radiograph in the BNPP test set (n=2691).

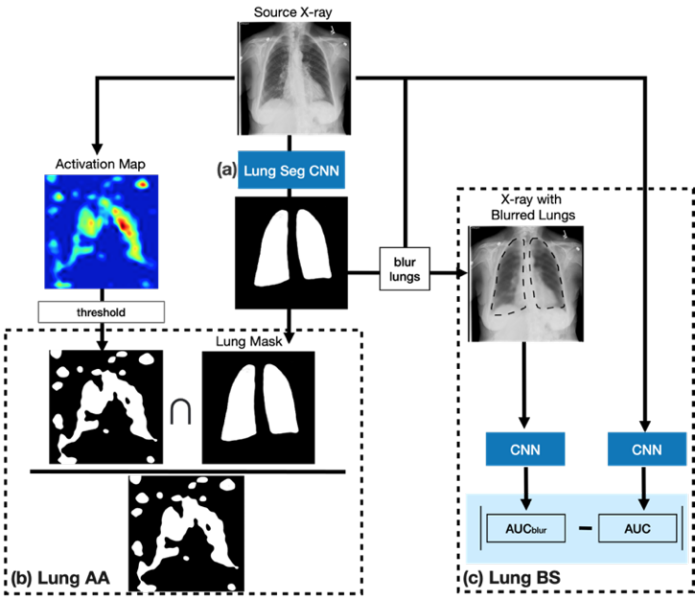

Fig 3. ***Proposed metrics to quantify level of CNN attention.*** In (a), a lung segmentation CNN is used to produce masks for input images. In (b), lung area attention (AA) calculates the proportion of pixels from an activation map that are also located within the lungs. In (c), lung blur sensitivity (BS) calculates the decrease in AUC caused when pixels within the lung mask are blurred by convolution with a Gaussian filter.

E. QUANTITATIVE ANALYSIS OF CNN ATTENTION

Activation maps were generated for each radiograph in the BNPP test set (n=2691). Currently, few metrics exist to quantitatively analyze deep learning saliency maps. To measure the degree of CNN attention within the lungs, we propose two metrics: lung area attention (AA) and lung blur sensitivity (BS), both of which utilize lung masks from a separately developed lung segmentation CNN. This lung segmentation module is part of a multi-organ segmentation U-net(Ronneberger, Fischer, and Brox 2015) developed using 302 radiographs and their corresponding lung masks, manually annotated utilizing an in- house radiograph annotation software developed in Python (Fig. 3a).

1. AREA ATTENTION

We define lung area attention (AA) as the proportion of the CNN activation map that overlaps with the lung segmentation mask (Fig. 3b). Lung AA is defined as the proportion of highly activated pixels in the activation map (and input image) that intersect with the lung mask:

$$AA(x) = \frac{heatmap(x) \cap mask(x)}{heatmap(x)}$$

where x is the input chest radiograph, heatmap(x) is the activation map from inference on x, and mask(x) is the lung mask. Activation maps were normalized to have values between 0-1, and thresholded based on the mean pixel value across all activation maps from a single model and technique.

Intuitively, a CNN with a high average lung AA value across the test set has focused mostly within the lungs rather than the rest of the image. In contrast, a CNN with an activation map concentrated on large regions of the image both inside and outside of the lungs will achieve a smaller lung AA.

2. BLUR SENSITIVITY

We define blur sensitivity (BS) as another way to estimate attention (Fig. 3c). Lung BS measures the sensitivity of the CNN to blurring of the image within the region denoted by a lung mask, generated by the separately developed lung mask CNN:

$$BS(\hat{y}, b) = AUC(\hat{y}, y) - AUC(blur(\hat{y}, b), y)$$

where $\hat{y}$ is a vector of the inferred values from a trained CNN for the entire test set, blur($\hat{y}$, b) are the inferred values from the CNN when every image in the test set has lungs blurred with a gaussian kernel of size b, and AUC($\hat{y}$,y) is the AUC computed from a vector of inferred values $\hat{y}$ and ground truth vector y. In this work, we applied gaussian kernel sizes of (3,3), (5,5), (11,11), (23,23), and (47,47) for models trained at 64x64, 128x128, 256x256, 512x512, and 1024x1024 input sizes, respectively. Gaussian kernel sizes were selected to increase proportionally with image size to ensure a similar effect relative to the field of view.

A CNN that relies on high resolution details within the lungs will have a large lung BS value. A CNN that overlooks such details within the lung area will have a smaller lung BS value.

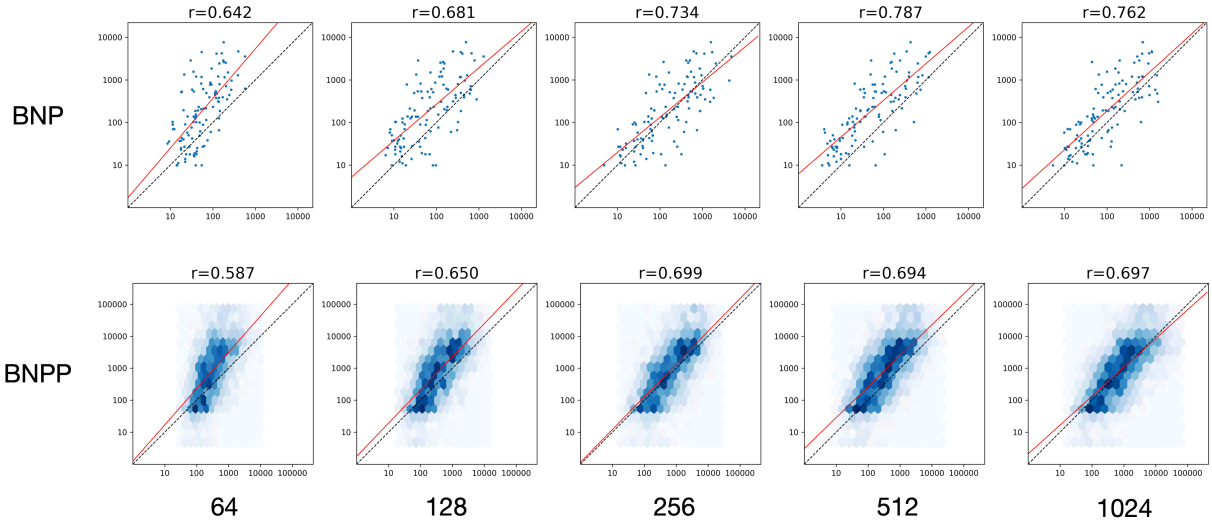

Fig 4. ***Plots showing the relationship of measured and inferred BNP and BNPP at multiple input resolutions.*** Plots are shown on a log scale. On the top row, correlation of measured and inferred B-type natriuretic peptide (BNP) increases with resolution (r=0.642-0.762, *n*=141). On the bottom row, correlation of measured and inferred NT-pro B-type natriuretic peptide (BNPP) increases with resolution (r=0.587-0.697, *n*=2691). Red lines indicate the line of best fit, while black lines indicate the line of best fit if all predictions were correct. To allow better visualization given the large BNPP test set size, a hex-bin plot was used for the bottom row.

A. CNN EVALUATION

The relationship between measured laboratory values and the respective inferred values by CNNs are shown in Figure 4, for both BNP and BNPP test sets at different input image resolutions. There was relatively stronger correlation between measured and inferred BNP values than measured and inferred BNPP values at all image resolutions, though the number of measurements used in CNN training and evaluation were much fewer (r=0.642-0.762 for BNP and r=0.587-0.697 for BNPP). Pearson correlation coefficient between measured and inferred laboratory values increased with input image resolution, having the greatest effect at lower image resolutions. There was little change in Pearson correlation at higher image sizes of 512 and 1024 with a tendency to a slight decline. For BNP, peak Pearson r was 0.787 for 512 image size and decreased slightly to 0.762 for 1024. For BNPP inference, peak Pearson r was 0.699 for 256, and plateaued at higher image sizes.

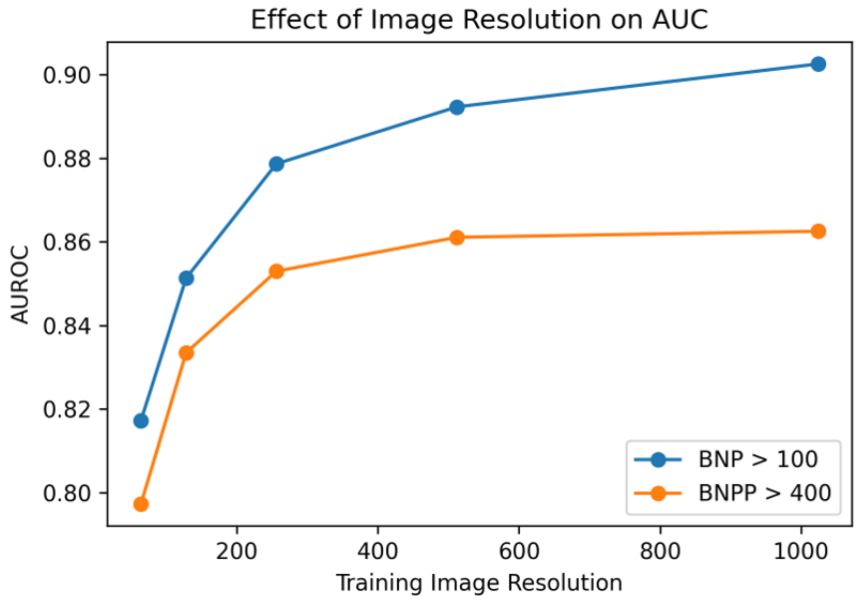

Fig 5. ***Relationship between CNN input resolution and AUC for identification of patients*** **exceeding thresholds for acute heart failure.** Greater CNN performance for detecting patients with BNP>100 and BNPP>400 is observed at the higher 512 and 1024 image sizes, despite observed declines in Pearson correlation.

The relationship between input image resolution and AUC obtained for prediction of BNP thresholded at 100 and BNPP thresholded at 400 are shown in Figure 5. Though Pearson *r* peaked earlier at smaller 256 and 512 image sizes, we observed continuous increments in AUC up to the 1024 image size. Increasing the image size from 64 to 1024 resulted in increased AUC (0.817 to 0.903 for BNP and 0.797 to 0.863 for BNPP). The greatest improvement in AUC was observed between the lowest resolutions.

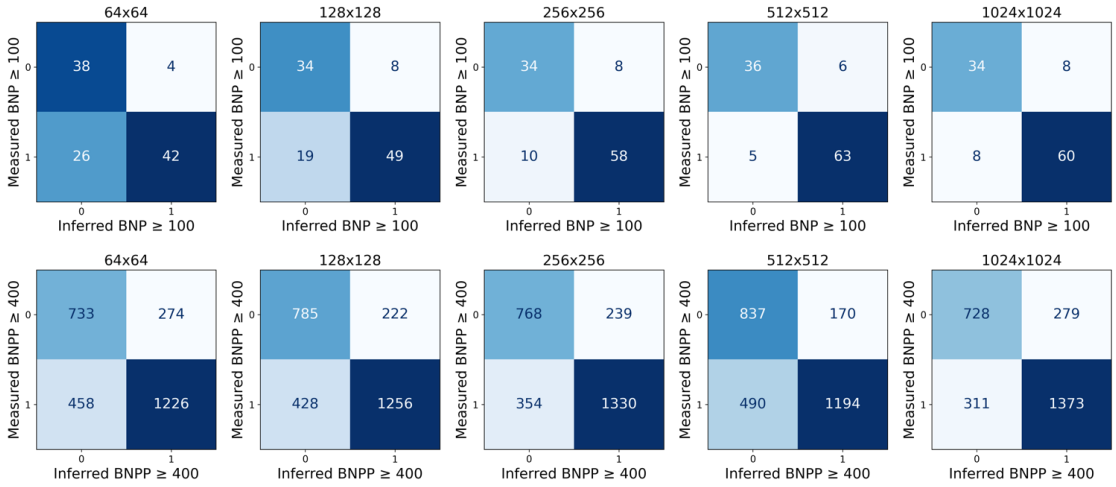

Fig 6. ***Confusion matrices for BNP and BNPP inference at thresholds for acute heart failure.***

TABLE II
EFFECT OF RESOLUTION ON OPTIMUM THRESHOLDS,
SENSITIVITY AND SPECIFICITY AT YOUDEN'S J INDEX.

Image Size	**64**	**128**	**256**	**512**	**1024**
BNP Threshold (pg/mL)	58	56	76	22	48
Sensitivity	0.618	0.721	0.852	0.926	0.882
Specificity	0.904	0.810	0.810	0.857	0.810
BNPP Threshold (pg/mL)	250	300	440	320	460
Sensitivity	0.728	0.746	0.790	0.709	0.815
Specificity	0.728	0.780	0.763	0.831	0.723

Confusion matrices, sensitivity and specificity values calculated from the Youden's j index of the AUROC at each input image size are shown in figure 6 and table II. For BNP inference, increased image size led to lower specificity and false negatives and higher sensitivity. For BNPP inference, increased image size did not lead to monotonic changes in sensitivity or specificity, although the sensitivity peaked at 1024 image size (0.815) and the specificity peaked at 512 image size (0.831).

TABLE III
PULMONARY EDEMA DETECTION PERFORMANCE COMPARISON AGAINST RADIOLOGISTS COMBINED AND SEPARATED BY INDIVIDUAL READER.

Image Size	**64**	**128**	**256**	**512**	**1024**
Inferred BNPP at Youden's	680	880	970	900	760
All Readers	0.774	0.783	0.780	0.801	0.795
Reader 1	0.834	0.835	0.852	0.832	0.847
Reader 2	0.760	0.784	0.754	0.790	0.771
Reader 3	0.598	0.602	0.598	0.649	0.625
Reader 4	0.838	0.853	0.836	0.863	0.867
Reader 5	0.839	0.841	0.862	0.869	0.865

To further investigate the clinical potential of our proposed method, we evaluated performance of the algorithm relative to each of five subspecialty cardiothoracic radiologists, who evaluated 250 randomly drawn images from the test set. CNN models exhibited modest improvements with increased input image size (table III). CNN performance varied relative to the ground truth provided by each radiologist, notably reader 2 and reader 3, but generally increased with input image size across each of the five readers (table III).

B. CNN ACTIVATION MAPPING

Grad-CAM activation maps were generated to visually assess the effect of resolution on CNN activation for each individual radiograph. Exemplar cases of patients with mild and severe pulmonary edema are shown in figures 7 and 8.

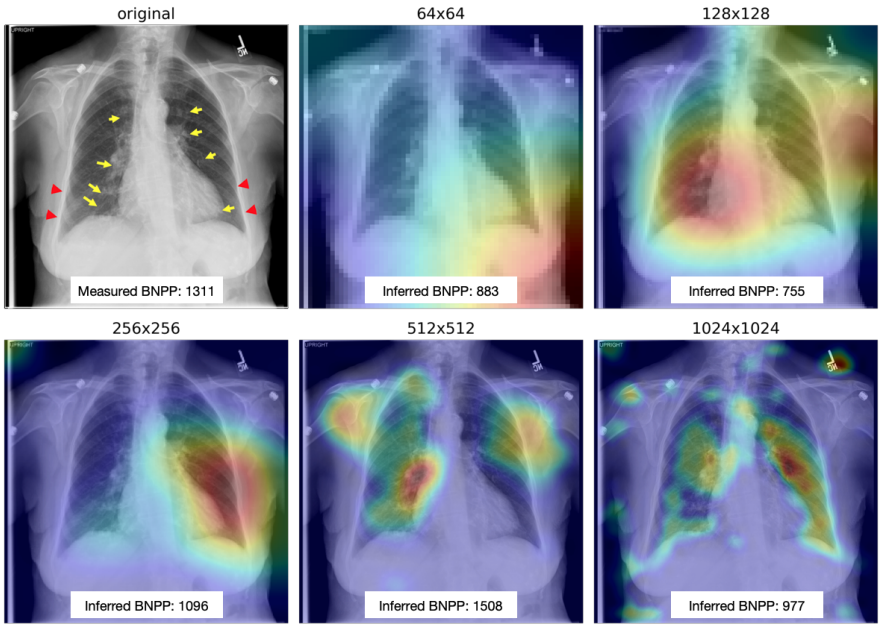

Fig 7. ***Comparison of grad-CAM heatmaps from models trained at different input resolutions on an exemplar case of mild pulmonary edema.*** The original image was independently annotated by cardiothoracic radiologist for peribronchial cuffing (yellow arrows) and Kerley B lines (red arrows), findings of mild pulmonary edema. Low resolution models (64, 128) show attention in large, indistinct regions on the chest X-ray. Higher resolution models (512, 1024) show greater attention to the lungs.

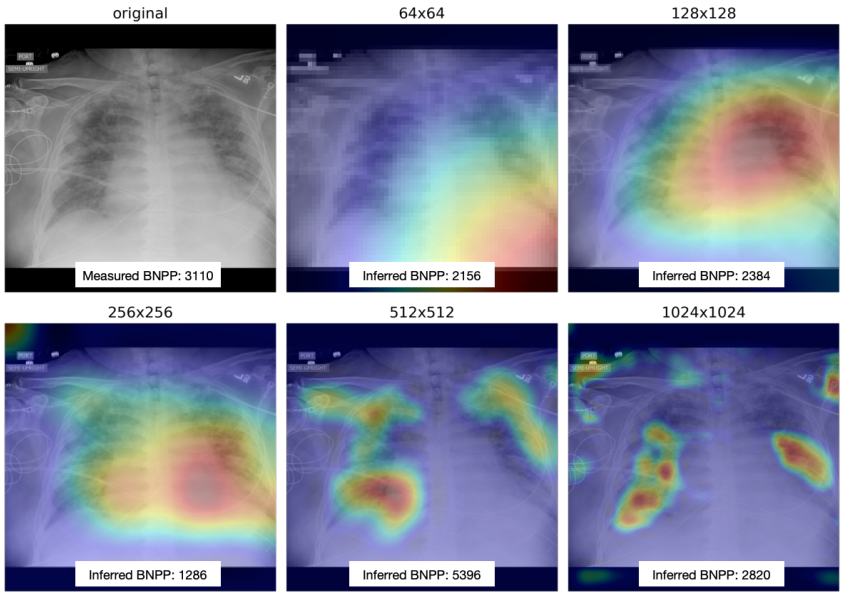

Fig 8. ***Comparison of grad-CAM heatmaps from CNNs trained at different training image resolutions on an exemplar case of severe pulmonary edema.*** Low resolution models (64, 128) show attention in large, indistinct regions of the chest X-ray. Higher resolution models (512, 1024) show greater attention to the lungs, where airspace opacities are observed.

In both examples, Grad-CAM shows diffuse and inconsistent activation at lower 64 and 128 image sizes, which increasingly focus on the lungs at higher 512 and 1024 image sizes. Multiple activation map strategies were then used to assess for consistent trends in CNN activation. This included saliency, grad-CAM, and XRAI, which were performed on CNNs trained at all image sizes. The resulting activation maps at each resolution were averaged over all test cases for each strategy, presented as average activation maps shown in figure 9.

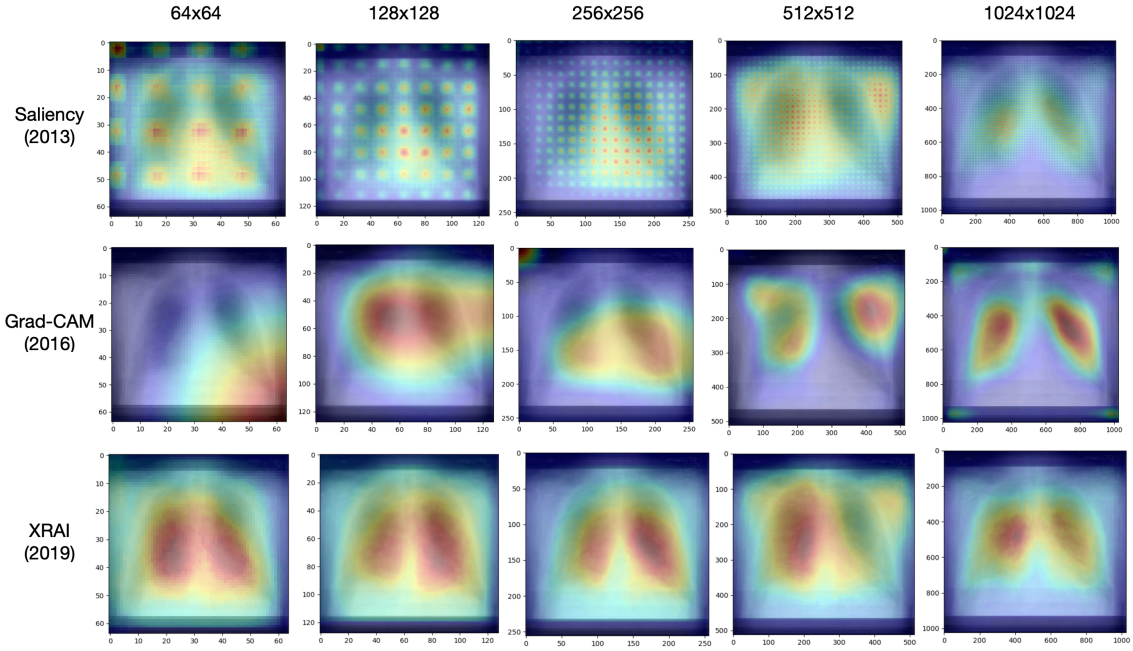

Fig 9. ***Average heatmaps from 3 visualization techniques applied to models trained on different input resolutions.*** Three model visualization methods were used: Saliency (top), grad-CAM (middle), XRAI (bottom). Resulting heatmaps were averaged over all images in the test set (n=2691), and overlayed on the average of chest radiographs from the test set. Areas of high activation are in red, and low activation are in blue. Visually, model activation is spread throughout the image when low resolution training images are used. As training image resolution increases, model activation is increasingly concentrated in the lung area.

Overall, average heatmaps from all three techniques show increasing attention on the lungs and decrease in attention outside of the lungs with greater image resolution. For the 64x64 model, saliency and XRAI show activations throughout the entire image, while grad-CAM activations are focused on the lower right corner of the image. For the 128x128 and 256x256

models, all three techniques show increased activations concentrated in the general lung area. Saliency and grad-CAM activation maps still focus on a single large region with no distinction between left and right lungs. For the 512x512 and 1024x1024 models, all three techniques show activation in two distinct regions: the left lung and right lung, with minimal activations outside of the lungs.

C. QUANTITATIVE ANALYSIS OF MODEL ATTENTION

To better quantify the level of CNN attention to the lungs, we proposed two metrics, lung area attention (AA) and lung blur sensitivity (BS). We used both lung AA and lung BS to quantify the effect of image resolution on model attention to the lung region.

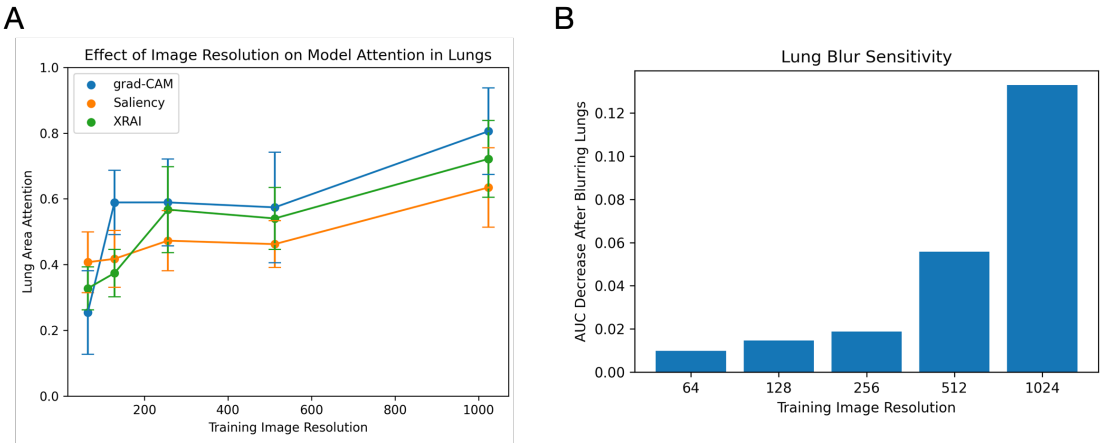

Fig 10. ***Plots showing the relationship between image resolution and CNN attention in the lungs.*** A) shows results based on average lung area attention (AA) over the test set. Error bars indicate standard deviation. B) shows results based on lung blur sensitivity (BS).

1. AREA ATTENTION

We calculated the average lung area attention over all images in the test set (n=2691). Fig. 10A plots the lung AA values for five CNNs, each trained at different input resolutions, using three activation mapping techniques (saliency, grad-CAM, XRAI). Overall, increasing input resolution from 64x64 to 1024x1024 led to increasing average lung AA (0.40 to 0.64 for

saliency, 0.26 to 0.80 for grad-CAM, 0.33 to 0.72 for XRAI). The greatest changes in average lung AA were observed when input resolution was increased from 512x512 to 1024x1024 (0.46 to 0.64 for Saliency, 0.58 to 0.80 for grad-CAM, and 0.54 to 0.72 for XRAI). At 64x64 image resolution, the average lung AA for all three techniques was less than 0.5, indicating that the 64x64 trained CNN. This trend contrasts with the results observed at 1024x1024 image resolution, where all three techniques yielded lung AA greater than 0.5, indicating that the 1024x1024 trained CNN mostly focused within the lungs.

TABLE IV
EFFECT OF MEASURED BNPP ON LUNG AREA ATTENTION CALCULATED FROM GRAD-CAM SALIENCY MAPS. CASES WITH MEASURED BNPP ABOVE AND BELOW 400 WERE COMPARED BASED ON LUNG AA.

Image Size	**64**	**128**	**256**	**512**	**1024**
BNPP ≥ 400					
Mean	0.247	0.589	0.591	0.566	0.810
SD	0.103	0.121	0.096	0.084	0.142
BNPP < 400					
Mean	0.262	0.634	0.593	0.585	0.811
SD	0.089	0.081	0.126	0.149	0.097

The relationship between measured BNPP level and lung AA is shown in table IV. There was no significant difference in measured attention in the lungs between samples with high (≥400) and low (<400) BNPP, which suggests that average lung AA is independent of BNP or BNPP values. Lung attention seemed to be consistent regardless of BNPP but varied greatly with input resolution.

2. BLUR SENSITIVITY

We calculated the lung BS based on AUC obtained from inference on all images in the test set (n=2691). Fig. 10B plots the average lung BS for five CNNs, each trained at different image resolutions. Overall, increasing input resolution from 64x64 to 1024x1024 resulted in increasing lung BS (0.01 to 0.13). For the models trained at lower image resolutions (64x64 –

256x256), lung BS was less than 0.02, indicating that blurring the lungs resulted in less than 2% decrease in AUC. The higher resolution models trained at 512x512 and 1024x1024 exhibit significantly higher lung BS values of 0.06 and 0.13 respectively, indicating that blurring the lungs resulted in AUC decreases of 6% and 13%.

DISCUSSION

In this work, we demonstrated the feasibility of inferring BNP and BNPP directly from chest radiographs. Resnet152V2 CNNs achieved optimal performance at larger input image sizes, which highlighted the importance of higher resolution spatial details for assessing pulmonary edema. At 1024x1024 image size, Pearson R values for BNP and BNPP were 0.762 and 0.697. Thresholding at BNP>100 and BNPP>400, AUROCs were 0.903 and 0.863 respectively. The trained CNNs also achieved strong performance for pulmonary edema detection when evaluated against radiologist labels. By applying three activation mapping techniques (saliency, grad-CAM, XRAI) and two proposed quantitative metrics (lung AA, lung BS) to our CNNs, we confirmed that increasing input resolution increased model attention to the lungs.

Few prior investigators have begun to explore the application of CNNs to assessing pulmonary edema, whether in isolation or as part of multiple class image classification. Prior investigators achieved similar AUROC for edema detection, ranging from 0.814-0.924(Rajpurkar et al. 2018; Cicero et al. 2017; Sabottke and Spieler 2020) with a variety of CNN architectures. While direct comparisons of AUROC are difficult due to differences in patient test populations, these are comparable to our approach of applying serum laboratory markers as ground truth, rather than NLP-derived labels. We show that training CNNs with serum laboratory markers (BNP and BNPP) achieved similar results, while providing additional information that can help to grade severity. A previous work(Seah et al. 2019) showed initial feasibility of BNP for this task, albeit at very low resolution, and observed CNN attention predominantly outside of the lungs. We suspected this might have been the result of lower image resolutions used by the prior authors. In another work(Sabottke and Spieler 2020), investigators

showed little improvement in AUROC above 128x128 resolution when using NLP-derived image labels, and left uncertainty about where CNN attention was for this inference.

We thus expanded on these pioneering works and show that while some performance is maintained at lower image sizes, CNNs require higher resolution images to ensure their inferences are the result of lung attention. Many known findings of pulmonary edema used by cardiothoracic radiologists, including Kerley B lines and peribronchial cuffing, are not visible at low resolution. We further devised two new metrics *area attention* and *blur sensitivity* to quantitatively measure lung attention. Our results provide insight into the effect of image resolution on CNN learning.

Currently, most deep learning models are benchmarked and optimized by several different metrics, including accuracy or AUROC. However, our findings suggest that global performance metrics such as accuracy or AUC do not always provide a full description of model performance or value. It is clear that assessment of pulmonary edema requires attention to the lungs. Thus, attention may be just as important for evaluation of CNNs, to confirm that models are not relying heavily on spurious relationships but instead focusing on areas of the image relevant for diagnosis, which may mitigate the problem of shortcut learning(DeGrave, Janizek, and Lee 2021).

As the image size increases, the filter size of the CNNs stay at a fixed size. This means that the CNNs are forced to look at smaller details when image size increases. At different image sizes, the CNNs are using different features to make their inference. It is entirely possible that the image features most indicative of BNP or BNPP at one image size are different from those at another image size, which might explain the change in performance with image size. For instance, the model could be looking at body habitus at smaller image size, heart size at medium

image sizes, and lung opacities at large image sizes. Each of these feature sets will have different diagnostic value, leading to different levels of performance.

Although BNP saw increasing performance with image size, BNPP seemed to peak after 512. This might be because of the reliability of the serum biomarkers themselves: BNPP values seemed to fluctuate more than BNP values. This might have made it more difficult to infer the values using higher levels of detail, resulting in the performance curve exhibited.

Using higher resolution images seemed to improve sensitivity and specificity for BNP and BNPP inference, as shown in figure 6. For BNP inference, increased image size led to lower specificity and false negatives and higher sensitivity. For BNPP inference, increased image size did not lead to monotonic changes in sensitivity or specificity, although the sensitivity peaked at 1024 image size (0.815) and the specificity peaked at 512 image size (0.831).

The ethical implications of this work are important to discuss(Char, Shah, and Magnus 2018; Harvey and Glocker 2019; Vollmer et al. 2020; Geis et al. 2019; Kohli and Geis 2018). First, if this would possibly threaten the role of radiologists in the clinic. Overall, there is a shortage of radiologists in our healthcare system, and the number of patients and medical imaging procedures is increasing, so they need all the help they can get. Also, this model is not descriptive or explainable enough to be standalone, if used it should provide assistance rather than make its own diagnoses. Second, there are issues of bias towards certain racial groups highlighted by recently published papers. In our case, we used data from UCSD Health Jacobs hospital for our training set, so it is unclear whether the results will generalize to other patient demographics. We believe that for this approach to be deployed in other hospitals, it should be retrained with data from that specific institution so that it is optimized for that patient population.

It must be noted that one of the main focuses of this work was assessing the effect of input resolution on CNN performance and attention. We wanted to ensure that the results we observed were generalizable and not the result of a specific architectural modification or technique. Therefore, we intentionally used a well-known and commonly used ResNet152v2 model architecture with minimal modifications. We chose this architecture over others due to its superior performance in preliminary experiments (appendix Fig. 11). Future work may focus on developing models structurally optimized for this task.

TABLE V
EFFECT OF RESOLUTION ON COMPUTATIONAL COST OF RESNET152V2 BASED MODEL (58M PARAMS), MEASURED IN FLOATING POINT OPERATIONS (FLOPs).

Image Size	**64**	**128**	**256**	**512**	**1024**
G-FLOPs	0.95	3.62	14.31	57.07	228.12

Another potential limitation of our work is that we did not experiment with resolutions higher than 1024x1024, even though the native resolution of our chest radiographs was as high as 4700x4700. For our experiments, we selected resolutions to encompass the entire gamut of commonly used input resolutions when training CNNs on chest radiographs. 1024x1024 was selected as the maximum resolution in our work for two reasons: (1) this is the maximum resolution of images from the commonly used public NIH ChestX-ray14 dataset(Jaeger et al. 2014) and RSNA-Pneumonia dataset(Pan, Cadrin-Chênevert, and Cheng 2019). (2) Compute resources required for training increase two-fold with resolution (Table V). Training a ResNet152v2 on 1024x1024 images pushed the memory limits of our available hardware. Future work may be directed at studying the performance gains at even higher resolutions.

Finally, we observe variable performance relative to each of five different subspecialty radiologists in the final study of algorithm performance. Similar performance was observed for three of the readers but had considerably lower agreement with two of the readers. We believe

this is due to variable thresholds that individual radiologists may set for binary classification of the presence or absence of pulmonary edema and may deserve further study. This also speaks to the benefit of using objective serum laboratory data to define ground truth for CNN training, which can prevent subjective thresholds of individual radiologists from limiting algorithm performance.

APPENDIX

A. Comparison to Other Architectures

We chose ResNet152v2 with MAE loss for its superior performance over other architectures for BNPP inference from radiographs. Figure 11 shows the ROC and AUC results using different methods using a threshold of 300 for measured BNPP and input image size of 512x512. We compared two loss functions (MAE, MSE), weight initialization schemes (random, ImageNet), and architectures (ResNet, DenseNet, Inception, InceptionResNet).

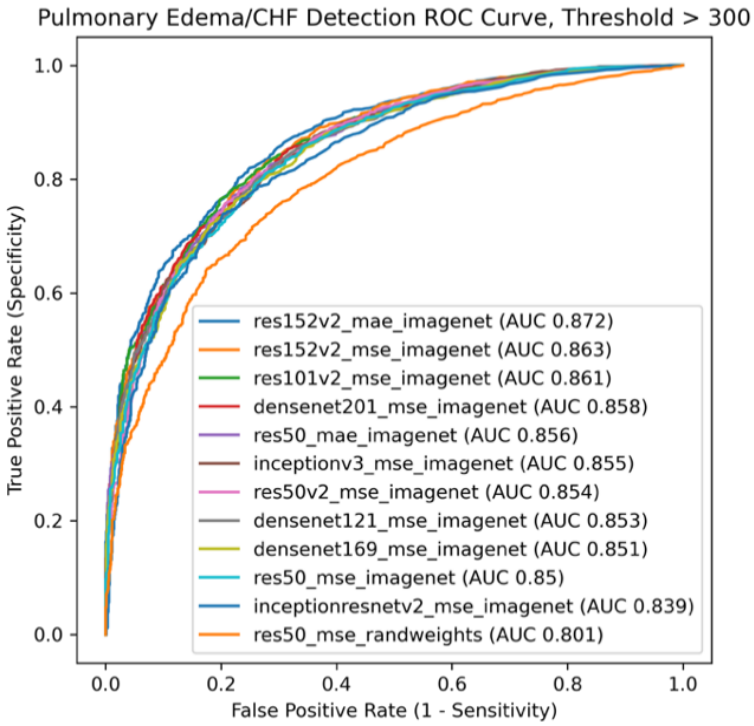

Fig 11. ***Comparison of AUROC on BNPP > 300 inference between different CNN architectures, weight initialization strategies, and loss functions when using 512x512 image size.*** ResNet152v2 with MAE loss was selected for subsequent experiments.

B. Comparing Various Thresholds of Measured BNPP

While in the manuscript we used a previously established screening threshold of 400 to detect acute heart failure from measured BNPP values, here we provided the ROC curves and their respective AUC computed for other potential thresholds in Figure 12.

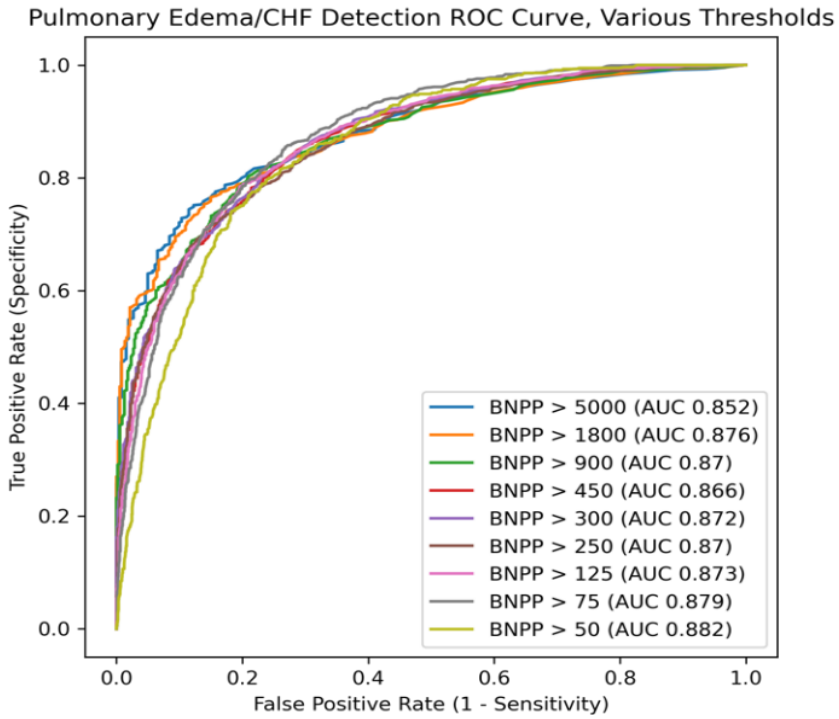

Fig 12. ***Comparison of CNN AUROC on various BNPP thresholds, when using ResNet152v2 model and 512x512 image size.***

ACKNOWLEDGEMENTS

This thesis, in full, is a reprint of the material as it appears in IEEE Access 2022. Huynh, Justin; Masoudi, Samira; Noorbaksh, Abraham; Mahmoodi, Amin; Kligerman, Seth; Yen, Andrew; Jacobs, Kathleen; Hahn, Lewis; Hasenstab, Kyle; Pazzani, Michael; Hsiao, Albert. *Deep Learning Radiographic Assessment of Pulmonary Edema: Optimizing Clinical Performance, Training with Serum Biomakers*. IEEE Access, 2022. The thesis author was the first author of this paper.

The author acknowledges research grant support from NSF RAPID 2026809 and DARPA N00173-21-Q-0141, in-kind support from Microsoft AI for Health, NVIDIA and Amazon Web Services for this work.

The author would like to especially thank research advisor and thesis chair Dr. Albert Hsiao for his guidance throughout the last two years. The author would like to thank Dr. Samira Masoudi for her close collaboration on every aspect of this work. The author would also like to thank all collaborators whom he had the pleasure of working with including Dr. Kyle Hasenstab, Dr. Abraham Noorbaksh, Dr. Michael Pazzani, Amin Mahmoodi, Dr. Lewis Hahn, Dr. Kathleen Jacobs, Dr. Andrew Yen, Dr. Seth Kligerman, Dr. Evan Masutani, Dr. Tara Retson, Dr. Sophie You, Dr. Shanmukha Srinivas, Rahul Chandruptala, Brendan Crabb, and others. The author would like to thank Professor Manmohan Chandraker for an amazing computer vision course and agreeing to be a thesis co-chair despite having different research areas. The author would like to thank Professor Laurel Riek for being on the thesis committee.

REFERENCES

Aberle, D R, J P Wiener-Kronish, W R Webb, and M A Matthay. 1988. “Hydrostatic versus Increased Permeability Pulmonary Edema: Diagnosis Based on Radiographic Criteria in Critically Ill Patients.” *Radiology* 168 (1). https://doi.org/10.1148/radiology.168.1.3380985.

Assaad, Sherif, Wolf B. Kratzert, Benjamin Shelley, Malcolm B. Friedman, and Albert Perrino. 2018. “Assessment of Pulmonary Edema: Principles and Practice.” *Journal of Cardiothoracic and Vascular Anesthesia* 32 (2). https://doi.org/10.1053/j.jvca.2017.08.028.

Barile, Maria. 2020. “Pulmonary Edema: A Pictorial Review of Imaging Manifestations and Current Understanding of Mechanisms of Disease.” *European Journal of Radiology Open* 7: 100274. https://doi.org/10.1016/j.ejro.2020.100274.

Char, Danton S., Nigam H. Shah, and David Magnus. 2018. “Implementing Machine Learning in Health Care — Addressing Ethical Challenges.” *New England Journal of Medicine* 378 (11): 981–83. https://doi.org/10.1056/NEJMp1714229.

Cicero, Mark, Alexander Bilbily, Errol Colak, Tim Dowdell, Bruce Gray, Kuhan Perampaladas, and Joseph Barfett. 2017. “Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs.” *Investigative Radiology* 52 (5). https://doi.org/10.1097/RLI.0000000000000341.

DeGrave, Alex J., Joseph D. Janizek, and Su In Lee. 2021. “AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal.” *Nature Machine Intelligence* 3 (7): 610–19. https://doi.org/10.1038/s42256-021-00338-7.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. IEEE. https://doi.org/10.1109/CVPR.2009.5206848.

Duggan, Gavin E, Joshua J Reicher, Yun Liu, Daniel Tse, and Shravya Shetty. 2021. “Improving Reference Standards for Validation of AI-Based Radiography.” *The British Journal of Radiology* 94 (1123): 20210435. https://doi.org/10.1259/bjr.20210435.

Geis, J. Raymond, Adrian P. Brady, Carol C. Wu, Jack Spencer, Erik Ranschaert, Jacob L. Jaremko, Steve G. Langer, et al. 2019. “Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement.” *Canadian Association of Radiologists Journal* 70 (4): 329–34. https://doi.org/10.1016/j.carj.2019.08.010.

Gluecker, Thomas, Patrizio Capasso, Pierre Schnyder, François Gudinchet, Marie-Denise Schaller, Jean-Pierre Revelly, René Chiolero, Peter Vock, and Stéphan Wicky. 1999. “Clinical and Radiologic Features of Pulmonary Edema.” *RadioGraphics* 19 (6). https://doi.org/10.1148/radiographics.19.6.g99no211507.

Gropper, Michael A., Jeanine P. Wiener-Kronish, and Satoru Hashimoto. 1994. “ACUTE CARDIOGENIC PULMONARY EDEMA.” *Clinics in Chest Medicine* 15 (3). https://doi.org/10.1016/S0272-5231(21)00946-1.

Halperin, B. D., T. W. Feeley, F. G. Mihm, C. Chiles, D. F. Guthaner, and N. E. Blank. 1985. “Evaluation of the Portable Chest Roentgenogram for Quantitating Extravascular Lung Water in Critically Ill Adults.” *Chest*. https://doi.org/10.1378/chest.88.5.649.

Hammon, Matthias, Peter Dankerl, Heinz Leonhard Voit-Höhne, Martin Sandmair, Ferdinand Josef Kammerer, Michael Uder, and Rolf Janka. 2014. “Improving Diagnostic Accuracy in

Assessing Pulmonary Edema on Bedside Chest Radiographs Using a Standardized Scoring Approach." *BMC Anesthesiology* 14 (1): 1–9. https://doi.org/10.1186/1471-2253-14-94.
Harvey, Hugh, and Ben Glocker. 2019. "A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology." In *Artificial Intelligence in Medical Imaging*, 61–72. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-94878-2_6.
He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Identity Mappings in Deep Residual Networks." In *European Conference on Computer Vision* , 630–45.
Huang, Ya Ting, Yuan Teng Tseng, Tung Wei Chu, John Chen, Min Yu Lai, Woung Ru Tang, and Chih Chung Shiao. 2016. "N-Terminal pro b-Type Natriuretic Peptide (NT-pro-BNP) - Based Score Can Predict in-Hospital Mortality in Patients with Heart Failure." *Scientific Reports* 6 (160). https://doi.org/10.1038/srep29590.
Hwang, Eui Jin, Ju Gang Nam, Woo Hyeon Lim, Sae Jin Park, Yun Soo Jeong, Ji Hee Kang, Eun Kyoung Hong, et al. 2019. "Deep Learning for Chest Radiograph Diagnosis in the Emergency Department." *Radiology* 293 (3): 573–80. https://doi.org/10.1148/radiol.2019191225.
Jaeger, Stefan, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. 2014. "Two Public Chest X-Ray Datasets for Computer-Aided Screening of Pulmonary Diseases." *Quantitative Imaging in Medicine and Surgery* 4 (6). https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.
Kapishnikov, Andrei, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. 2019. "XRAI: Better Attributions Through Regions." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4948–57. https://github.com/PAIR-code/saliency.
Kim, Han-Na, and James L. Januzzi. 2011. "Natriuretic Peptide Testing in Heart Failure." *Circulation* 123 (18). https://doi.org/10.1161/CIRCULATIONAHA.110.979500.
Kohli, Marc, and Raym Geis. 2018. "Ethics, Artificial Intelligence, and Radiology." *Journal of the American College of Radiology* 15 (9): 1317–19. https://doi.org/10.1016/j.jacr.2018.05.020.
Krämer, Bernhard K, Frank Schweda, and Günter A.J Riegger. 1999. "Diuretic Treatment and Diuretic Resistance in Heart Failure." *The American Journal of Medicine* 106 (1). https://doi.org/10.1016/S0002-9343(98)00365-9.
Krebs, H A. 1950. "Chemical Composition of Blood Plasma and Serum." *Annual Review of Biochemistry* 19 (1): 409–30. https://doi.org/10.1146/annurev.bi.19.070150.002205.
Lakhani, Paras, and Baskaran Sundaram. 2017. "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks." *Radiology* 284 (2). https://doi.org/10.1148/radiol.2017162326.
Milne, E. N.C., M. Pistolesi, M. Miniati, and C. Giuntini. 1985. "The Radiologic Distinction of Cardiogenic and Noncardiogenic Edema." *American Journal of Roentgenology* 144 (5): 879–94. https://doi.org/10.2214/ajr.144.5.879.
Murray, J. F. 2011. "Pulmonary Edema: Pathophysiology and Diagnosis." *International Journal of Tuberculosis and Lung Disease* 15 (2): 155–60.
Pan, Ian, Alexandre Cadrin-Chênevert, and Phillip M. Cheng. 2019. "Tackling the Radiological Society of North America Pneumonia Detection Challenge." *American Journal of Roentgenology* 213 (3). https://doi.org/10.2214/AJR.19.21512.
Rajpurkar, Pranav, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, et al. 2018. "Deep Learning for Chest Radiograph Diagnosis: A Retrospective

Comparison of the CheXNeXt Algorithm to Practicing Radiologists." *PLOS Medicine* 15 (11). https://doi.org/10.1371/journal.pmed.1002686.

Ray, Patrick, Martine Arthaud, Sophie Birolleau, Richard Isnard, Yannick Lefort, Jacques Boddaert, and Bruno Riou. 2005. "Comparison of Brain Natriuretic Peptide and Probrain Natriuretic Peptide in the Diagnosis of Cardiogenic Pulmonary Edema in Patients Aged 65 and Older." *Journal of the American Geriatrics Society* 53 (4). https://doi.org/10.1111/j.1532-5415.2005.53213.x.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 9351:234–41. Springer Verlag. https://doi.org/10.1007/978-3-319-24574-4_28.

Sabottke, Carl F., and Bradley M. Spieler. 2020. "The Effect of Image Resolution on Deep Learning in Radiography." *Radiology: Artificial Intelligence* 2 (1): e190015. https://doi.org/10.1148/ryai.2019190015.

Seah, Jarrel C.Y., Jennifer S.N. Tang, Andy Kitchen, Frank Gaillard, and Andrew F. Dixon. 2019. "Chest Radiographs in Congestive Heart Failure: Visualizing Neural Network Learning." *Radiology* 290 (3): 514–22. https://doi.org/10.1148/radiol.2018180887.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In *Proceedings of the IEEE International Conference on Computer Vision* , 618–26. http://gradcam.cloudcv.org.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. "Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, 1–8.

Staub, N C. 1974. "Pulmonary Edema." *Physiological Reviews* 54 (3). https://doi.org/10.1152/physrev.1974.54.3.678.

Vollmer, Sebastian, Bilal A Mateen, Gergo Bohner, Franz J Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, et al. 2020. "Machine Learning and Artificial Intelligence Research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics, and Effectiveness." *BMJ*, March, l6927. https://doi.org/10.1136/bmj.l6927.

Wang, Linda, Zhong Qiu Lin, and Alexander Wong. 2020. "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images." *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-76550-z.

Ware, Lorraine B., and Michael A. Matthay. 2005. "Acute Pulmonary Edema." *New England Journal of Medicine* 353 (26). https://doi.org/10.1056/NEJMcp052699.