Predicting Pulmonary Edema Using Deep Learning and Image Segmentation

David Davila-Garcia[1], Yash Potdar[1], and Marco Morocho[1]

[†]DSC 180B: Section A14

14 March 2023

Abstract

Background

This project investigates the potential of convolutional neural networks (CNN) to identify cardiogenic pulmonary edema (CPE) from chest radiographs and NT-proBNP biomarker measurements. Building upon the methods used by Huynh in "Deep Learning Radiographic Assessment of Pulmonary Edema," our project aimed to expand upon his findings by modifying CNN architectures to intake additional parameters [1]. Specifically, we evaluated the impact of including (1) clinical data and (2) heart and lung image segmentation on CNN model performance: we hypothesized that incorporating these features would improve the ability of the CNN classifier to identify CPE.

Methods

Chest radiographs and clinical data obtained from 16,619 UC San Diego Health patients were randomly split into train, validation, and test sets at a ratio of 80%/10%/10%. We trained four modified ResNet152 CNN architectures with differing inputs: (A) Original Radiographs only, (B) Original Radiographs + Clinical Data, (C) Original Radiographs + Heart & Lung Segmentations, and (D) Original Radiographs + Heart & Lung Segmentations + Clinical Data. Early stopping was implemented to save the models with the minimum loss on the validation set (n = 1,662). The four model's accuracy and AUC on the test set (n = 1,662) were used to compare model performance.

Results

Model (B) had the highest accuracy (0.787) and AUC (0.869). Model (D) performed marginally worse than Model (B), with accuracy and AUC of 0.783 and 0.866, respectively. Models (A) and (B) performed considerably worse than both Models (B) and (D).

Conclusions

This project emphasizes the need to consider confounding factors, clinical data, and image segmentation when training CNN models for medical imaging classification tasks. While an

increase in CNN model performance was observed from adding clinical data, a performance increase was not observed in the models that included heart and lung segmentations. Further research is needed to determine the optimal use of image segmentations in CNN models.

1 Introduction

Pulmonary edema is a serious and potentially life-threatening condition with increased extravascular lung water [2]. The condition is most commonly caused by two major conditions: cardiogenic and noncardiogenic. Cardiogenic pulmonary edema (CPE) typically results from acute decompensated heart failure due to increased filling pressures in the heart's left chambers, causing pressure backflow to the lung vasculature [3]. Noncardiogenic pulmonary edema is caused by increased permeability of the lung capillaries resulting from injury, infection, or trauma.

NT-proBNP is a natriuretic hormone released primarily from the heart in response to increased chamber size due to increased filling pressures. Elevated levels of NT-proBNP can be used to determine if findings of pulmonary edema from chest radiographs are due to cardiogenic versus noncardiogenic causes [4].

However, NT-proBNP levels are influenced by confounding factors such as renal failure, age, sex, and BMI [5]. First, patients in renal failure tend to have higher concentrations of NT-proBNP due to increased intravascular volume [6]. As such, we incorporated laboratory measurements of Creatinine, which is the most widely used biomarker for measuring kidney function [7]. Second, patients with obesity tend to have lower concentrations of NT-proBNP compared to non-obese patients. However, it is not known whether this disparity is due to obesity or other confounding variables including age, sex, and heart failure severity [5]. For this project, we included BMI measurements for each patient so the CNN model could account for this confounding variable.

The typical chest radiograph findings of pulmonary edema range from mild to severe, including redistribution of pulmonary blood vessels, cardiomegaly, Kerley B-lines, peribronchial cuffing, consolidations, air bronchograms, and pleural effusions. These are the primary identifiers for examining chest radiographs to diagnose CPE.

CNNs have been effective in classifying diseases from medical images. However, the consistency and accuracy of manually labeled medical image datasets is an area of concern. In addition, highly accurate and generalizable CNNs for diagnosing diseases from medical images have a strong application, particularly in Low- and Middle-Income Countries (LMIC) where there may not be enough radiologists. To address these points, Huynh proposed the method of training CNNs to predict NT-proBNP concentrations; this biomarker provides a continuous and objective measure used for diagnosing CPE.

2 Literature Review and Prior Work

Huynh's study included 26,667 radiographs with a corresponding NT-proBNP value and 1,423 radiographs with a BNP value. These BNP values are blood serum biomarkers indicative of acute heart failure and cardiogenic pulmonary edema. The datasets were partitioned into 80% train,

10% validation, and 10% test sets, and the distribution of NT-proBNP and BNP values remained relatively constant across the sets. Based on scientific literature, the threshold for diagnosing cardiogenic pulmonary edema is NT-proBNP $\geq 400pg/mL$ [4]. The ResNet152v2 model trained on a two-stage training process of transfer learning. The first stage was used to infer NT-proBNP values, while the second layer used ResNet152v2 to infer BNP values. This model architecture was trained, and performance was measured across different image resolutions. The correlation between the true and predicted values peaked as smaller image sizes for both BNP and NT-proBNP. However, the AUC scores increased to 1024x1024, indicating greater CNN performance. Additionally, the study found that BNP predictions had stronger correlations to the measured BNP values than NT-proBNP predictions to the measured NT-proBNP values.

The scientific paper "Lung Field Segmentation in Chest Radiographs: a Historical Review, Current Status, and Expectations from Deep Learning" by Sanjeev Sofat et al. explored a modified U-Net architecture for biomedical image segmentation. The model was designed to effectively segment the lung field, clavicles, and heart using 18 convolutional layers. Although this model did not perform as well as an existing InvertedNet architecture, the researchers pointed out the shift from feature engineering to improving architecture engineering.

One of the primary challenges the researchers face is the limited availability of annotated medical data for training. To address this issue, they have looked to transfer learning and are also attempting to crowdsource large medical datasets with the assistance of radiologists. Despite the challenges, the researchers remain optimistic about the potential of deep learning techniques in medical image segmentation and hope to improve the accuracy and efficiency of these models.

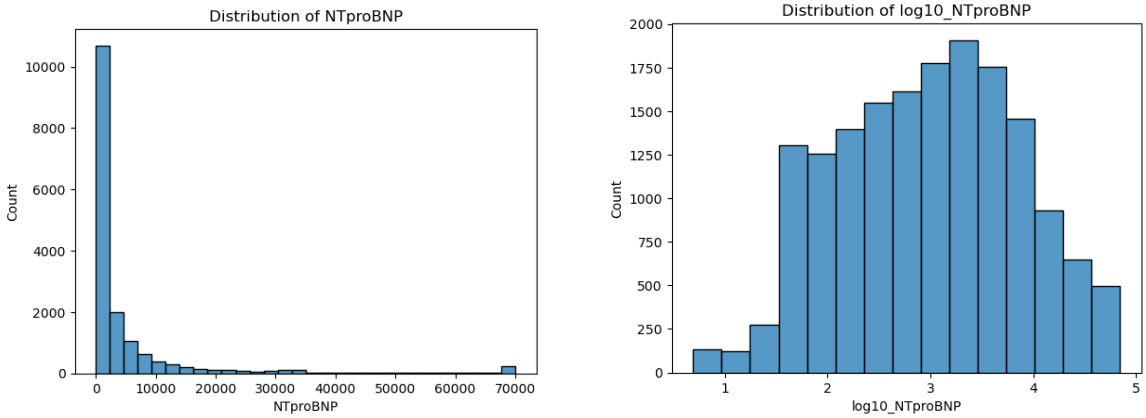
In the paper *Comparing different deep learning architectures for classification of chest radiographs* by Keno Bressem et. al., the researchers used 15 CNNs of five architectures: ResNet, DenseNet, VGG, SqueezeNet, Inception v4, and AlexNet. This project was conducted in a similar domain as our project, although the scope was broadened to include other diseases like cardiomegaly, atelectasis, and pleural effusion. They hypothesized that deeper CNNs do not necessarily have better performance than shallower networks. They noted that with the increased complexity of a CNN, the higher the number of resources required to train a model. To account for larger images, we can reduce batch sizes to avoid memory errors. They concluded that AlexNet, ResNet34, and VGG16, which are shallower CNNs, gave an accurate classification of radiographs. These results are insightful since it shows that with networks with fewer layers, we may have more hyperparameter tuning and a faster training process, while being able to efficiently handle larger images.

3 Data Description

We constructed a dataset of 16,619 records from UCSD Health patients. The NT-proBNP column represents the NT-proBNP value, a continuously valued biomarker measured from blood serum samples. The NT-proBNP values are bottom and top coded as their minimum possible value is 0 pg/mL, and the maximum is 70,000 pg/mL. As shown in Figure 1, there is a strong right skew for NT-proBNP values. Due to this, we followed Huynh's study and performed a log (base 10) transformation to create the log10 NTproBNP column. The histogram for log10 NT-proBNP in

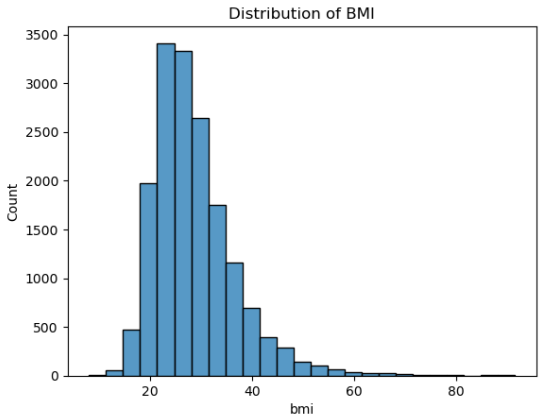
Figure 1 shows that the distribution is approximately normal due to the log transformation. Using the threshold for pulmonary edema consistent with academic literature, we classified patients with an NT-proBNP value of at least 400 pg/mL as positive for cardiogenic pulmonary edema and patients under the threshold as normal cases. Similarly, on the logarithmic scale, any records with a log10 NT-proBNP value of at least 2.602 are considered positive for cardiogenic pulmonary edema. These binary cases were encoded in the 'cardiogenic_edema' column. Around 64.7% of the records in our dataset had cardiogenic pulmonary edema based on this threshold.

The 'bmi' column contains the patient's body mass index (kg/m^2), derived from a patient's mass and height. The 'creatinine' column contains a continuous Creatinine concentration (mg/dL) measured from blood serum samples. The normal range for adult males is 0.7 to 1.3 mg/dL and 0.6 to 1.1 mg/dL for adult females [8]. The 'pneumonia' and 'acute_heart_failure' columns indicate the presence or absence of these physician clinical findings. In the dataset, 12.0% of patients have pneumonia, and 17.2% have acute heart failure. The dataset initially provided by the UCSD Artificial Intelligence and Data Analytics (AIDA) Lab had 18,900 deidentified records.

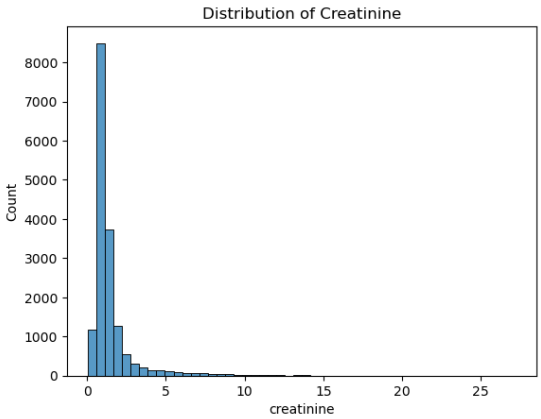


(a) NT-proBNP

(b) log10_NT-proBNP



(c) BMI



(d) Creatinine

Figure 1: Distributions of Quantitative Labels

4 Methods

4.1 Clinical Data

To ensure high-quality data for our project, we excluded patients with missing values for columns containing clinical data, specifically Creatinine, BMI, Acute Heart Failure, and Pneumonia. This exclusion was performed on the initial dataset of 18,900 patient records, resulting in 2,281 patients being removed from the final dataset used for model training, validation, and testing. The remaining dataset contained 16,619 patients. The potential mechanisms of missingness were not evaluated for the excluded patients.

4.2 Lung & Heart Image Segmentation

The UC San Diego AIDA laboratory provided a pre-trained U-Net CNN, which we used to create predicted binary masks of the right and left lungs, heart, right and left clavicle, and spinal column for each patient's chest radiograph. These masks were generated by applying the pre-trained model to the full dataset of patient chest radiographs. We then used the binary masks to create a new image of only the heart and another with only the lungs (Figure 2).

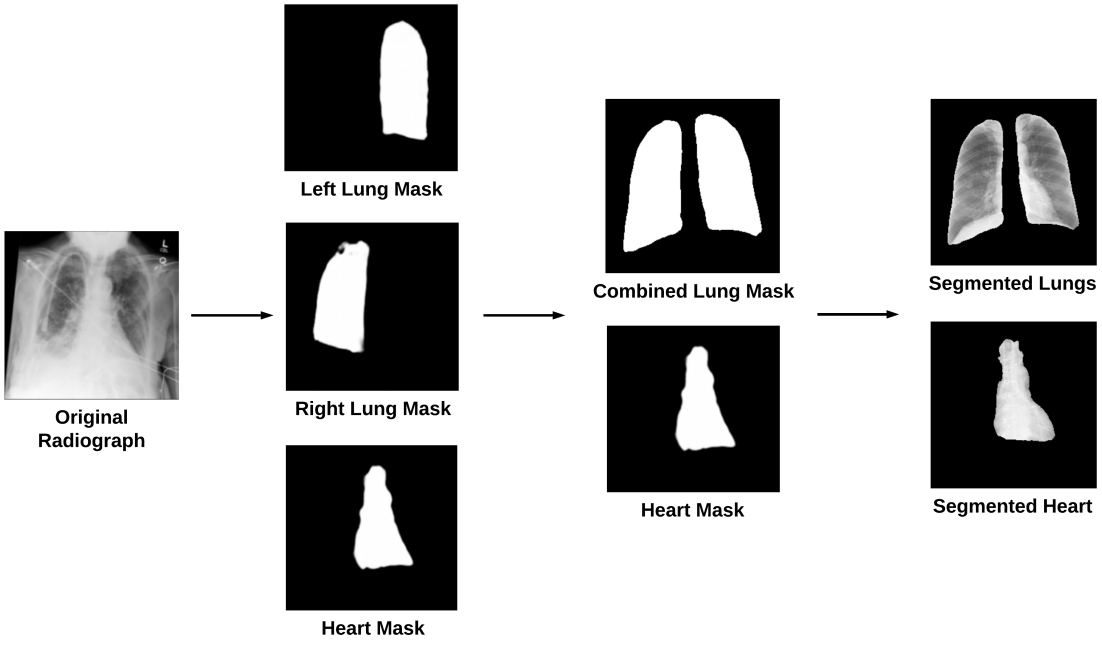


Figure 2: Heart and Lung Segmentation Diagram

4.3 ResNet152 Architectures

We used the default PyTorch ResNet152 model, with only one (regression) output class instead of ten (classification), to train four models. Model A took only the original radiographs as inputs. Model

B incorporated the clinical data by modifying the forward method of the model to concatenate the clinical data with the output from the convolutional layers. Model C took in the original radiographs, lung segmentations, and heart segmentations (three image channels) as inputs, passing each through the model's convolutional layers. Model D took in the original radiographs, lung segmentations, and heart segmentations (three image channels), and the clinical data as inputs. The forward method of the model was modified to concatenate the clinical data with the output from the convolutional layers (Figure 3).

4.4 Model Training/Validation

All four models were trained on the training set with a batch size of 12 for 20 epochs using the NAdam optimizer with a learning rate of 0.001 [hyperparameters: betas=(0.9, 0.999), eps=1e-08, weight_decay=0], and the mean absolute error (MAE) loss function. The MAE on the validation set was computed after each epoch, and early stopping was implemented such that the model with the minimum MAE on the validation set was saved. The overall purpose of the validation set was to optimize the model's parameters during training. (Figure 7)

4.5 Model Testing

After 20 epochs, the MAE on the unseen test set was computed for the four models with the minimum MAE on the validation set. We also saved each patient's predicted log10(NT-proBNP) values in the test set. We plotted these predictions against the laboratory-measured values for log10(NT-proBNP) and calculated the Pearson correlation coefficient (r). We used the threshold for cardiogenic pulmonary edema diagnosis (log10(400 pg/mL) = 2.602) to binarize each model's predicted values of log10(NT-proBNP) and calculated accuracy and AUC using the test set. We created confusion matrices and AUC-ROC curves for all four models and used these metrics to compare performance (Figures 4 and 5). The unseen test set aimed to provide an unbiased estimate of model performance after training.

5 Results

Table 1 includes all four ResNet152 model performances by input data, including their respective Train L1-Loss, Test L1-Loss, Accuracy, AUC, and Pearson R scores. This table highlights that Model (B) performed the best with an accuracy of 0.787 and AUC of 0.869.

Figure 4 displays the ROC curves of each ResNet152 model with different model inputs. Model (B) achieved the highest AUC score of 0.869 compared to Model A (AUC: 0.824), Model C (AUC: 0.828) and Model D (AUC: 0.866). Figure 5 presents the four outcomes to visualize the performance rates for each model. As previously mentioned, the matrices showcase that Model (B) outperformed all other models.

The performance of the models can also be seen in the Pearson correlation scatterplots, as shown in Figure 6. The red lines with the equation y=x represent a perfect regression model in which the predicted values perfectly align with the observed values. The scatterplots show that Models (B)

ResNet152 Model Performance by Input Data					
Input Data	**Train L1-loss**	**Test L1-loss**	**Accuracy**	**AUC**	**Pearson R**
Model A: Original X-rays	0.442	0.531	0.756	0.824	0.646
Model B: Original X-rays with Clinical Data	**0.392**	**0.455**	**0.787**	**0.869**	**0.739**
Model C: Original X-rays, Lung & Heart Segmentations	0.518	0.514	0.768	0.828	0.656
Model D: Original X-rays, Lung & Heart Segmentations, with Clinical Data	0.428	0.468	0.783	0.866	0.738

Table 1: ResNet152 Model Performance by Inputs

and (D) follow the red line most closely. The correlation coefficients for these two models are similar (r = 0.739 and r = 0.738, respectively), thus showing that they outperform the models that do not include Clinical Data.

6 Discussion

From the above results, it was apparent that a more complex and preprocessed input including segmentations did not appear to improve the ability of the classifier to identify cases of pulmonary edema. We hypothesized that image segmentation of X-rays would help focus the neural network on the lungs and heart regions, where cardiogenic pulmonary edema is most identifiable from radiographs. We proposed that by reducing noise in the input, the classifier could better distinguish cardiogenic pulmonary edema and normal cases. However, heart and lung segmentation did not improve performance, which is likely because the neural network was able to learn the necessary features from the original X-rays. Providing segmented images is not necessarily adding information to the ResNet152 model, but is instead cropping information out when it is passed to the network.

Another reason we believe segmentation was not as beneficial as we had hypothesized is due to the dataset size. By training on around 13,000 images, the neural network will most likely have sufficient data to distinguish normal and edema cases and determine that these differences lie in the heart and lung regions. If there were a smaller set of training images. In that case, segmentation might yield a better classifier due to the more focused input provided and the possible lack of the

neural network's ability to have enough data to focus on the heart and lungs.

Our results reinforce our hypothesis that the addition of confounding factors as features improves the performance of a classifier. This suggests that when creating a classifier, it is crucial to understand the features that truly correlate with the target feature. In this case, the clinical labels have shown that they have a high impact on the presence of CPE.

We were also able to create a model that performed better than the model in Huynh's paper trained on 256x256 images. The AUC obtained by the model in the paper was roughly 0.85, whereas our highest-performing model had an AUC of 0.869.

7 References

[1] J.Huynh, S.Masoudi, A.Noorbaksh, K.Hasenstab, and A.Hsiao,"Deep learning radiographic assessment of pulmonary edema: Training With Serum Biomarkers," in Proc. Med. Imag. Deep Learn., 2022.

[2] Milne, E. N. et al., (1985). The radiologic distinction of cardiogenic and noncardiogenic edema. *AJR. American journal of roentgenology*, 144(5), 879–894.

[3] Garan AR et al., "Pathophysiology of cardiogenic pulmonary edema", 2023.

[4] Welsh, P., et al., "Reference Ranges for NT-proBNP (N-Terminal Pro-B-Type Natriuretic Peptide) and Risk Factors for Higher NT-proBNP Concentrations in a Large General Population Cohort."

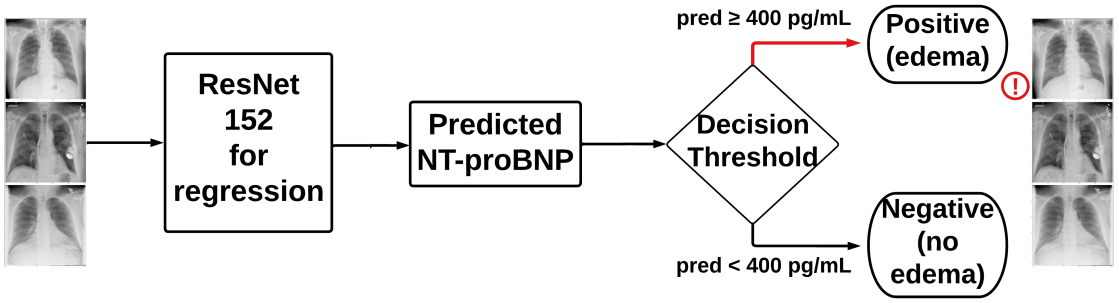
[5] Chen, H. H. et al., "Natriuretic peptide measurement in heart failure."

[6] Pornpen S. et al, "The Effect of Renal Dysfunction on BNP, NT-proBNP, and Their Ratio, American Journal of Clinical Pathology", Volume 133, Issue 1, January 2010, Pages 14–23.

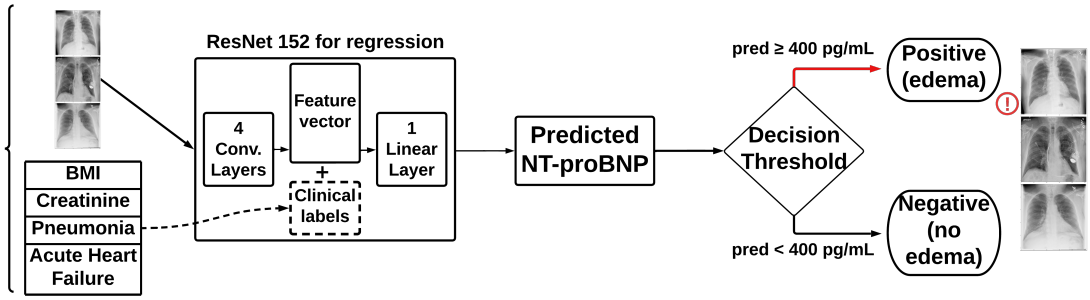
[7] *Creatinine blood test.* Mount Sinai Health System. (n.d.). Retrieved March 14, 2023, from https://www.mountsinai.org/health-library/tests/creatinine-blood-test.

[8] Sofat, S. et al., "Lung Field Segmentation in chest radiographs: A historical review . . . "

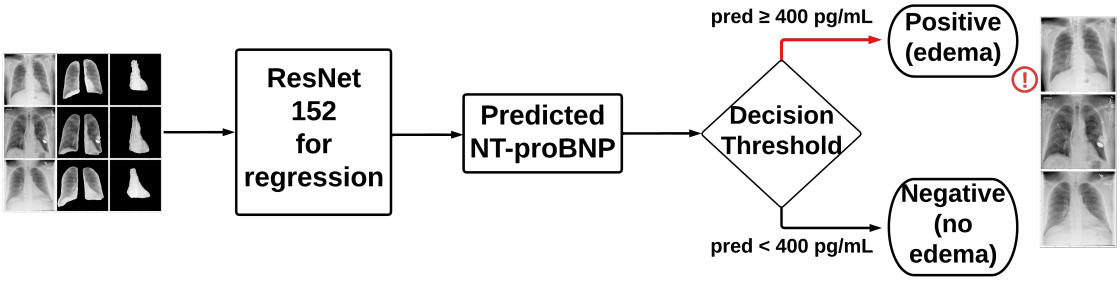
[9] Han-Na K et al., "Natriuretic peptide testing in heart failure". Circulation, 123(18):2015–2019, 2011.



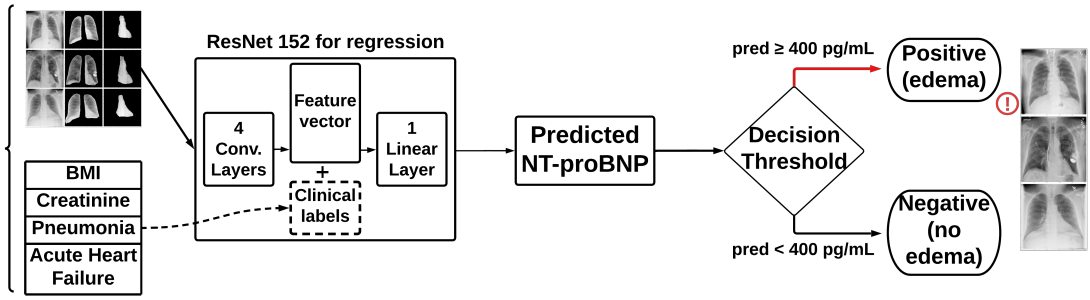
(a) Model A Architecture



(b) Model B Architecture



(c) Model C Architecture



(d) Model D Architecture

Figure 3: ResNet152 Architectures

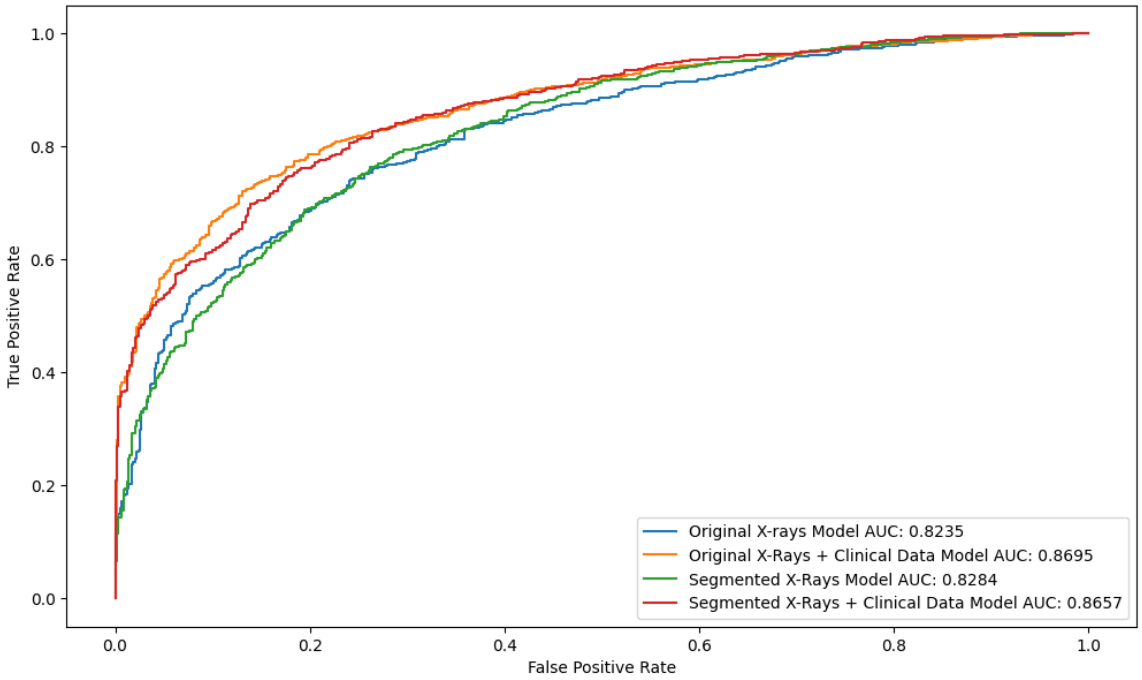
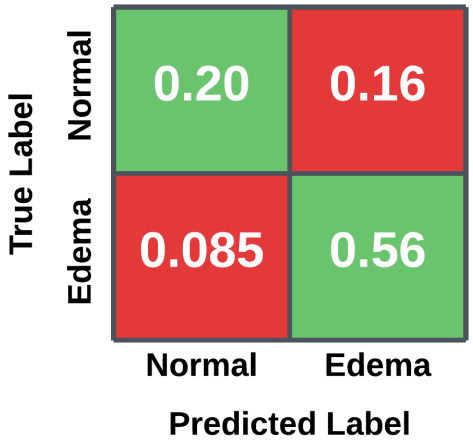
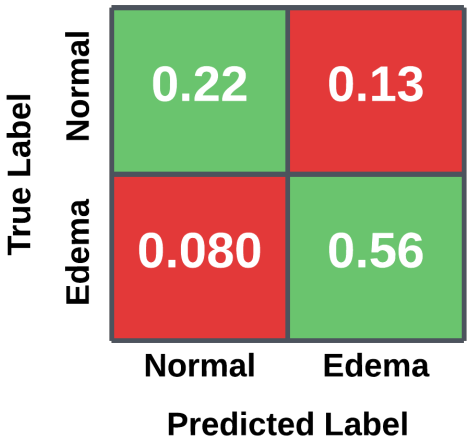


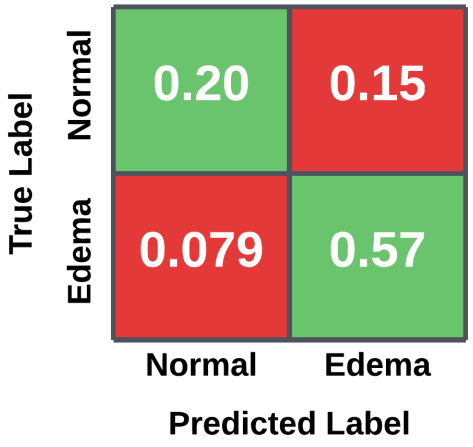
Figure 4: ROC Curves for each ResNet152 model by Input Data



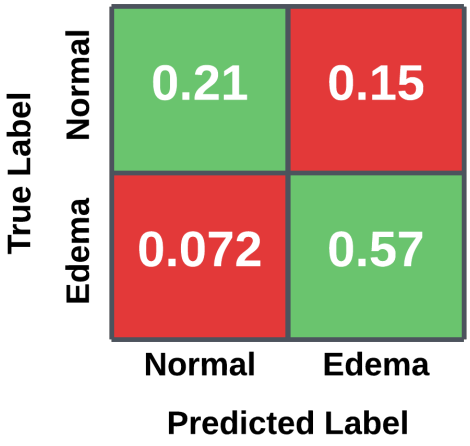
(a) Model A



(b) Model B

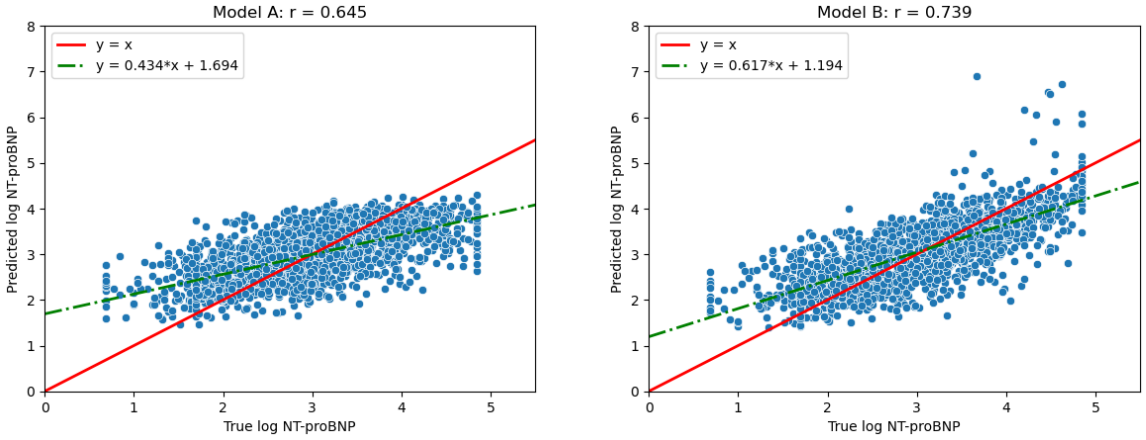


(c) Model C



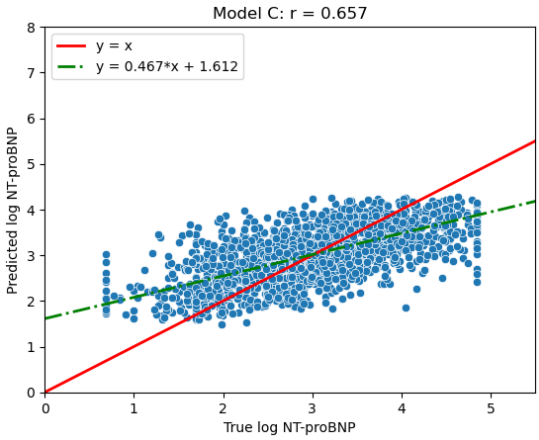
(d) Model D

Figure 5: Confusion matrices for each model

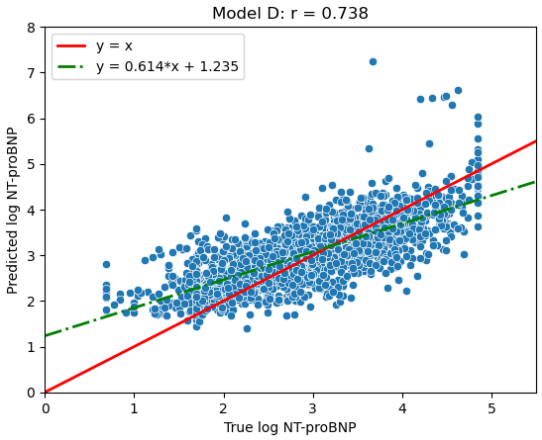


(a) Model A

(b) Model B

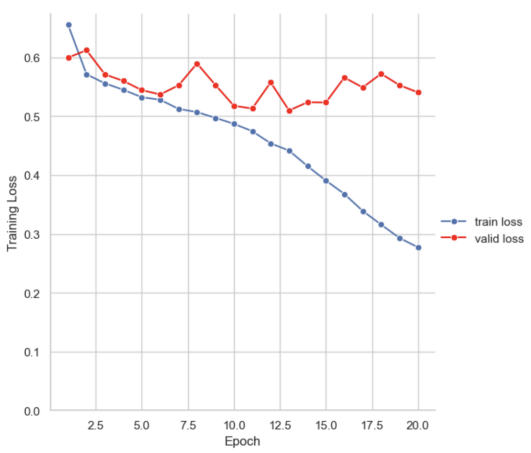


(c) Model C

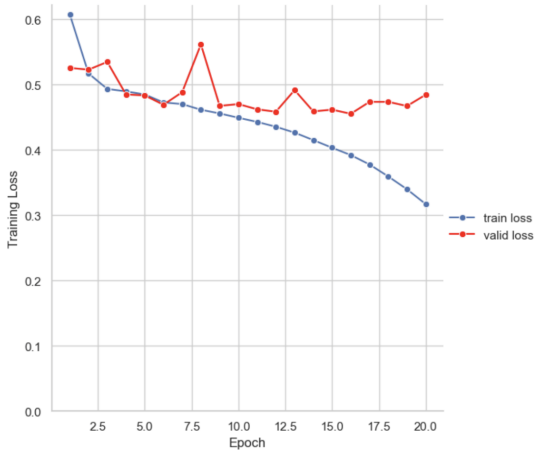


(d) Model D

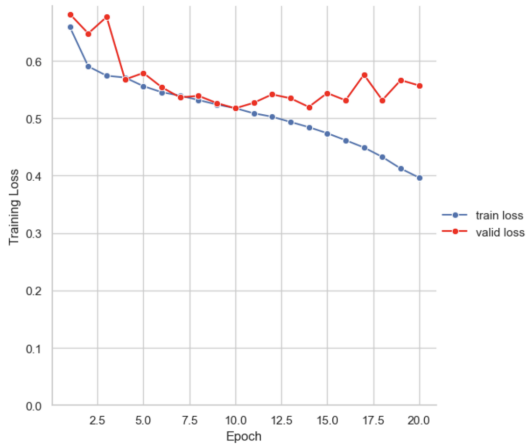
Figure 6: Scatterplots of Measured (x-axis) vs. Predicted (y-axis) values for log10(NT-proBNP)



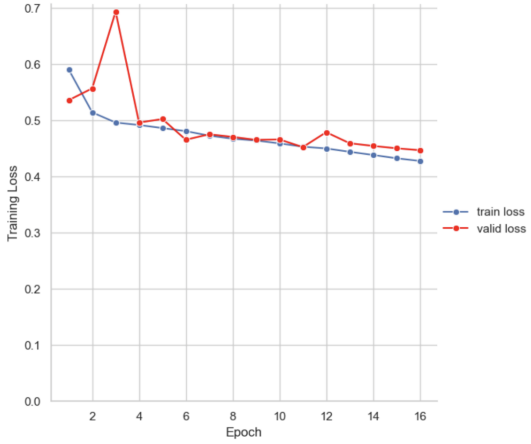
(a) Model A



(b) Model B



(c) Model C



(d) Model D

Figure 7: Train and Validation L1 Loss Curves for each model