Classification of Pulmonary Edema: Combining Physician Assessment and Machine Learning Inference, Integrating Neural Networks with Large Language Models

Dallas Plunkett dmplunkett@ucsd.edu
Jeru Paul Balares j1balares@ucsd.edu
Kendall Underwood keunderwood@ucsd.edu
Albert Hsiao a3hsiao@health.ucsd.edu

Abstract

Neural Networks (NN) have become increasingly useful for medical diagnosis; however, the NN predictions need a physician's assessment to be meaningful for diagnosis. Previous work developed a Convolution Neural Network (CNN) that predicted a T-pro B-type natriuretic peptide (BNPP) level, a clinical biomarker that indicates cardiogenic pulmonary edema, from X-ray images. However, this value would need a radiologist to assess the meaning of the BNPP level to determine the presence of pulmonary edema. Thus, we extend this work by integrating radiologist reports in order to provide meaning to the BNPP levels. In this work, we finetune a Large Language Model (LLM) to interpret radiologist reports to provide researchers a method to label the classification of pulmonary edema from the reports themselves without the need of a physician being present. Using these reports in conjunction with the predicted levels, we also investigated the threshold of the BNPP levels that determine the presence of pulmonary edema, yielding a threshold of 1027 compared to the previous consensus of 400.

Website: https://dallasplunkett.github.io/dsc-180a/
Code: https://github.com/dallasplunkett/dsc-180a

1 Introduction

Pulmonary edema is a condition characterized by the accumulation of excess fluid in the lungs, leading to impaired gas exchange and respiratory distress. Clinically, its presence is typically assessed using chest X-ray imaging in conjunction with BNPP levels as a biomarker indicative of pulmonary edema. While CNNs have been developed to predict BNPP levels directly from chest X-ray images, the clinical interpretation of these predicted biomarker values remains unclear. In particular, although a BNPP level of approximately 400 is commonly cited as a threshold suggestive of pulmonary edema, image-derived predictions do not inherently establish a clinically validated cutoff for diagnosis.

Prior work introduced a CNN capable of predicting BNPP levels directly from chest X-ray images. While this approach automates BNPP level estimation, it does not directly determine the presence of pulmonary edema nor clarify how CNN-predicted BNPP values correspond to clinically meaningful thresholds. In particular, the relationship between predicted BNPP levels and radiologist-confirmed pulmonary edema remains underexplored.

Because radiology reports are unstructured, variable in format, and ultimately do not provide a clear classification of pulmonary edema, we employ and finetune an LLM to extract and standardize the pulmonary edema status into a binary classification (present vs. absent), which achieved a precision of .90 for presence and 1.00 for absent. By comparing these classifications with the BNPP values, we clarify the validity of the LLM's predictions and investigate whether it aligns with the BNPP levels found in the X-Ray Images. We also find a different threshold from the general consensus of 400, as a byproduct of this investigation.

2 Methods

2.1 Dataset

There are two datasets used in this study. One dataset utilizes medical imaging data along with BNPP levels, while the other dataset consists of radiologist reports that correspond to those medical images.

2.1.1 X-Ray Images Dataset

The data for the X-Ray images included a BNPP level, the age of the patient, and the image itself. For the data preparation of the images and BNPP levels, we noticed in the dataset that the documented BNPP levels were highly skewed with values ranging from 0 to 70,000. To help reduce the variance, a logarithmic transformation was applied on the BNPP values prior to model training. This step that was taken follows the approach used by Huynh (2022). After the preprocessing process, the medical images' resolution size was 256x256 pixels. This helped ensure uniform input dimensions for the CNN.

2.1.2 Radiologist Reports Dataset

For the radiologist reports, the training set given had pre-existing labels attached. The labels attached are in a JSON format with uniform fields relating to pulmonary edema. For this study, the relevant field for training and testing the LLM will be the pulmonary edema presence field. Due to the significantly smaller size of the training set compared to the testing and validation sets, we decided to split the dataset to provide the model more training examples for a better performance. In addition, the dataset was skewed where the class labels were not evenly represented. Thus, we created the data to have a 50-50 representation for both present and absence of pulmonary edema. The dataset's pulmonary edema presence field consists of three results: present, absent, and unknown. We decided to drop the "unknown" class as it comprised only an insignificant portion of the whole dataset compared to the rest of the classes.

Note that while the reports are from the same source as the X-Ray Images (each X-Ray Image has a code connected to its corresponding radiologist report), the reports used for the training of our LLM are not of the same subset as the X-Ray Images Dataset. We will use the actual corresponding radiologist reports for inference after the LLM is trained (for our analysis of the intersection between reports and BNPP levels).

2.2 Convolutional Neural Network

2.2.1 Training

We implemented a CNN that utilizes a pretrained ResNet architecture. Learning was employed by initializing the network with ImageNet weights, which allowed the model to leverage previously learned lower level image features. Training was carried out using PyTorch with the Adam optimizer. A batch size of 4 was used along with a learning rate of 1e-5. The model was then trained for 16 epochs where the images were passed through the neural network multiple times. Mean Absolute Error (MAE) was used as the loss function.

2.2.2 Training Evaluation

The model performance for the CNN was evaluated by comparing the predicted BNPP levels with the ground truth (the BNPP levels attached with the X-ray images). The MAE was then calculated to measure the loss which we used to compare the different ResNet Models (ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152) to ultimately select a model for our work. We also used Pearson R to evaluate the performance of the model after selecting our model post-comparison.

2.3 Large Language Model

2.3.1 Training

We used the MedGemma 27B model which was pretrained exclusively on medical texts. We first conducted Zero-Shot Learning, a method where the model processes the data without any given examples or labels outside its pre-training. In this case, the MedGemma processed the given radiologist reports, given only a prompt for instruction. After conducting Zero-Shot Learning, we finetuned the LLM on the radiologist reports with the addition of the labels for pulmonary edema presence. We utilized Low-Rank Adaptation (LoRA) for better efficiency in time and computation and for better performance in LLM inference.

2.3.2 Training Evaluation

The model performance for the LLM was evaluated by comparing the inferred classification of the radiologist report with the ground truth (the classification labels attached to the radiologist reports).

3 Results

3.1 CNN - Image Signal

3.1.1 Model Selection

Because there are multiple versions of ResNet, we conducted several trainings to determine the model with the best loss. ResNet18, ResNet34, ResNet50, ResNet101 and ResNet152 were the models for selection. After testing, we chose ResNet34 as it had the lowest loss as seen in the following graphs.

Figure 1: Train Loss and Validation Loss of ResNet Models.

Figure 2: Loss Bar Plot to compare ResNet Models.

3.1.2 Model Evaluation

We further trained the ResNet34 model by increasing the epochs past the previous epoch of 16, yielding a Pearson R of 0.705. We also compared the distributions of the actual BNPP levels to the predicted BNPP levels as seen in the following graph, which reveals a somewhat bimodal distribution in the predicted BNPP levels compared to the unimodal distribution of the actual BNPP levels.

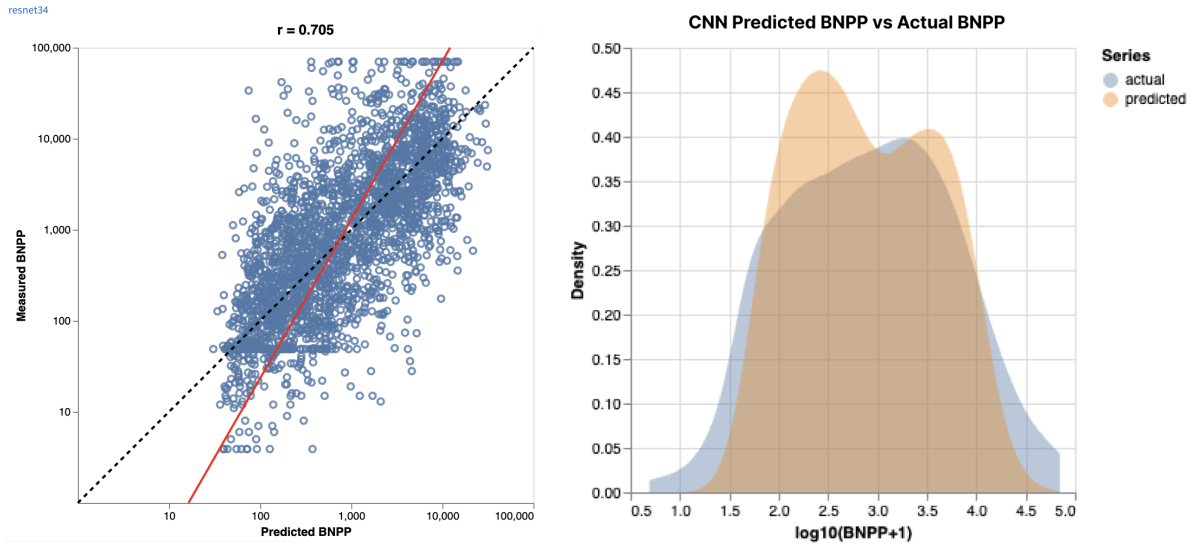

Figure 3: Pearson R of the ResNet34 model after increasing epochs in training (left) and the distributions of the predicted and actual BNPP levels (right)

3.2 LLM - Language Signal

3.2.1 Model Selection

The following confusion matrix is the result of our zero-shot learning using the MedGemma 27B Model before we dropped the “unknown” class.

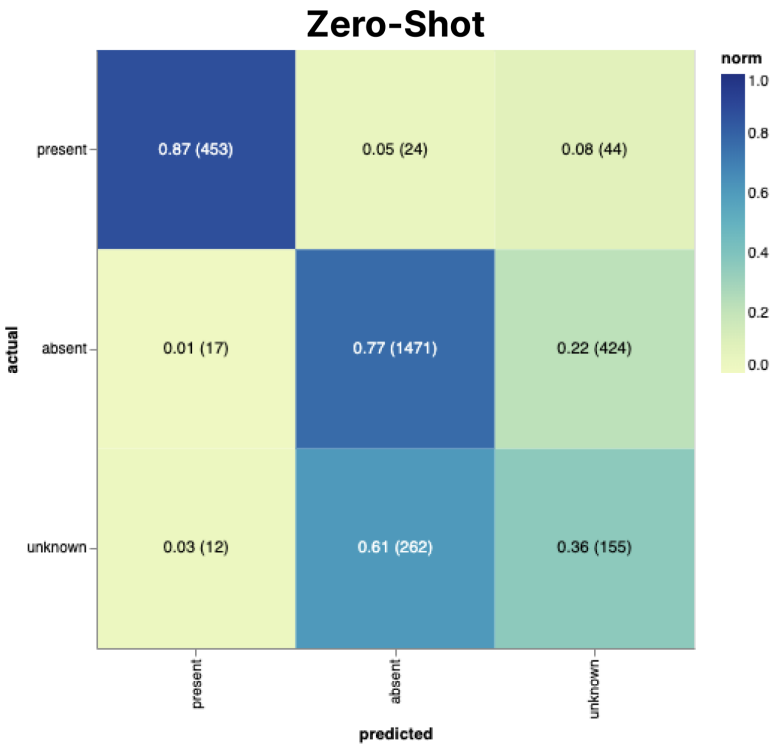

Figure 4: The confusion matrix of the result of the LLM’s labeling. As seen with unknown labels (the last row of the confusion matrix), the LLM very rarely labels an unknown as present, it is only ever between unknown or absent.

To determine the best ratio between the present and absent class labels for training, we conducted a comparison between the ratios of 50-50 and 75-25 using the following metrics: AUC (Area Under Curve), Youden J, and weighted F1.

Table 1: Results from Different Ratios Between Present and Absent Labels.

Ratio	AUC	Youden J	Weighted F1
No Change	0.76	0.39	0.88
50-50	0.79	0.44	0.98
75-25	0.77	0.43	0.96

3.2.2 Model Evaluation

After dropping the "unknown" class and choosing the ratio of absent and present in the training, we created a confusion matrix comparing our current solution to the zero-shot. The performance for both the zero-shot and our solution after the 50-50 data splitting improved upon our previous solution (See Figure 4).

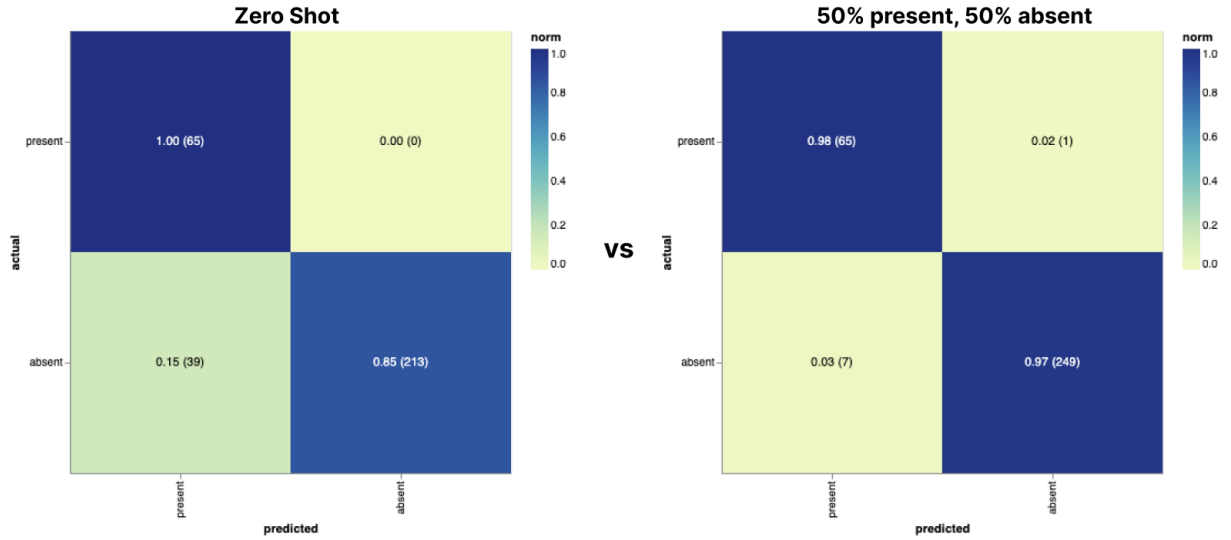

Figure 5: Although the predictions between the Zero Shot and the 50-50 ratio regarding the present label was slightly lower, the increased accuracy for the absent label was significant.

We then evaluated the model's precision and recall with the results being shown in the following table:

Table 2: Our LLM's Precision and Recalls Regarding Each Class Label

	Present	Absent
Precision	0.90	1.00
Recall	0.98	0.97

3.3 Interaction Between Signals

After training the CNN on the X-Ray Images, and tuning the LLM on the radiologist reports (again, remember that the training set for the reports are not connected to the X-Ray Images), we label the radiologist reports that actually correspond to the X-Ray Images. As a result, we find a threshold of a BNPP level of 1027, which is different from the general consensus of 400.

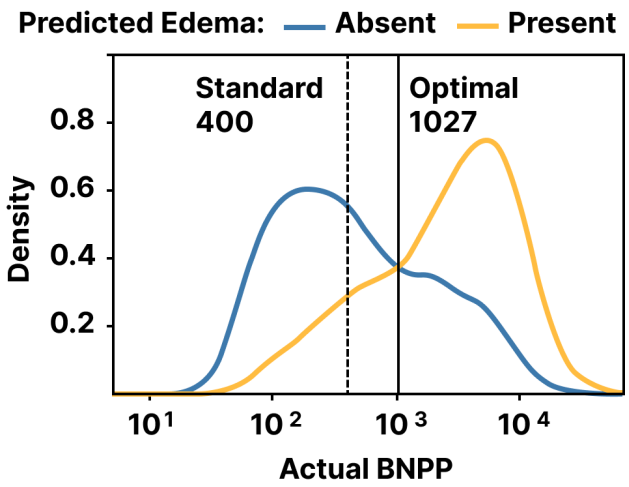

Figure 6: Youden J was used to find the optimal cut off point between the two distributions.

We also analyze the relationship between the BNPP levels (both ground truth and predicted) pulmonary edema labels among the different age groups.

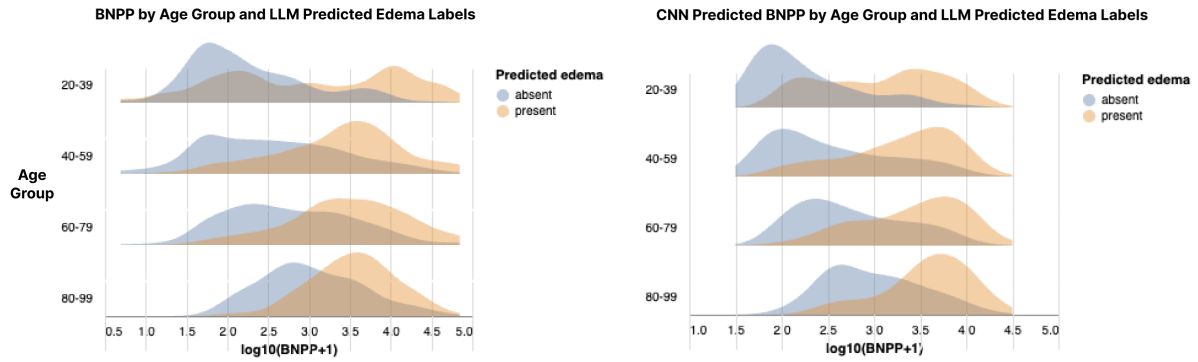

Figure 7: By using the labels from the LLM, we see the different distributions of the BNPP levels for presence and absence of pulmonary edema among different age groups. The distribution of the predicted BNPP levels follow a similar trend to the actual BNPP levels.

4 Discussion

As seen from Table 2, the precision and recall of the LLM's labeling of the radiologist reports shows that the LLM has performed well. This is also backed up with the fact that its labels seem to follow the trends of the distributions for both the predicted and actual BNPP levels, though the predicted BNPP levels themselves have a relatively weak correlation (0.7) with the ground truth. With these analyses, the LLM should be sufficient enough to be trusted to interpret the radiologist reports for researchers who are unable to interpret the reports themselves due to a lack of knowledge regarding medical terms. As mentioned before in the paper, a CNN can be used to automate the labeling of the X-Ray Images which helps

the efficiency of the pipeline, allowing radiologists to work on other tasks instead of being bottle-necked on a tedious task. A similar logic can be applied with the LLM. There is the possibility for the LLM to be used to automate the labeling of radiologist reports which would help researchers work on other tasks. This also provides an advantage for them to continue their work without being hindered by the absence of a physician.

As a byproduct of us analyzing the relationship between the LLM and CNN's results, we found a new threshold of 1027 from the generally agreed 400 BNPP level. While the threshold is different, it does not exactly mean that it is a better threshold. An increase of a threshold means that there is a chance that we can miss some positive pulmonary edema cases. This could showcase that the threshold of 400 is conservative to cover those cases.

A possibility to extend this work in the future is to further investigate this threshold and see if it is indeed an improvement or if the 400 is still the optimal threshold. In addition, we saw that the thresholds change in different age groups, increasing as the age becomes older. This difference in thresholds can be applied in physician work to encourage using different thresholds for different age groups instead of using a single number for all ages.

5 Conclusion

In conclusion, an LLM can be used to help automate the labeling of radiologist reports. Because of the medical terminology used in the reports, it can be difficult for researchers to label them without the help of a physician. This can be a bottleneck which the LLM can help alleviate. It is shown that different thresholds correspond to different age groups which could be a better application in using thresholds for pulmonary edema classification. The labels provided by the LLM, while coupled with the CNN-predicted BNPP levels, can be helpful in pulmonary classification and its research.

References

[1] Ozturk, Tuba Cimilli et al. "Can NT-proBNP be used as a criterion for heart failure hospitalization in emergency room?." Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences vol. 16,12 (2011): 1564-71.

[2] J. Huynh et al., "Deep Learning Radiographic Assessment of Pulmonary Edema: Optimizing Clinical Performance, Training With Serum Biomarkers," in IEEE Access, vol. 10, pp. 48577-48588, 2022, doi: 10.1109/ACCESS.2022.3172706.

Appendices

A.1 Additional Figures

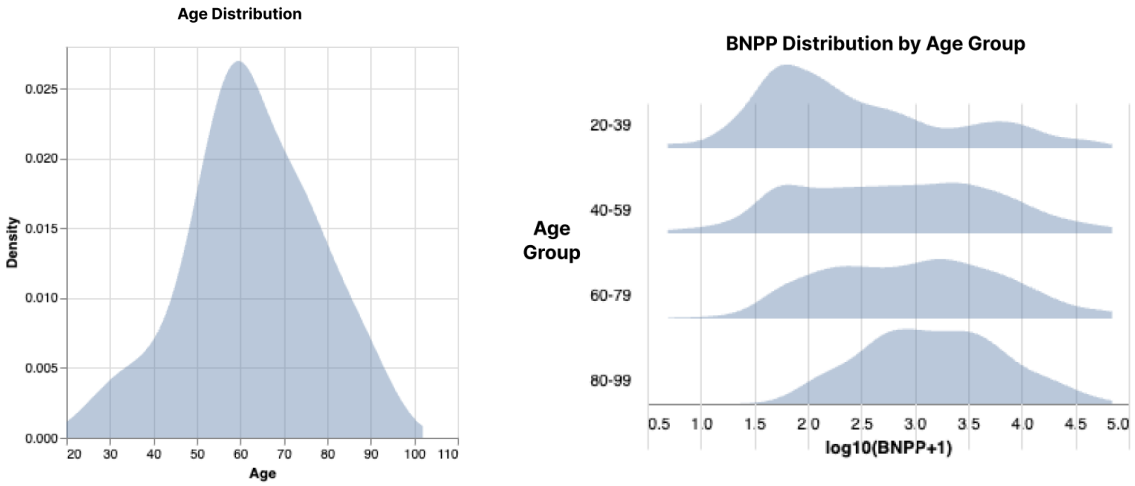

Figure A 1: Age distribution of the dataset that we used (left) with a separate graph showcasing the distribution of BNPP levels among different age groups (right).

A.2 Additional Tables

50% present, 50% absent Performance

Counts and Proportions

	total	present	absent
all	3015	681 (22.6%)	2334 (77.4%)
train	1094 (80%)	547 (50.0%)	547 (50.0%)
valid	303 (10%)	68 (22.4%)	235 (77.6%)
test	322 (10%)	66 (20.5%)	256 (79.5%)

Category Performance

	present	absent
support	66.00	256.00
precision	0.90	1.00
recall	0.98	0.97

Variable Performance

	model	low	random	high	lift
micro	0.98	0.63	0.67	0.72	1.45
macro	0.96	0.45	0.50	0.56	1.93
weighted	0.98	0.64	0.67	0.71	1.45

Figure A 2: A screenshot of metrics and additional information regarding the LLM's performance.