Problem Set 4 (56 Points)

MLHC 2025

April 8, 2025

Submission Instructions

Due 4/17/2025, 23:59 on Gradescope

Very Important: For each question (and subquestion), please submit the code for it in the pdf so we can quickly verify your results. Meaning for question 1.1 (and all others), show your code cells and then write your answer.

Please make a submission only if you are registered as a regular student. Submit your write-up as `[mit_email]_pset4.pdf` (*e.g.,* `sophiejg_pset4.pdf`), with all written work in a single PDF file following the problem set structure. In addition, please append your full code at the end of the report. For those who prefer to typeset in LaTeX, the source code of this file is here .

Submit your own work and note any collaborators - if none, state "Collaborators: none." If using external sources, cite them properly. You should be able to explain your solutions verbally. Please do not share your code or report with anyone inside or outside of the class, nor post them publicly online. Course staff welcomes any questions about these policies. Please see the "Late Policy" on the course website (`https://mlhcmit.github.io/`) Make sure to specify how many slack days you are using in your pdf writeup.

In this problem set, we will work with a subset of the MIMIC-CXR (1) dataset, distributed through PhysioNet as in previous PSets. We will present you with a simulated experience of evaluating a potential model for deployment. In this case, the model in question (which we will provide to you, fully trained) is being evaluated for its use in the detection of Pneumothorax (PTX); a potentially dangerous condition often colloquially known as a collapsed lung. Ultimately, your desired use case for this model is to use it as a diagnostic aid for new patients to the ED.

However, being an experienced MLHC researcher, you're concerned that perhaps this model is not actually leveraging viable diagnostic information to make its PTX decision–instead, it may be confounded by other factors in the data.

Colab notebook https://colab.research.google.com/drive/139utghM_52rL0dO6s5mssndmNEaxhbag?usp=sharing

1 Error Auditing [12 points]

In this first problem, we will try to understand the data and the performance of our model on a high level.

1.1 Compute Accuracy and AUC on the evaluation set. Do you think this AUROC is good, or bad? If you saw this on a model used to treat patients, how would you feel about that? (3 points)

In the notebook, a pandas dataframe named 'evaluation_set' contains a column 'Pneumothorax' with the labels, use that and the output of "eval_model".

1.2 Describe but do not implement one way to compute confidence intervals for each of AUC and accuracy (2 points). Bonus (+2) points: if you implement and compute the confidence intervals correctly.

Feel free to look up different methods online and briefly summarize them.

1.3 Compute the confusion matrix of our classifier (FP,FN,TP,TN) and the class balance (% of labels that are 1). Compute rates instead of counts (i.e. normalized confusion matrix) (2 points)

1.4 Investigate the chest xrays and reports in the evaluation set, the "display_study" function is helpful here. Report some of your observations on the variation of images. Report some of your observations on the notes with pneumothorax. Do you notice any patients who were labeled as pneumothorax but who might not be so? Can you say why the label is incorrect and contradicts the report? (5 points).

In this question we are not looking for specific observations, but want to see an effort in you looking at the data.

2 Interpretability Visualizations [15 points]

One serious concern is that the model may not be diagnosing Pneumothorax at all, but instead largely relying on a treatment-based confounder, such as the presence of a "pigtail cathater", a small-bore tube used in the case of pneumothorax to allow reinflation of the lung or escape of air from the pleural space (note that pigtail cathaters have numerous other uses as well, but this is the use case of interest to us here), or a chest tube. Both of these are very evident on an x-ray, and are strong indicators that the patient did at one time have Pneumothorax; however, predicting them offers no diagnostic benefit to us as they are only seen if the Pneumothorax is already known and under treatment. For the rest of our analyses, we'll focus on trying to determine if this is the case or not. First, we'll look at interpretability analyses via Saliency maps and Class Activation Maps.

2.1 Implement Saliency Maps. In the code, the body of the saliency map is empty but there are TODOs we have written for you to follow. Please also copy the code here for your solution (10 points)

2.2 **Visualize different images. Use the 'plot_interpretability_measures' to plot images with both intepretability viz methods for images that are true positives and images that are true negatives. Do regions highlighted by the saliency map and CAM suggest that the model may be largely leveraging treatment devices, such as pigtail catheters and chest tubes, rather than diagnostic criteria for predicting pneumothorax? You may also look at FN and FP cases. [5 points]**

3 Leveraging Radiology Reports [19 points]

Another oft-understated tool to understand a model's decision making process is to leverage additional, related modalities of data to better explore the underlying data and determine if there are any unexpected confounders in the data. Here, we'll leverage the radiology reports co-released with these images in MIMIC-CXR to do just that.

3.1 Using the reports in evaluation set, report the AUC and accuracy of the model by gender and the count by gender (how many in each gender). You will need to extract gender from the reports, if for a report you are unsure of the gender, label it as unknown and evaluate as its own category. [4 points]

3.2 read below for 3 parts [10 points]

1) Find the top 10 words that appear the most in the reports that have pneumothorax compared to the reports that are labeled 0.
To do this, for each word that exists in the positive reports, find how many times it appears in total, then subtract how many times it appears in the negative reports.
2) Do the same thing but now for the top 10 words that appear the most in reports without pneumothorax compared to reports with.
The CountVectorizer from sklearn with unigrams will be useful: "cv = CountVectorizer(ngram_range=(1, 1), binary = True)".
3) What do you note? Do these terms appear to be diagnostically relevant to Pneumothorax? Or are they more about other medical concepts? Are there any relationships related to treatments, as we fear?

3.3 Evaluate the AUC of the model on the reports that include either of the words "drain", "pigtail", "catheter" or "tube". Then evaluate the AUC of the model on reports that don't include all of the four words. What does this suggest about what is guiding our model's performance? [5 points]

3.4 [Optional, Bonus 2 points] Describe but do not implement a scheme that can allow the model to learn how to diagnose without relying on the treatment indicator.

4 Zero-Shot Image Classification using BioMedCLIP [10 points]

While we use a traditional convolutional model in Problem 1, we were required to pre-train the model on existing images for our specific task in order to achieve such performance.

Here, we will evaluate the performance of BiomedCLIP (2), a multimodal foundation model on zero-shot evaluation (no pre-training) of our pneuomothorax classification task.

While the original CLIP model was trained on 400 million general internet image-text pairs to enable zero-shot classification by matching images with text descriptions, BiomedCLIP extends this approach to the biomedical domain. It was trained on PMC-15M, a massive dataset of 15 million biomedical image-text pairs collected from 4.4 million scientific articles, which is two orders of magnitude larger than existing biomedical multimodal datasets. BiomedCLIP incorporates domain-specific adaptations including using PubMedBERT for text encoding and larger vision transformers to better handle biomedical imagery.

The model achieved state-of-the-art results across various tasks including cross-modal retrieval, zero-shot image classification, and visual question answering. Here, we will be benchmarking the model on it's zero-shot image classification task using our MIMIC-CXR evaluation dataset.

4.1 Compute and report the Accuracy and AUC on the evaluation set. How does this compare to the VGG-16 model? Justify the model performance. [6 points]

4.2 Experiment with 4 different variations of template and labels, including the one you used in 4.1. Hint: Look at supplementary table 10 in the BioMedCLIP paper for inspiration. Report each variation and their respective metrics. Which variation has the best performance? [4 points]

References

[1] Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-Ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019.

[2] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025.