How Well Do Bio-Markers Work for CNN Training: A Comparison of Model Results

Jacob M. Ryan

Halıcıoğlu Data Science Institute, University of California, San Diego, La Jolla, CA 90293, USA

(Dated: March 14, 2023)

I. ABSTRACT

Utilizing correlative biomarkers from inexpensive blood tests as ground truths could decrease the cost of obtaining training data for deep learning dramatically. This work aims to analyze the success of utilizing BNPP levels as a heuristic for pulmonary edema for training data by testing these chest radiograph models on radiologist confirmed pulmonary edema radiographs from a public dataset. It also explores the use of anatomical segmentation for images to help focus the model on biologically important aspects of the image.

II. INTRODUCTION

Medical imaging requires highly specialized analysis by imaging professionals. This process is costly and time consuming. In recent years, there has been a push to use machine learning and a form of Artificial Intelligence called Deep Learning to analyze medical images without the need of specialized persons involvement. These models have gained popularity as their effectiveness are proven time and time again. However, the training process for these models requires tens of thousands of medical images and their lab diagnoses.

In an effort to reduce the burden on training the models, this paper utilizes biomarkers from blood streams that are representative of pulmonary edema, instead of actual pulmonary edema diagnoses. This work expands on past work performed by AiDA Lab ([2]), which found success in predicting biomarker levels from blood tests using chest radiographs. These blood tests are far cheaper and faster. Therefore, this paper presents models trained on these biomarkers, compares different training techniques using conventional statistics, and implements saliency techniques. Furthermore, this paper explores the success of this model on publicly available datasets. Additionally, this paper utilizes a lung segmentation model to create new lung-and-heart only images to be used for a new training model. The goal of this new model is to compare the effectiveness of model prediction on using focusing attention on key anatomical elements we are predicting diagnoses for.

III. METHODS

A. UCSD Data

Utilizing the dataset provided by AiDA Lab at UCSD Health ([2]), there was a total of 16,000 chest radiograph images, each with a correspond BNPP value. The mean and standard deviation of BNPP values across the set was 4919 ± 11039 pg/mL. This total set was randomly split into train and validation sets with an 80%, 20% split, respectively. Using a threshold of 400 pg/mL, the BNPP values are classified into positive and negative classes as a heuristic for pulmonary edema. The set contained 64.6% positive labels. These images were downsampled to a size of 224 by 224.

B. MIMIC-CXR Data

MIT's MIMIC-CXR study provides their data publicly ([3]). 22,000 chest radiograph images were sampled from their set. The set contained 48.7% positive labels. These images were downsampled to a size of 224 by 224.

C. Image Segmentation

All images went through a U-Net based anatomical segmentation model, kindly provided by AiDA Lab ([2]). The segmentation model provides a heart, left lung, and right lung segmentation. The lungs were combined into a singular image and then the lungs and the hearts were added as separate channels to create segmentation datasets.

D. Model Training

All Neural Networks used were built with a ResNet 152v2 pre-trained on ImageNet ([1]). Each model was trained using at max 30 epochs (Early Stopping based on change in validation AUROC), a static learning rate of 1e-5, a weight decay of 0.9, and a batch size of 32. The loss function utilized is weighted cross balanced entropy.

All layers of the models were unfrozen for training. Models were analyzed using conventional statistics including Area Under the Receiver Operator Curve (AUROC), Area Under the Precision Recall Curve (PRC), and Accuracy.

E. Model Evaluation

After training, each model was tested on the held back MIMIC data. The same conventional statistics were applied to the test sets. The best performing model was then further analyzed using XRAI, a saliency technique ([4]).

Each model was evaluated based on Area Under ROC (AUROC) and a Precision-Recall Score (PRC).

IV. RESULTS

In total, four models were trained, two UCSD training data based (one full image, one segmentation) and two MIMIC training data based (one full image, one segmentation). All four models were evaluated on the same withheld MIMIC test set.

For each style of images (full or segmented), MIMIC trained models outperformed their respective UCSD biomarker trained models by a considerable amount, allowing for the rejection of Hypothesis 1 (VII, 1, 2). Currently, models trained on radiologist ground truths outperform models trained on heuristic biomarkers for the prediction of pulmonary edema presence.

The model trained on MIMIC segmented images performed the best overall. However, between the UCSD trained models, the segmented images did not improve performance more than the full images.

The best performing model (MIMIC based segmentation) was further analyzed using XRAI ([4], 3).

V. DISCUSSION

From these results, it appears that biomarkers as a heuristic are not a quality substitute for radiologist confirmed labels for the identification of pulmonary edema. The UCSD biomarker trained models consistently under-performed compared to their MIMIC radiologist confirmed data models. The differences in each performance statistic are vast, MIMIC based models were at least 10% better.

There are a few explanations for these results. First, biomarkers are a good correlative feature for pulmonary edema identification, but with enough error to affect model performance significantly. In this case, radioligist confirmed ground truths are to be preferred. Second, models trained on biomarkers need additional hyperoptimization to perform at the level of models trained on radiologist confirmed ground truths. This experiment trained all models using the same techniques for fairness; however, this could have the exact opposite effect by not giving enough accommodation to a model that is performing a harder task. Third, models trained on the MIMIC set were also tested on the MIMIC set, whereas UCSD biomarker models were tested on a different set (MIMIC). It is possible that there is a homogeneity across MIMIC set and a different homogeneity across UCSD set due to the actual radiographic images themselves. MIMIC trained models would have the advantage of learning those features across the dataset, while UCSD trained models would be punished for learning the features inherent to the UCSD set. Overall, In this experiment, Hypothesis 1 can be rejected.

Alternatively, segmented images had a nonzero improvement in MIMIC based models, while it had a negative effect on UCSD biomarker based models. Interestingly, MIMIC trained models performed better than UCSD trained models. It could be possible that segmentation helped improve the models that better represented the the feature space of pulmonary edema by targeting focus to the anatomical structures that mattered. On the same note, the UCSD models did not represent the pulmonary edema feature space appropriately and thus segmentation may have led it further astray. Of course, the improvement in MIMIC based models and the worsening of UCSD models due to segmentation was minimal and could thus be completely random.

For both experiments, future work is necessary. Better training procedures for biomarker based models would most likely improve results. Additionally, the results for segmentation are unclear. Further experimentation is necessary to see if improvement was an artifact of randomness or a significant result.

VI. REFERENCES

[1] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[2] Huynh, J., Masoudi, S., Noorbakhsh, A., Mahmoodi, A., Kligerman, S., Yen, A., Jacobs, K., Hahn, L., Hasenstab, K., Pazzani, M., et al. (2022). Deep learning radiographic assessment of pulmonary edema: Optimizing clinical performance, training with serum biomarkers. *IEEE Access*, 10:48577–48588.

[3] Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019). Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

[4] Kapishnikov, A., Bolukbasi, T., Viégas, F., and Terry, M. (2019). Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957.

VII. FIGURES AND TABLES

	UCSD Full Image	MIMIC Full Image	UCSD Segmented Image	MIMIC Segmented Image
AUROC	0.755	0.877	0.754	**0.889**
PRC	0.590	0.799	0.559	**0.816**
Accuracy	0.652	0.793	0.616	**0.802**

TABLE I. Performance of all four models on withheld MIMIC test set.

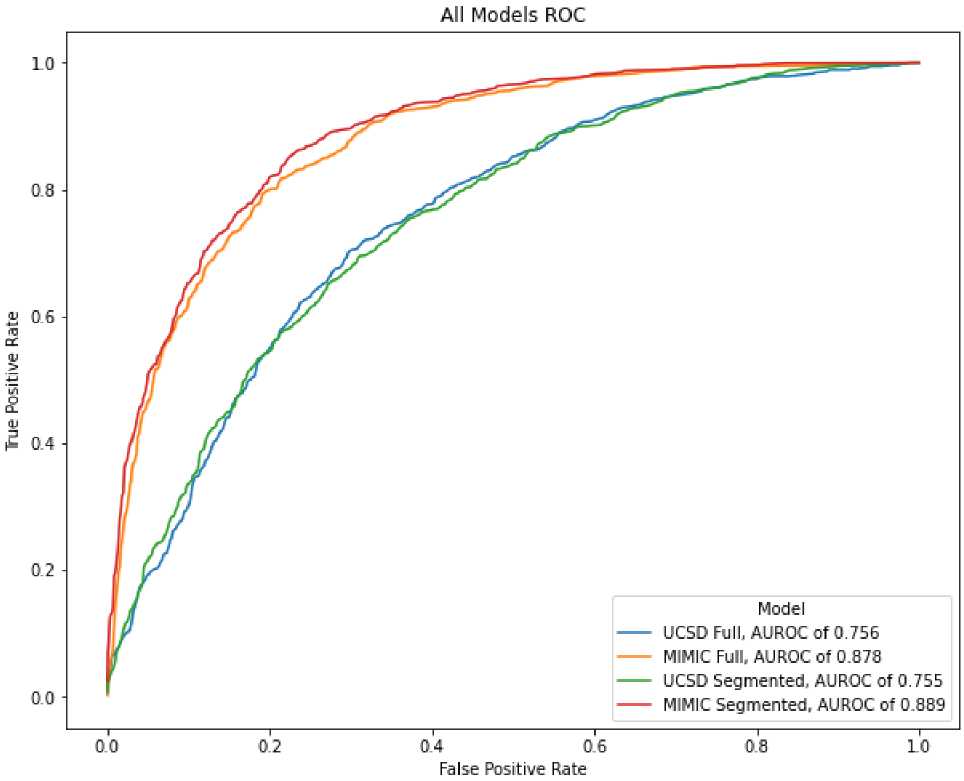

FIG. 1.

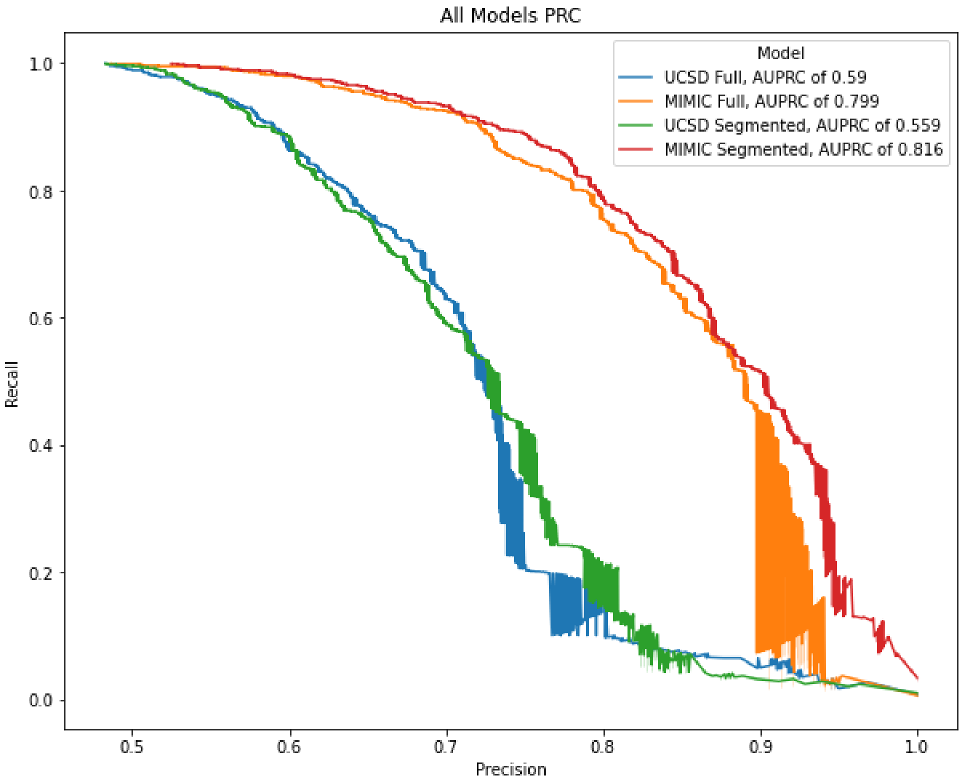
All Models PRC
Model
UCSD Full, AUPRC of 0.59
MIMIC Full, AUPRC of 0.799
UCSD Segmented, AUPRC of 0.559
MIMIC Segmented, AUPRC of 0.816
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.5
0.6
0.7
0.8
0.9
1.0
Precision

FIG. 2.

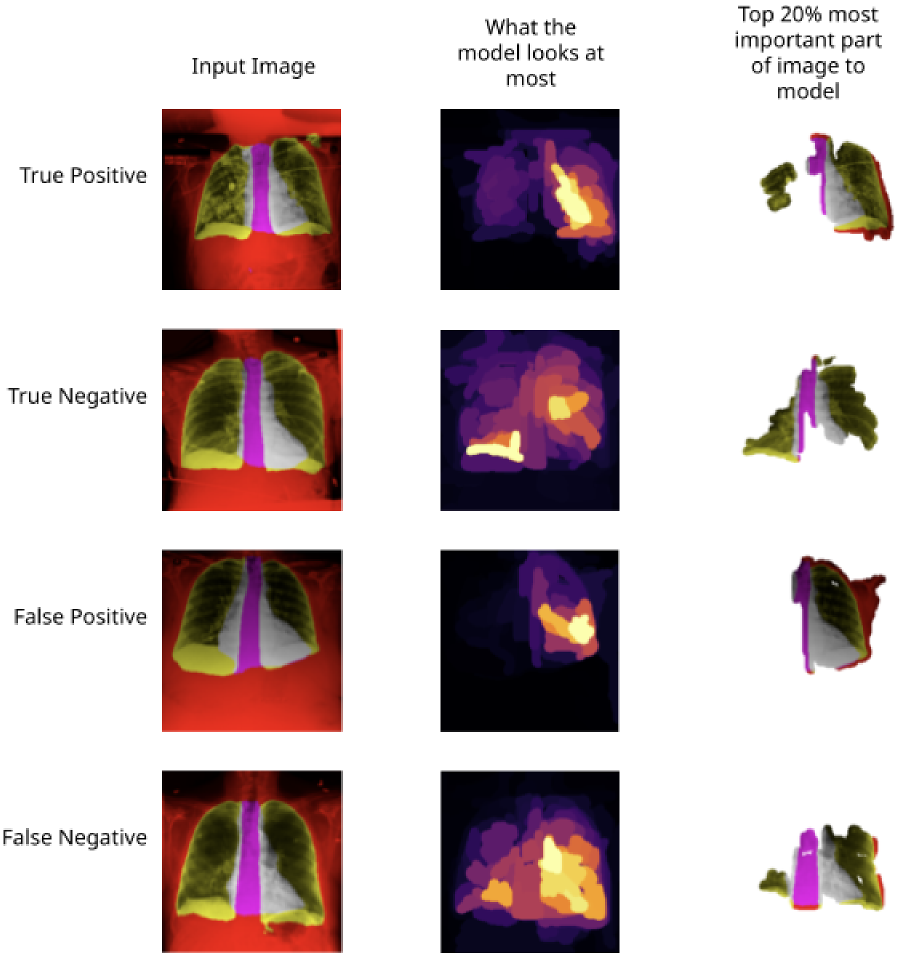
Input Image
What the model looks at most
Top 20% most important part of image to model
True Positive
True Negative
False Positive
False Negative

FIG. 3.