Analysis of Confounding Variables Effecting CNN Performance on MIMIC-CXR

Niki Vasan
Quantitative Theory and Methods
Emory University
Atlanta, GA
niki.vasan@emory.edu

Mert Ozbay
Computer Science
Emory University
Atlanta, GA
mert.ozbay@emory.edu

Zach Zaiman
Computer Science
Emory University
Atlanta, GA
zzaiman@emory.edu

***Abstract*—This paper seeks to build upon previous applications of deep learning models in healthcare by evaluating potential biases in a convolutional neural network (CNN) trained to classify comorbidities in chest X-rays using the MIMIC-IV and MIMIC-CXR datasets. The model is evaluated in terms of overall raw performance and stratified performance along five different clinical and demographic confounders relevant to healthcare: race, age, sex, ICU status and insurance type. We find that the model performs better on average for male patients and older patients, which is reflective of skews in the training data, but find inconclusive results when stratifying by ICU status. This study demonstrates that diverse training datasets are important when building models for clinical deployment, as representation biases in the data are indeed reflected in performance. Further research is needed to confirm possible biases with regards to race, ICU status and insurance type. The code and other resources are available at our github repository.**

I. Introduction

Artificial Intelligence (AI) and more broadly Machine Learning (ML) provide powerful tools that have the potential to greatly enhance the efficacy and efficiency of various processes within many industries [17]. When properly trained and used correctly, these models can improve an organization's resource utilization, and aid in human worker's day to day tasks. One field that has witnessed steep growth in AI research and development over the last ten years is healthcare [31]. The use of these models in a healthcare setting can greatly benefit the industry and a patient's quality of care. Models can be developed to predict a disease based on a patient's symptoms [18], likelihood of an intervention given a patient's past medical history [5], or even a comorbidity based on a chest X-ray (CXR) [24]. Despite the potential benefits, models of this nature, specifically deep learning (DL) models, are referred to colloquially as "black box models" [19] since it is difficult to explain what factors contribute to such a model's prediction on a given input sample. There are many instances where DL models have been found to learn artifacts or encoding of confounding variables as opposed to its intended task. For example, a study published in 2016 by Ribeiro et al. discovered that a convolutional neural network (CNN) trained to differentiate images of wolves and images of huskies [21] was actually learning an artifact in the images of huskies: snow. Every image of a husky had snow in the background, so the model was learning the presence of snow rather than the image of a husky. While this is just a trivial example of this phenomenon, similar scenarios have been shown to occur in the healthcare space [7]. Because of this, it is extremely important that all models trained with intention of clinical deployment be evaluated not only in terms of raw performance metrics, but also in relation to explainability and bias.

In this project, we attempt to conceptualize and analyze many of the possible confounding variables that could influence a DL model trained to identify various comorbidities on chest X-rays. To do this, we first attempt to replicate existing studies that have shown good performance training a CNN model on chest-Xrays [24]. After iterating through that process and picking the best model, we analyze our results globally and also through subgroups on a subset of many demographic and clinical features that may cause the model to be biased.

We use the MIMIC-IV dataset [11] to analyze these biases. The MIMIC IV dataset is a multimodal healthcare dataset from Beth Israel Hospital in Boston, MA. More specifically, it contains tabular information about patients on a hospital level, including their lab tests and medication information and contains data on patients at the intensive care unit (ICU) level, including all clinical events occurring in that unit. In addition, it includes tables for the emergency department (ED) and an accompanying chest X-ray (CXR) dataset connected to the existing patients in the other tables of the dataset through unique anonymized identifiers. Because of the extensiveness and the granularity of this dataset, it has been used in countless academic papers attempting to derive patterns, train machine learning models, and analyze various biases that occur in patient care. Most of the data is tabular with numeric values in fields like lab results, and categorical fields in fields like medication type. A unique aspect of this dataset is the CXR component. The diagnosis labels for each CXR were provided to us in the dataset and were extracted using the open source CheXpert labeler. These labels serve as our ground truth values for assessing the performance of our model on the CXR dataset.

In this project we use the tabular data in conjunction with the associated CXRs and free text reports to train a multi-label classification model, and analyze possible bias introducing confounding variables using a variety of machine learning and statistical techniques

II. Background

CNNs are the current go-to architecture for building state-of-the-art computer vision models. CNNs have been used in a wide array of problems that deal with image data, including facial recognition [10], document analysis [6], self-driving cars [2], and image captioning [12]. Several CNN algorithms with different trade-offs and advantages have been rolled out, including VGG16, ResNet, and EfficientNet [8], [25], [27]. For our project, we are using DenseNet [9].

As a result of their success, researchers have combined the power of CNNs with available medical imaging technologies to explore numerous potential use cases in the healthcare sector. Such applications of CNNs range from the early detection of Osteoarthritis from MRI scans [15] to the classification of Brain Tumors [22], among many uses.

X-rays are the most widely used and available medical imaging technique [1], so leveraging ML to aid and expedite diagnosis and screening processes has been an active area of scholarly research. Specifically, chest X-rays receive a lot of attention, with some models showcasing impressive performances in diagnostic tasks [14], [20], [30].

Given that these models are now achieving strong performance that motivates their application in real-life medical contexts, the fairness of these ML algorithms and subsequent biases of the models have also become important areas of research, as decisions made in medical contexts usually have very significant implications for quality of care and general ethics [4], [29]. A model that scores highly on most performance metrics can still produce adverse outcomes for a subset of the population, especially in the case of minority and protected groups.

A recent landmark paper that aims to investigate biases in ML applications is *CheXclusion: Fairness gaps in deep chest X-ray classifiers* by Seyyed-Kalantari et al. [24]. The article uses three different popular public chest X-ray datasets (including MIMIC) to train and evaluate CNN models and stratifies model performance results according to categories of race, sex, age, and insurance provider. The metric used to measure bias is TPR disparity, and the experiment reveals there are performance disparities for every category and disease diagnosis. We follow the steps of this inquiry to use TPR as a performance disparity metric. In addition to the categories described above, we look at Intensive Care Unit (ICU) status as another category. ICU status is an important indicator of the severity of a patient's condition and the amount of medical attention they are receiving. Thus, it is useful to understand how these models perform differently for ICU and non-ICU patients.

The biological basis of socially constructed categoris such as race and the extent to which these categories manifest in patient X-rays are contentious topics. But recent research has shown that AI models can successfully infer the race of a patient based on medical imaging [3], [7]. Whether the underlying cause is biological or socioeconomic, current paradigms suggest that these different categories lead to disparities in model performance.

III. Methods

The broader algorithm we chose to explore for this project is a convolutional neural network (CNN). A CNN is a variation of the Artificial Neural Network (ANN) which uses linear transformations and matrix operations to extract sequential or spatial encodings from the input data [13]. These models have been shown, particularly in the field of computer vision (CV), to be particularly useful in various tasks. Since the dataset we chose is an imaging dataset, a CNN was the natural choice.

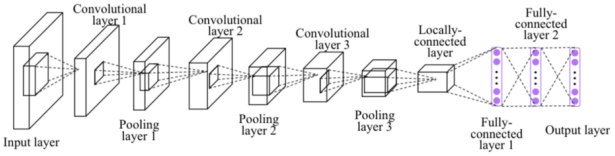

Fig. 1: CNN Basic Architecture

Figure 1 shows the basic architecture of a CNN. There are three main components of a CNN: convolutional layers, pooling layers, and fully connected dense layers. Convolution refers to the matrix operation element-wise multiplication. In a convolutional layer, the input and a mask of a specific, usually smaller, size are element-wise multiplied together, resulting in a set of features often smaller than the input. Since the mask is usually smaller than the image, a stride value must be specified which refers to the step size of the kernel when moving across the inputs. The goal of convolution is to extract important features from the image, and filter out unimportant features. Between each convolutional layer, there is also normally a pooling layer. Pooling is another way to reduce the dimensionality of the features. Using optimization strategies like taking the maximum or the average to extract the most important features, the dimensionality can be further reduced. Finally, after all convolutional and pooling layers are applied sequentially to the input, the extracted features are fed into an ANN or fully connected dense layers to make the actual prediction. There are many APIs available that aid in the construction of these models like Tensorflow, Keras, and Pytorch. These APIs make experimenting with different model architectures very straightforward. Since Yann LeCun proposed CNNs before the turn of the century, they have been a huge area of research with different types of model architectures benchmarked on open source data [16]. Picking an architecture for a CNN is much like model selection in traditional machine learning, however depending on the amount of data and the complexity of the model, CNNs can take anywhere from a few hours to many weeks to train so for the scope of this project, we pre-selected a base model called DenseNet121 [9].

A. DenseNet

DenseNet was proposed by researchers from Cornell, Tsinghua University, and Facebook AI research in 2016. This international group of academic and industry partners recognized an issue called the vanishing gradient problem. This occurs predominantly in deep CNNs during the back propagation phase. As the weights get updated, the gradients in towards the beginning of the model become arbitrarily small. Because of this, the weights do not get updated significantly and the model ceases to learn. DenseNet architecture combats this issue.

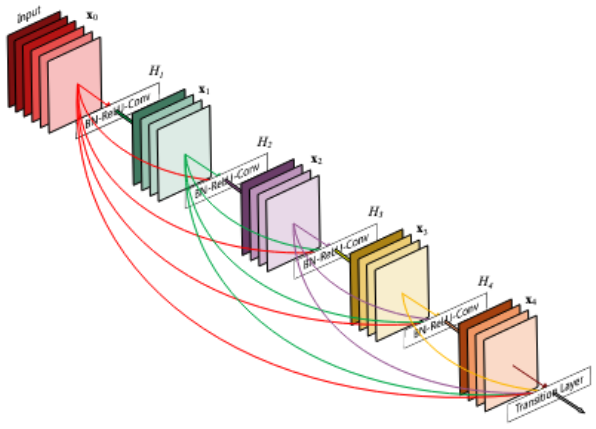

Fig. 2: Dense Block Architecture [9]

The intuition behind this architecture can be seen in Figure 2. This image, taken from the seminal paper on the DenseNet, depicts the architecture of a single dense block which makes up a DenseNet. The main thing that separates this architecture from that of a traditional CNN is that each convolutional layer is connected to all following layers. This allows features learned in previous layers to propagate through the network, whereas in a traditional CNN, those would be lost through repeated convolution. The authors showed that at the time, DenseNets outperformed similar state of the art models with fewer parameters.

B. Multi-label Classification

In class, we discussed the primary types of classification: binary and multi-class classification. In both of these types of classification, the vector of probabilities that the model outputs should sum to one. However, the healthcare domain is increasingly complex. It is possible, even likely, for patients to not only have one disease but many comorbidities. While it is possible to treat each label independently and train a binary classifier on each one, it would be a long and computationally expensive process. Because of this, we built a different kind of model: a multi-label classification model. A multi-label classification model is similar to a multi-class classification model. However, unlike multi-class classification, it treats each of the possible prediction values independently. Rather than discriminating between classes, a multi-label classification model outputs a probability of each of the possible labels in isolation.

From an implementation perspective, this was accomplished very simply using the Keras API. The convolution section of the model remained the same. The only thing that changed is the ANN portion of the CNN. Instead of having one output label at the end of the network with a sigmoid function and one neuron for binary classification, or n neurons (proportional to the number of classes) and a softmax activation function, this multi-label model had n output layers all concatenated to the convolutional section of the model, each with one neuron and a sigmoid activation function. Figure 3 depicts the differences between the three types of classifiers where each circle represents its own dense fully connected layer with one neuron. Another implementation consideration that was taken into account is that every classification layer needs its own loss function. Because of this, we leveraged different loss functions or class weights and applied them to each dense classification layer individually.

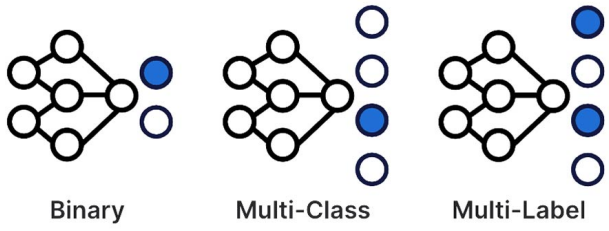

Fig. 3: Types of Classification [28]

C. Handling Class Imbalance

Another consideration when building this model was how to handle class imbalance. As seen in table I, in every case, the normal class greatly outnumbered the comorbidity class. This posed an issue when training and evaluating a model, as we wanted to ensure that the model learnt how to identify these comorbidities rather than just learning what is a "normal" chest X-ray.

We first trained a baseline model to see what the performance would look like if there were no strategies in place to handle the imbalance. For this, we used binary cross entropy(BCE), also called log loss, on each of the fully connected dense layers. After analyzing the performance of our baseline model, we attempted to use a new loss function that boasts better performance on highly imbalanced data: focal loss [14].

$$\mathrm{FL}(p_\mathrm{t}) = -\alpha_\mathrm{t}(1-p_\mathrm{t})^\gamma \log(p_\mathrm{t}).$$

Fig. 4: Focal Loss Equation

Figure 4 shows the equation for focal loss. Focal loss differs from regular BCE loss in two places. The first place is the

		Overall
n		243324
Atelectasis, n (%)	0.0	194534 (79.9)
	1.0	48790 (20.1)
Cardiomegaly, n (%)	0.0	195651 (80.4)
	1.0	47673 (19.6)
Consolidation, n (%)	0.0	231799 (95.3)
	1.0	11525 (4.7)
Edema, n (%)	0.0	213993 (87.9)
	1.0	29331 (12.1)
Enlarged Cardiomediastinum, n (%)	0.0	235667 (96.9)
	1.0	7657 (3.1)
Fracture, n (%)	0.0	238543 (98.0)
	1.0	4781 (2.0)
Lung Lesion, n (%)	0.0	236692 (97.3)
	1.0	6632 (2.7)
Lung Opacity, n (%)	0.0	188555 (77.5)
	1.0	54769 (22.5)
No Finding, n (%)	0.0	162207 (66.7)
	1.0	81117 (33.3)
Pleural Effusion, n (%)	0.0	185603 (76.3)
	1.0	57721 (23.7)
Pleural Other, n (%)	0.0	241241 (99.1)
	1.0	2083 (0.9)
Pneumonia, n (%)	0.0	226102 (92.9)
	1.0	17222 (7.1)
Pneumothorax, n (%)	0.0	232089 (95.4)
	1.0	11235 (4.6)
Support Devices, n (%)	0.0	170030 (69.9)
	1.0	73294 (30.1)

TABLE I: Label Distribution

alpha hyperparameter. Alpha in this case refers to the weight of the positive class. The other hyperparameter that is different in focal loss is gamma. The gamma parameter is referred to in the paper as a modulating term. This hyperparameter allows the loss function to penalize hard-to-classify samples more than easy-to-classify samples. In situations with large class imbalance, it is almost always the minority class that is the hardest to classify. Focal loss has been shown in many instances to improve the performance of models with high class imbalance in the training data. Because of this, we also trained a model using focal loss where gamma was equal to two, as that parameter performs the best empirically. We also tuned the class weight for each of the potential labels respective of the each instantiation of the focal loss object.

D. Learning Rate Schedulers

To improve the robustness of a deep learning model, it is often best practice to implement some type of learning rate scheduler to ensure the learning rate is not held constant throughout the model training phase. Since these models use a variation of stochastic gradient descent, the learning rate plays a pivotal role in finding the global minimum for the loss function. With a constant learning rate, it is likely that the algorithm would miss the global optimum and settle at a local optimum (large learning rate) or take a very long time to train if the learning rate was too small. To avoid this, there are many different types of commonly used learning rate schedulers. The two that were used in this project are called reduce learning rate on plateau scheduler and a cyclical learning rate scheduler [26]. As the name suggests, reduced learning rate on plateau multiplies the learning rate by a constant factor any time the validation loss does not improve for a user specified number of epochs. The assumption is that if the validation loss doesn't improve after that number of epochs, the learning rate is too large and could be skipping over the optimum. A cyclical learning rate scheduler tackles this issue differently by oscillating the learning rate between a user specified minimum and maximum learning rate according to a scaling function over a number of steps. The number of steps is normally set to be 2 to 10 times the number of steps in one epoch. The variation from smaller to larger learning rates avoids local optima by allowing the optimizer to exit a region it would have skipped over or gotten stuck in with a constant learning rate.

E. Experiments

A combination of all of the components discussed above were used in our experiments. In total we trained three models. Table II shows the the three experiments we ran and the different hyperparameters we used to try and handle class imbalance and improve the robustness of our model performance.

	Baseline Model	Model 1	Model 2
Architecture	Densenet121	Densenet121	Densenet121
Loss	Binary Cross Entropy	Focal	Focal
Batch Size	64	64	64
Optimizer	Adam	Adam	Adam
Initial Learning Rate	0.0001	0.0001	[0.00001, 0.00000001]
Learning Rate Scheduler	Reduce on plateau	Reduce on plateau	Cyclical
Early Stopping	Yes	Yes	Yes
Augmentation	No	Yes	Yes

TABLE II: Model Hyperparameters

The baseline model is used as a comparison point for the subsequent models. Because of this, the more state of the art techniques for model training were not used. In models 1 and 2, we incrementally added these techniques and observed the effect on model performance. Furthermore, each model was trained, validated, and evaluated on only frontal chest X-rays (CXRs)in order to avoid possible confusion by included frontal and lateral CXRs, as they are visually different.

IV. Results

A. Data Description

The model is evaluated on overall raw performance on the testing data as well as stratified performance across five relevant confounders. The model predictions are contained in a result dataset, where each row represents a given subject ID and study ID combination. The study ID represents an

CXR instance, and each subject can have multiple CXRs. The next 14 columns represent each of the 14 diagnoses labels produced by the CheXpert labeler(Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices) and constitute the ground truth values for this study. They contain values of either 1.0, 0.0 or -1.0. A value of 1.0 indicates the presence of the labeled comorbidity, 0.0 indicates no presence and -1.0 indicates that the model is unable to be confident of either result. Each record can have more than one comorbidity. To simplify our assessment process, we replace all instances label values of -1.0 to 0.0 so that we are working with binary values. The last 14 columns represent the model's probability of diagnosis for each label [0,1], indicating how confident the model is in the given prediction.

The result data for each of the three models is then merged with four other tables from the MIMIC-IV dataset: *admissions*, *patients*, *icustays* and *metadata*. This enables us to join clinical and demographic information for each subject-CXR combination, adding the following variables to the data: hospital admit time, hospital discharge time, admission type, admission location, discharge location, insurance, expiration flag, language, marital status, whether the image was taken in the ICU, race, age, date of birth (anonymized using a delta) and date of death. To narrow the scope of this project, we elect to focus on five confounders: race, age, sex, ICU status (binary), and insurance type. Thus, the final results data for each model includes 2 column identifiers for the subject and CXR, 14 columns representing the disease labels produced by CheXpert, 14 columns representing the label predictions produced by our model, and 5 columns for each of the confounders. Figures 5-9 represent the distributions of each confounder for each of the given labels in the testing dataset.

Figure 5 shows the racial distribution of the patients across the different labels. Racial information on patients is stored in the *admissions* table in MIMIC-IV. This dataset does not classify race in alignment with the US Census racial categorization schema and includes more specific information on ethnicity. We choose to bucket the races to loosely align with US social definitions of race and to mitigate underrepresentation of granularized minority groups. The racial categories used in this paper are: White (White American and European ethnicities), Asian (includes South Asian, East Asian, Southeast Asian and Pacific Islander ethnicities), Black/African American, Other (Unknown, Multiple Ethnicities, Native American/American Indian). Table III presents the racial breakdown of the testing data. The data is overwhelmingly white, but there doesn't appear to be any significant anomalies in the plots in Figure 5.

Figure 6 shows the distribution of age across the labels. Age is clustered into 8 age buckets ranging from 20 to 90+. The plots demonstrate the test data is overwhelmingly older, with the majority of patients being over the age of 50. Figure 7 shows the distribution of sex across the labels. The data is 45% female and 55% male. Lung Opacity is the only disease label where there is a higher proportion of males diagnosed than females; there appear to be no other anomalies in the data.

Figure 8 shows the distribution of insurance type across the labels. The four insurance types are Medicare, Medicaid, Other (private) and Unknown. Medicare is public insurance for people 65 and older. Medicaid is public insurance for people who are low income. Because the data is skewed towards older populations, 50% of patients are on Medicare. Another 35% of patients have some type of private insurance. This makes sense since those who go to hospitals tend to need robust insurance. This may also be why we see low distributions of Medicaid patients in our data (6%). There are also around 10% of patients for whom their insurance type is unknown. Figure 9 shows the distribution of ICU status across the labels. ICU status is encoded in a binary variable, where 1.0 indicates the CXR was taken while the subject was in the ICU and 0.0 indicates the CXR was taken while the subject was not in the ICU. There are more CXRs taken outside of the ICU. The plot also shows that all ICU patients in the data have a Support Device and almost all ICU patients have Pleural Effusion.

Race	Percent
White	2098 (61.65%)
Black	644 (18.92%)
Hispanic/Latino	97 (2.85%)
Asian	191 (5.61%)
Other	373 (10.96%)
Total	3403

TABLE III: Test Dataset Racial Breakdown

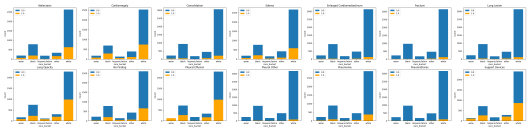

Fig. 5: Race Distribution Plots

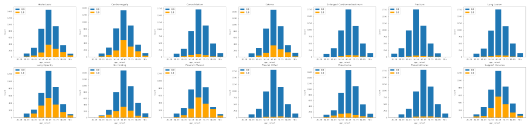

Fig. 6: Age Distribution Plos

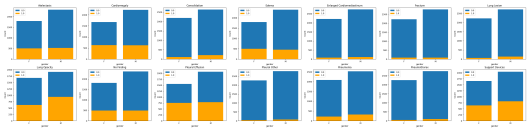

Fig. 7: Sex Distribution Plots

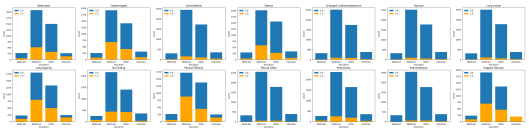

Fig. 8: Insurance Status Distribution Plots

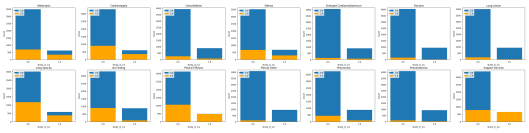

Fig. 9: ICU Distribution Plots

B. Overall Performance

The performance of the model is evaluated in three ways: 1) overall performance 2) AUPRC performance within each confounder 3) TPR disparities within each confounder. Overall performance is assessed by plotting the loss function, ROC and PRC curves for each of the three models and their respective strata. Figure 14 shows the training loss for the baseline model as a function of the number of epochs. The training loss converges for 12 out of 14 labels, but Pleural Other and No Finding do not improve. However, the validation loss spikes consistently across all labels, particularly around epoch 4. This is likely due to the fact that the learning rate scheduler used on this model decreases the learning rate at this point, resulting in the weights being updated in smaller increments during backpropagation and the loss subsequently increasing.

Figure 15 shows the training loss for Model 1. This model uses the same learning rate scheduler as the baseline model, but uses focal loss as the loss function to mitigate the class imbalance shown in the distribution plots. This means the hard-to-classify samples are penalized more, which is why the model does not converge for most labels. The validation loss curve is also not smooth, indicating the learning rate can still be improved. Figure 16 shows the training loss for Model 2. This model uses focal loss as the loss function but uses a cyclical learning rate scheduler. This yields the best performance out of the three models - the model converges on both the training and validation data for most labels. Pleural Other and Lung Lesion have slightly worse performance on the validation data, but it is significantly improved from the other two models. Thus, Model 2 is selected for the remainder of our analysis.

1) Metrics: The second aspect of evaluating overall model performance is choosing the appropriate assessment metric. Table IV demonstrates the AUROC and AUPRC scores for each label in Model 2. The AUROC scores are relatively high (¿ 0.70) for all labels except Enlarged Cardiomediastinum, Lung Lesion, Lung Opacity and Pneumonia. However, Table V demonstrates that there is significant class imbalance in the training data that disproportionately impacts some labels more than others. For this reason, AUPRC scores are computed, as they are more robust against class imbalance. Class imbalance also explains the model's poor AUROC scores on the aforementioned labels. For example, only 0.8% of subjects in the training data have Pleural Other, resulting in an artificially inflated AUROC score (0.86) and a poor AUPRC score (0.07). This pattern holds for both Lung Lesion and Enlarged Cardiomediastinum. Figures 13 and 14 visualize the ROC curves and PRC curves for each label. The ROC curves indicate strong performance across the labels, but the PRC curves reveal the model struggles on labels with high class imbalance. Thus the evaluation metric we choose is the AUPRC score.

Label	AUROC	AUPRC
Atelectasis	0.71472	0.38048
Cardiomegaly	0.74311	0.46406
Consolidation	0.72524	0.15331
Edema	0.81683	0.53226
Enlarged Cardiomediastinum	0.68907	0.08597
Fracture	0.72137	0.080859
Lung Lesion	0.68689	0.08396
Lung Opacity	0.66996	0.46870
No Finding	0.78063	0.46429
Pleural Effusion	0.85214	0.72078
Pleural Other	0.86017	0.07111
Pneumonia	0.66348	0.18681
Pneumothorax	0.76605	0.13195
Support Devices	0.83938	0.73343

TABLE IV: Model 2 Metric Comparison

Label	% Positive Class
Atelectasis	20
Cardiomegaly	19.5
Consolidation	4.7
Edema	11.9
Enlarged Cardiomediastinum	3.1
Fracture	2.0
Lung Lesion	2.7
Lung Opacity	22.4
No Finding	33.6
Pleural Effusion	23.6
Pleural Other	0.8
Pneumonia	7.0
Pneumothorax	4.6
Support Devices	30.0

TABLE V: Training Class Imbalance

2) Gradcams: In addition to the subgroup analysis that explored the model's performance, we also want to explore a more visual explanation of the model's performance. To do this, we generate gradcam mappings [23]. A gradcam attempts to display what physical feature on the image a model is using to make a decision. Since this is a multi-label classification model, every image generates one gradcam associated with that target label. Figure 19 shows the activation maps for the first seven labels. The heatmap shows which areas of the image activate the model the most during inference phase, eventually influencing the model's final decision.

C. Stratified Performance

1) Race: Another way we evaluate the model is by observing its performance within the strata of each confounder. The PRC curves stratified by race bucket are not smooth (see Figure 20). This is because the large number of strata (5) combined with the overrepresentation of white patients in the testing data yields small subclasses within each label, exponentially magnifying the existing class imbalance in certain labels. This means the model tends to not be confident in its predictions of *all* strata in these labels, demonstrated by the fact the confidence intervals of the AUPRC score for labels like Pneumonia, Lung Lesion or Enlarged Cardiomediastinum are extremely wide. Thus, most of the following analyses will focus on the other labels.

Alternatively, the sample size per label-race combination can be so small that all samples may be predominantly of one class, resulting in an extremely high AUPRC, such as in Pleural Effusion-Asian (0.814). Table **??** conveys this information more cleanly in a table. As expected, the model does perform well on white patients across labels without significant class imbalance. Among these labels, the AUPRC scores and respective confidence intervals of all three minority groups (black, Hispanic/latino, Asian) are similar to each other and are all above 0.2, indicating strong and consistent performance. The 'other' category, which includes patients of unknown races and Native Americans, has extremely high AUPRC scores, even in less-imbalanced labels like Pleural Effusion, but this again likely has to do with small sample size post-stratification. Clearly, for a testing dataset of this size (3,000 observations), it appears number of samples plays a large role in the model's performance on different races, but there does not appear to be any noticeable anomalies among these racial groups.

2) Sex: The model performance is also stratified by sex to highlight any potential biases. The PRC curves (see Figure 21 are significantly more interpretable, and there is only a marginal overrepresentation of male patients (55%) in the testing data. The model doesn't appear to be biased across all of the labels, but it is important to note that the model is significantly better at classifying comorbidities in males for No Finding, Pleural Effusion and Edema. This could be due to the slight overrepresentation of males, which more than likely is a possible source of bias for this model.

3) Age: Similar to the race stratification results, the age PRC curves with the original age buckets were highly impacted by the large number of strata, skew towards one age bucket (60-69) and subsequent smaller sample sizes. To ensure interpretability of our analysis, we consolidated the groups into two buckets: over 50 and under 50. The PRC curves of this stratification are shown in Figure 22. We can see that the PRC curve of the 50+ age bucket is much smoother and closely follows that of the overall PRC curve for each label, which makes sense given the skew in the data towards older populations. The confidence intervals for the over 50 group are narrower for all labels except those with massive class imbalance along these groupings (e.g. Fracture, Pneumothorax, Pleural Other), indicating that the model performs better on CXRs of older people. However, the model is significantly worse at classifying instances of No Finding in older populations (see Figure 22), which is likely because only 20% of the over 50 population have No Finding while 38% of the under 50 population have No Finding. Thus, the model does have better overall performance on older populations.

4) Insurance: The stratified analysis by Insurance type is largely inconclusive. The data is skewed towards Medicare, but the model does not consistently perform the best on Medicare patients compared to other patients, even in the labels with balanced classes. Similar to the other confounders, the PRC plots (see Figure 23) suffer from low sample sizes within strata, resulting in harsh curves for the Unknown and Medicaid insurance types for certain labels. This indicates that stratifying insurance in this way or in this particular use case may not be very telling of potential biases in model performance.

5) ICU Status: The model on average performs better on ICU CXRs (see Figure 24), confirming our initial hypothesis. There are a couple of notable exceptions to this, specifically the labels No Finding and Edema, where the model actually performs worse on ICU CXRs. This intuitively makes sense with regards to No Finding, since normally patients who are sent to the ICU have some sort of serious issue. The fact that the model performs worse on Edema may present an opportunity for further analysis. However, the ROC curves actually indicate that the model performs better on non-ICU patients, even on the labels with minimal class imbalance (see Figure 25). This may be due to the fact that there are more samples in the non-ICU strata, but regardless, highlights an area for future research.

D. TPR Disparity

Another way we measured and compared model performance across the different categories (race, sex, age, insurance, ICU status) is to look at the differences in True Positive Rate (TPR) for each disease label. A higher TPR would mean the model is doing better at detecting comorbidities for a given group, in which case TPR disparities across groups imply a bias in the model.

For our TPR calculations, we took the largest group in each category as the reference (TPR disparity = 1.00): white for race, male for sex, 60-69 for age, Medicare for insurance provider, 0.0 (non-ICU) for ICU status. Our plots depict the TPR disparities and sample size differences for each group. Figure 10 is an example of such plot. Bigger groups take larger space in the plot, and the color indicates the TPR disparity, with darker blue tones signaling lower relative TPR and brown tones signaling higher relative TPR.

1) Race: TPR disparities greater than 20% are observed in all disease classes except Cardiomegaly, Pneumonia, and Support Devices. Most drastic disparities are observed in groups with smaller sample size: Asian and Hispanic. Interestingly, while both of these groups are similar in size, Asian patients

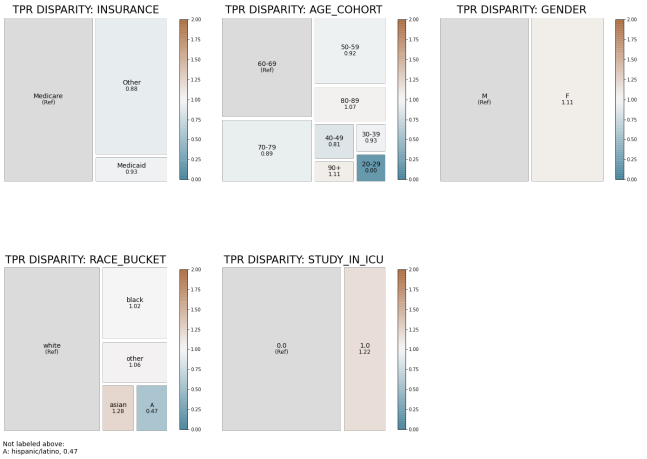

Fig. 10: Atelectasis TPR disparity across different categories

have a higher TPR than the reference for a majority of the diagnoses while Hispanic patients have a lower TPR. Figure 11 shows the case of Lung Opacity, where, per figure 26, both groups have a high incidence of the disease, but Hispanic patients are correctly diagnosed at a significantly lower rate.

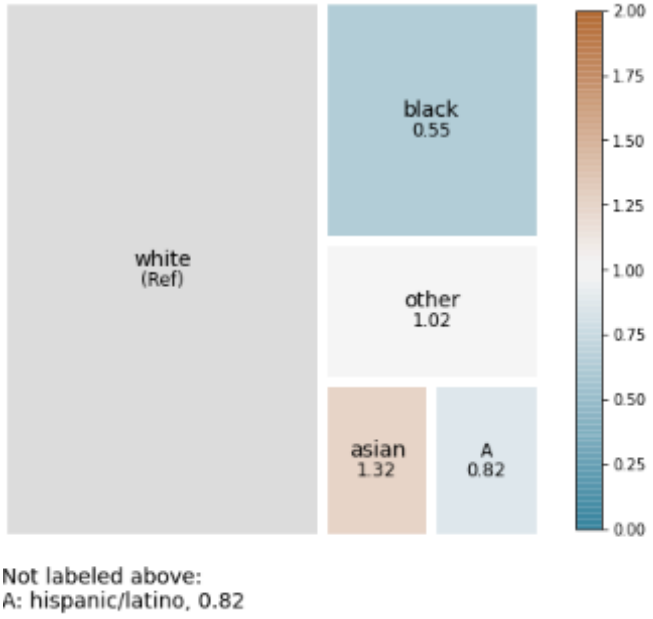

Fig. 11: Lung Opacity TPR Disparity across Race

2) Sex: TPR disparities greater than 20% are present in four disease classes: Fracture, No Finding, Pneumonia, and Pneumothorax. On all four of these except No Finding, male patients have a higher TPR.

3) Age: Disregarding the 20-29 age cohort which is not meaningfully represented in the dataset, all but four of the disease labels show TPR disparities greater than 20% across the original age buckets. Biggest variance in TPR is observed in some of the disease categories with the biggest class imbalance (Pneumothorax, Fracture, Lung Lesion) and in age groups with smaller sample size (30-39, 40-49, 90+). For the label No Finding, the model performs better for younger patients (30-39, 40-49, 50-59) and worse for older patients (70-79, 80-89, 90+) compared to the majority group (60-69), as seen on Figure 12. This overlaps with the finding from Section C and is probably due to the higher incidence of No Finding in younger patients.

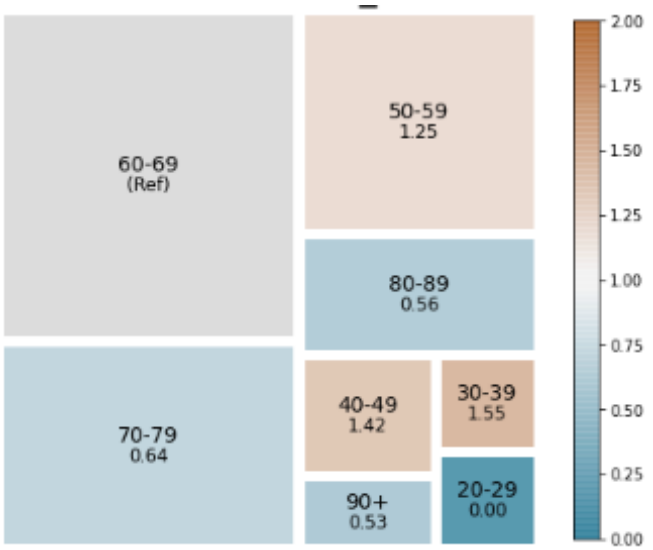

Fig. 12: No Finding TPR Disparity across Age

4) ICU Status: TPR disparities greater than 20% are observed in seven disease labels. In four of these disease classes, the model performs better for ICU patients, even though they constitute a smaller sample. Similar to the findings of section C, one of the labels the model performs significantly worse for ICU patients is No Finding, for which the model scored 88% worse than the TPR of non-ICU patients.

5) Insurance Type: When stratifying by Insurance, 6 out of 14 disease classes have TPR disparities greater than 20%. In three of these cases, Medicaid patients have the lowest TPR: Pneumothorax, Pneumonia, Fracture. The only instance where the model performs the worst with the majority group, Medicare patients, was with the label No Finding, as seen on figure 13. As older patients tend to have Medicare, this overlaps with our findings on age.

V. Discussion

This paper analyzes relevant confounders that may affect CNN performance on MIMIC-CXR data. Our goal is to build upon previous applications of deep learning models in this domain but with a specific focus on identifying the potential impact of confounders on model performance. We take two different approaches for our analysis. The first approach is stratifying the overall model performance metrics (AUPRC) by each of the confounders, and the second approach is assessing the TPR disparities within a given confounder for each label.

In general, we find that the TPR disparity results corroborate the stratified PRC results. The PRC analysis indicates that

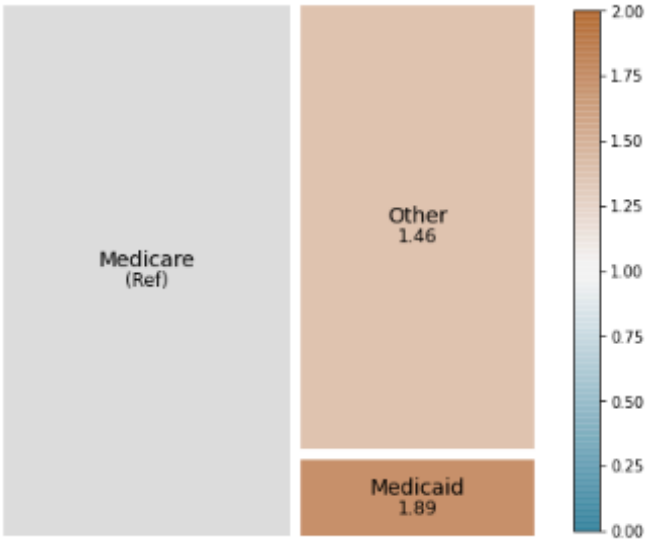

Fig. 13: Lung Opacity TPR Disparity across Race

the model performs well on white patients, which aligns with the bias in the data. Minority group performance is relatively consistent across the three minority groups (black, white, Asian), with anticipated fluctuations in the 'other' category. TPR disparities greater than 20% are observed in all labels, and Asian patients have a higher TPR than Hispanic patients. This is likely due to the fact that in the US, Hispanic/Latino is an ethnicity, not a race, meaning it is not necessarily referent to a specific phenotype or gene pool. Future studies should occur in conversation with existing literature on race and biology to perhaps determine more salient categories for analyzing racial bias.

Furthermore, both the PRC and TPR analyses on sex demonstrate that the model may be better at classifying co-morbidities in males compared to females. The PRC analyses indicate that the model performs better on males in No Finding, Pleural Effusion and Edema. The TPR analyses suggest that males have a higher TPR on Fracture, Pneumonia and Pneumothorax; however, these labels have significant class imbalance, so the TPR is likely inflated. The training data does contain more male CXRs (55%), but the disparity in sample size is not extremely significant. This implies that there is some residual bias against females in the model that should be explored further.

The PRC analysis using the consolidated age buckets (over 50 and under 50) indicates that the model performs better on older populations, which substantiates the skew observed in the training data. The TPR analysis on the original buckets yields inconclusive results, likely due to the large number of strata. This issue also renders the PRC analysis for Insurance Type largely inconclusive, although the TPR analysis indicates that there is a TPR disparity of greater than 20% for around half of the labels. However, because the Medicaid and Unknown strata are so small, these results are not very compelling.

Lastly, the PRC analysis suggests that the model performs better on CXRs taken in the ICU compared to CXRs taken outside of the ICU. However, the ROC curves imply the opposite, indicating that the model actually performs better on non-ICU CXRs. These ROC results are consistent across all labels, including the ones with minimal class imbalance. However, the TPR disparity analysis indicates that the model performs better on ICU patients, with the exception of the No Finding label. Thus, the combination of these three analyses give the impression that the model is in fact better at predicting ICU CXRs, but future clinician-backed research is needed to uncover the PRC/ROC discrepancy.

Thus, our model performs reasonably well, but the stratified analysis suggests that model performs better on males, older populations and ICU patients. We believe this is primarily due to the imbalanced distributions of these clinical and demographic features in the MIMIC-IV dataset. There are a few important limitations to our study that encourage future work. First, the size of the testing dataset needs to be expanded to stratify at this level of granularity. This will mitigate against extremely small sub-sample sizes that magnify existing class imbalances. To accomplish this, future research might consider externally validating our model against data from another open source dataset or institution, or enrich our training data with those datasets to improve our model. Additionally, the dataset that is used is over 88% white and heavily biased towards older and male populations. Future work should use a more diverse dataset, particularly in terms of race and sex in order to fully evaluate the potential biases that may occur when deploying a model like this in a clinical setting. Hospital data is inherently going to be biased against older patients, which presents an interesting challenge on how best to train an unbiased model with limited representation of young people and children, especially given data privacy and HIPPA regulations around working with childrens' data. Ultimately, our model is simply a starting point for engaging in a much deeper conversation on bias in AI and healthcare.

CONTRIBUTIONS

Person	Task
Niki	model evaluation, stratified analysis (race, sex), repo organization, results section, abstract, slide deck construction
Mert	background section, TPR disparity analysis stratified analysis (age, insurance) association analysis (Pearson, Chi Squared)
Zach	model training, gradcams, stratified analysis (ICU), introduction and methods section, code modularization and formatting

REFERENCES

[1] X-rays, ct scans and mris. https://orthoinfo.aaos.org/en/treatment/x-rays-ct-scans-and-mris, 2017.

[2] Akhil Agnihotri, Prathamesh Saraf, and Kriti Rajesh Bapnad. A convolutional neural network approach towards self-driving cars. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4, 2019.

[3] Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis Pyrros, Luke Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang, and Judy W. Gichoya. Reading race: AI recognises patient's racial identity in medical images. *CoRR*, abs/2107.10356, 2021.

[4] Danton S. Char, Nigam H. Shah, and David Magnus. Implementing machine learning in health care — addressing ethical challenges. *New England Journal of Medicine*, 378(11):981–983, 2018. PMID: 29539284.

[5] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[6] Arpita Dutta, Arpan Garai, Samit Biswas, and Amit Das. Segmentation of text lines using multi-scale cnn from warped printed and handwritten document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 24:1–15, 12 2021.

[7] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 2022.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[10] Rondik J.Hassan and Adnan Mohsin Abdulazeez. Deep Learning Convolutional Neural Network for Face Recognition: A Review. *International Journal of Science and Business*, 5(2):114–127, 2021.

[11] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV, 2022.

[12] Saloni Kalra and Alka Leekha. Survey of convolutional neural networks for image captioning. *Journal of Information and Optimization Sciences*, 41(1):239–260, 2020.

[13] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, Winter 1989.

[14] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[15] Rabbia Mahum, Saeed Ur Rehman, Talha Meraj, Hafiz Tayyab Rauf, Aun Irtaza , Ahmed M. El-Sherbeeny, and Mohammed A. El-Meligy . A novel hybrid approach based on deep cnn features to detect knee osteoarthritis. *Sensors*, 21(18), 2021.

[16] Mustafa Alghali Elsaid Muhammed, Ahmed Abdalazeem Ahmed, and Tarig Ahmed Khalid. Benchmark analysis of popular imagenet classification deep cnn architectures. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 902–907. IEEE, 2017.

[17] Isaac Kofi Nti, Adebayo Felix Adekoya, Benjamin Asubam Weyori, and Owusu Nyarko-Boateng. Applications of artificial intelligence in engineering and manufacturing: a systematic review. *Journal of Intelligent Manufacturing*, 33(6):1581–1601, 2022.

[18] Madhumita Pal, Smita Parija, Ranjan K Mohapatra, Snehasish Mishra, Ali A Rabaan, Abbas Al Mutair, Saad Alhumaid, Jaffar A Al-Tawfiq, and Kuldeep Dhama. Symptom-based COVID-19 prognosis through AI-based IoT: A bioinformatics approach. *Biomed Res. Int.*, 2022:3113119, July 2022.

[19] Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141, 2020.

[20] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis P. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017.

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[22] J Seetha and S Selvakumar Raja. Brain tumor classification using convolutional neural networks. *Biomedical & Pharmacology Journal*, 11(3):1457, 2018.

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019.

[24] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[26] Leslie N. Smith. No more pesky learning rate guessing games. *CoRR*, abs/1506.01186, 2015.

[27] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019.

[28] Kanyes Thaker. Data-centric approaches to multi-label classification, Jul 2022.

[29] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, 15(11):1–4, 11 2018.

[30] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315, 2017.

[31] J Zhang, S Whebell, J Gallifant, S Budhdeo, H Mattie, P Lertvittayakumjorn, M P Arias Lopez, B J Tiangco, J W Gichoya, H Ashrafian, L A Celi, and J T Teo. An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. *medRxiv*, 2021.

Label	PRAUC				
	Asian	Black	Hispanic/latino	Other	White
Atelectasis	0.34 [0.256, 0.496]	0.36 [0.3, 0.444]	0.36 [0.012, 1.0]	0.52 [0.442, 0.612]	0.36 [0.325, 0.404]
Cardiomegaly	0.36 [0.277, 0.491]	0.48 [0.42, 0.556]	0.39 [0.264, 0.551]	0.48 [0.38, 0.59]	0.47 [0.433, 0.52]
Consolidation	0.14 [0.085, 0.278]	0.18 [0.075, 0.317]	N/A	0.24 [0.159, 0.365]	0.15 [0.112, 0.204]
Edema	0.58 [0.439, 0.715]	0.59 [0.496, 0.67]	0.47 [0.285, 0.702]	0.37 [0.273, 0.498]	0.53 [0.486, 0.578]
Enlarged Cardiomediastinum	0.06 [0.03, 0.124]	0.11 [0.03, 0.257]	N/A	0.14 [0.066, 0.31]	0.09 [0.069, 0.133]
Fracture	0.1 [0.029, 0.275]	0.01 [0.002, 0.038]	0.1 [0.011, 0.5]	0.2 [0.105, 0.38]	0.08 [0.048, 0.166]
Lung Lesion	N/A	0.13 [0.062, 0.294]	0.25 [0.032, 0.591]	0.14 [0.062, 0.311]	0.08 [0.052, 0.121]
Lung Opacity	0.5 [0.39, 0.625]	0.39 [0.331, 0.471]	0.57 [0.435, 0.739]	0.53 [0.453, 0.621]	0.47 [0.434, 0.511]
No Finding	0.22 [0.122, 0.417]	0.53 [0.442, 0.629]	0.5 [0.301, 0.715]	0.31 [0.187, 0.462]	0.48 [0.431, 0.532]
Pleural Effusion	0.81 [0.726, 0.895]	0.61 [0.537, 0.692]	0.36 [0.042, 1.0]	0.76 [0.683, 0.824]	0.73 [0.695, 0.762]
Pleural Other	N/A	0.15 [0.084, 0.307]	0.3 [0.077, 0.794]	0.16 [0.016, 0.698]	0.06 [0.039, 0.088]
Pneumonia	0.16 [0.095, 0.297]	0.21 [0.138, 0.326]	0.42 [0.138, 0.671]	0.14 [0.092, 0.235]	0.2 [0.165, 0.254]
Pneumothorax	0.16 [0.065, 0.356]	0.12 [0.047, 0.277]	0.03 [0.011, 0.087]	0.36 [0.136, 0.596]	0.09 [0.061, 0.135]
Support Devices	0.76 [0.673, 0.839]	0.67 [0.598, 0.732]	0.57 [0.217, 0.868]	0.8 [0.742, 0.864]	0.73 [0.695, 0.762]

TABLE VI: Model 2 PRAUC

Label	ROCAUC				
	Asian	Black	Hispanic/latino	Other	White
Atelectasis	0.69 [0.602, 0.77]	0.72 [0.672, 0.763]	0.54 [0.146, 1.0]	0.65 [0.597, 0.711]	0.71 [0.689, 0.736]
Cardiomegaly	0.64 [0.554, 0.717]	0.72 [0.676, 0.76]	0.61 [0.48, 0.717]	0.69 [0.628, 0.756]	0.77 [0.747, 0.787]
Consolidation	0.62 [0.478, 0.734]	0.72 [0.617, 0.808]	N/A	0.8 [0.723, 0.859]	0.7 [0.662, 0.745]
Edema	0.77 [0.693, 0.84]	0.83 [0.786, 0.859]	0.72 [0.587, 0.85]	0.76 [0.699, 0.826]	0.83 [0.807, 0.845]
Enlarged Cardiomediastinum	0.53 [0.38, 0.69]	0.66 [0.511, 0.788]	N/A	0.71 [0.585, 0.829]	0.7 [0.648, 0.749]
Fracture	0.79 [0.603, 0.942]	0.66 [0.303, 0.879]	0.55 [0.104, 0.979]	0.78 [0.659, 0.879]	0.7 [0.624, 0.768]
Lung Lesion	N/A	0.81 [0.703, 0.898]	0.62 [0.355, 0.882]	0.63 [0.48, 0.768]	0.67 [0.603, 0.724]
Lung Opacity	0.66 [0.57, 0.735]	0.66 [0.608, 0.697]	0.7 [0.591, 0.799]	0.63 [0.577, 0.693]	0.68 [0.655, 0.703]
No Finding	0.68 [0.548, 0.809]	0.81 [0.76, 0.853]	0.74 [0.6, 0.856]	0.72 [0.625, 0.805]	0.78 [0.759, 0.806]
Pleural Effusion	0.83 [0.772, 0.887]	0.83 [0.794, 0.857]	0.84 [0.635, 1.0]	0.84 [0.796, 0.878]	0.86 [0.84, 0.872]
Pleural Other	N/A	0.93 [0.903, 0.956]	0.9 [0.747, 0.99]	0.86 [0.685, 0.997]	0.82 [0.783, 0.862]
Pneumonia	0.7 [0.587, 0.809]	0.66 [0.585, 0.741]	0.81 [0.666, 0.931]	0.6 [0.485, 0.713]	0.67 [0.631, 0.704]
Pneumothorax	0.65 [0.462, 0.823]	0.91 [0.846, 0.96]	0.38 [0.083, 0.702]	0.66 [0.486, 0.83]	0.77 [0.716, 0.822]
Support Devices	0.74 [0.663, 0.806]	0.8 [0.765, 0.835]	0.94 [0.874, 0.984]	0.83 [0.788, 0.87]	0.85 [0.828, 0.862]

TABLE VII: Model 2 ROCAUC

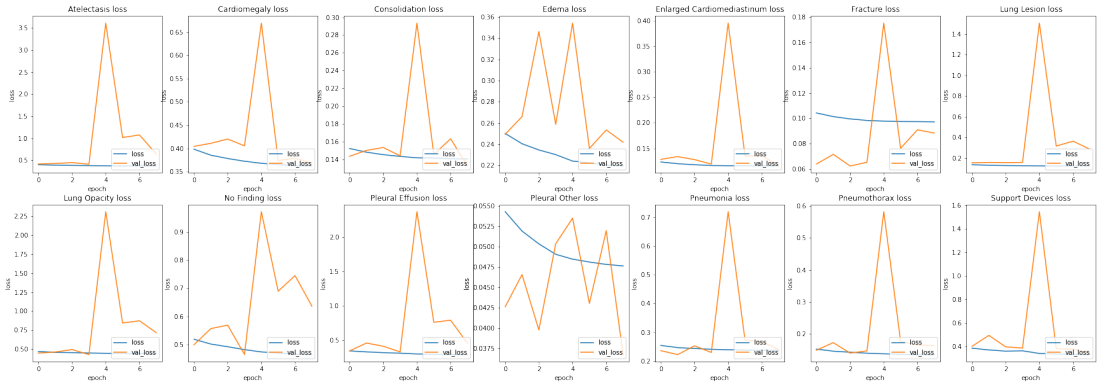

Fig. 14: Baseline Loss Plot

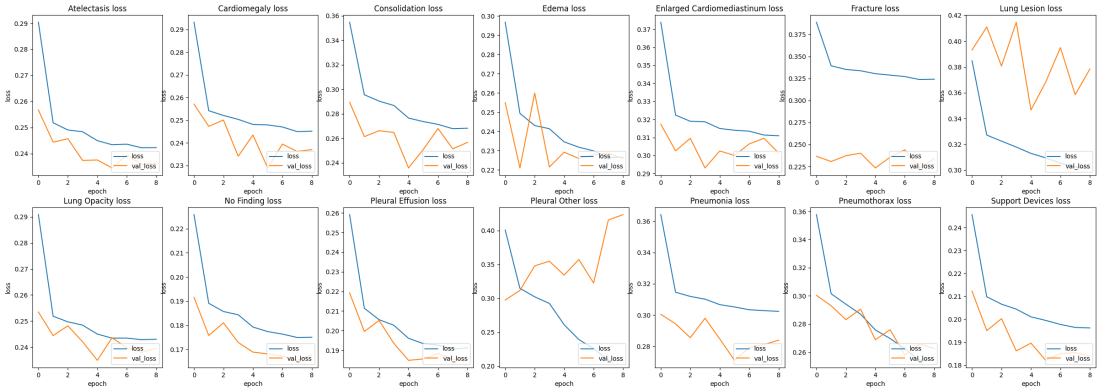

Fig. 15: Model 1 Loss Plot

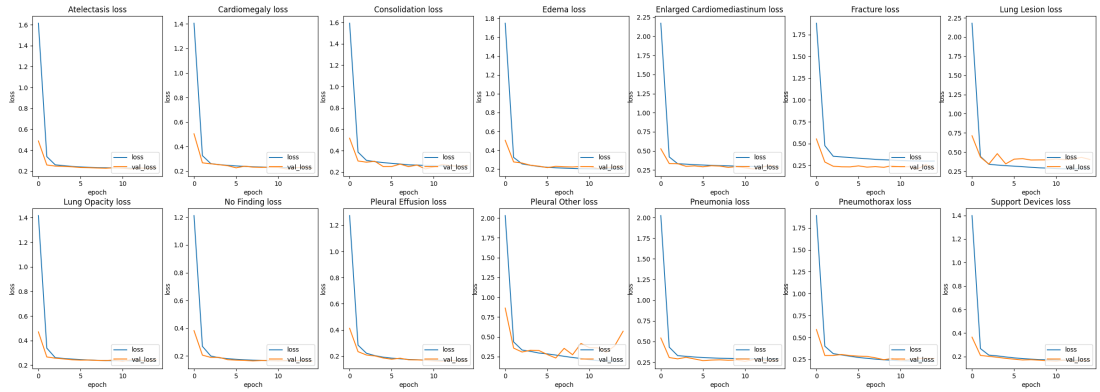

Fig. 16: Model 2 Loss Plot

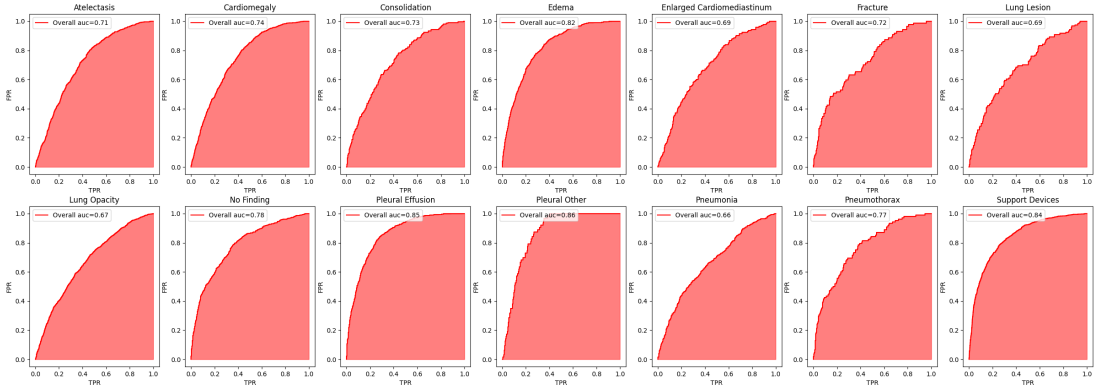

Fig. 17: Model 2 ROC Plots

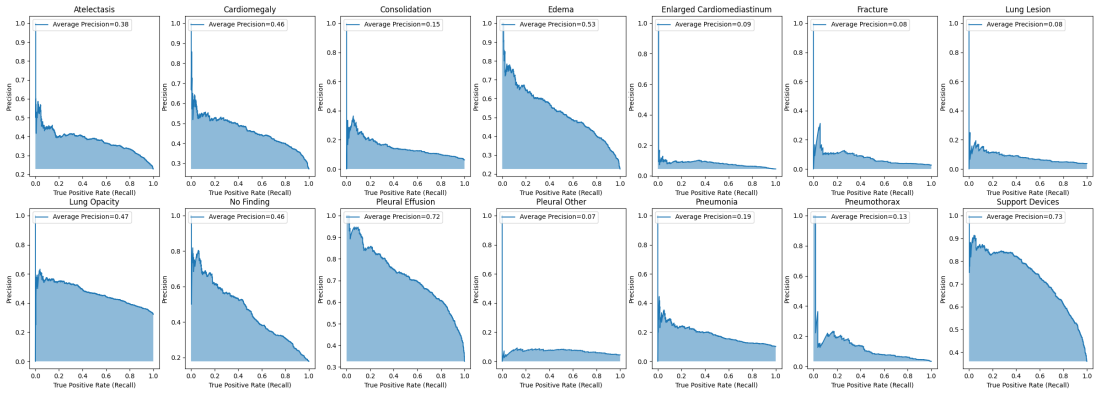

Fig. 18: Model 2 PRC Plots

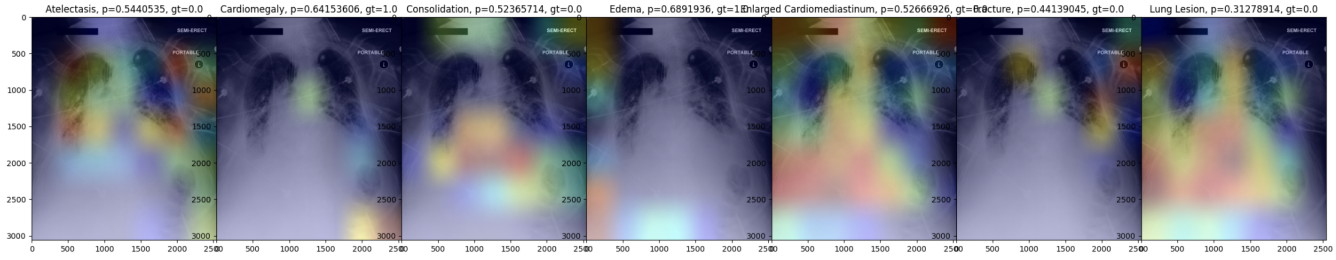

Fig. 19: Example Gradcam

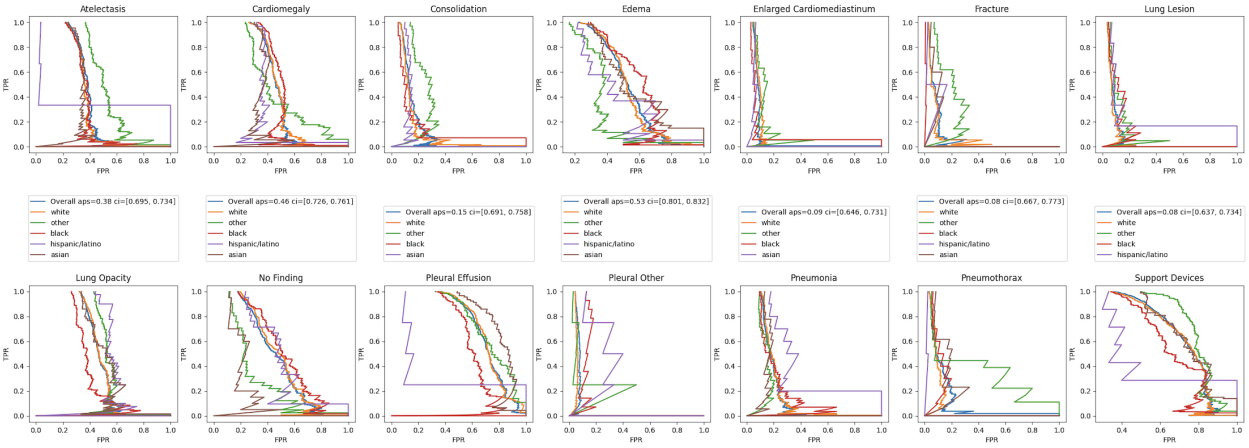

Fig. 20: PRC Plots by Race

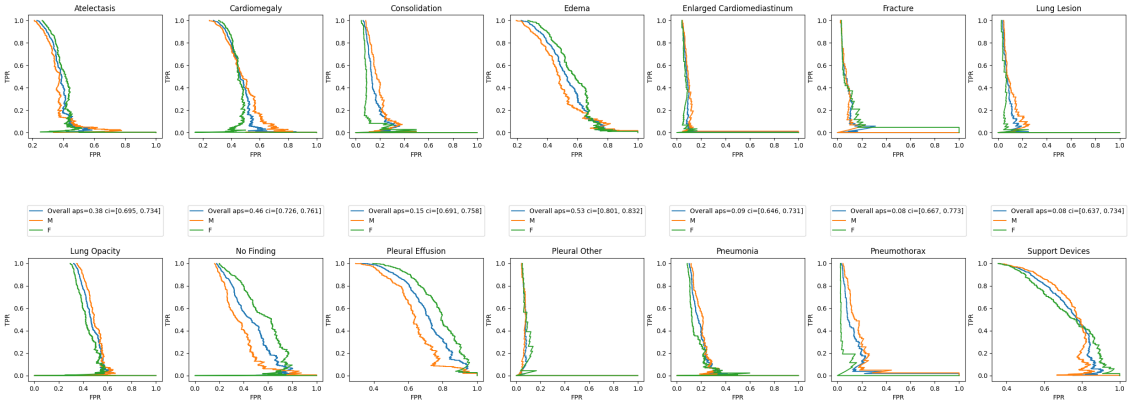

Fig. 21: PRC Plots by Sex

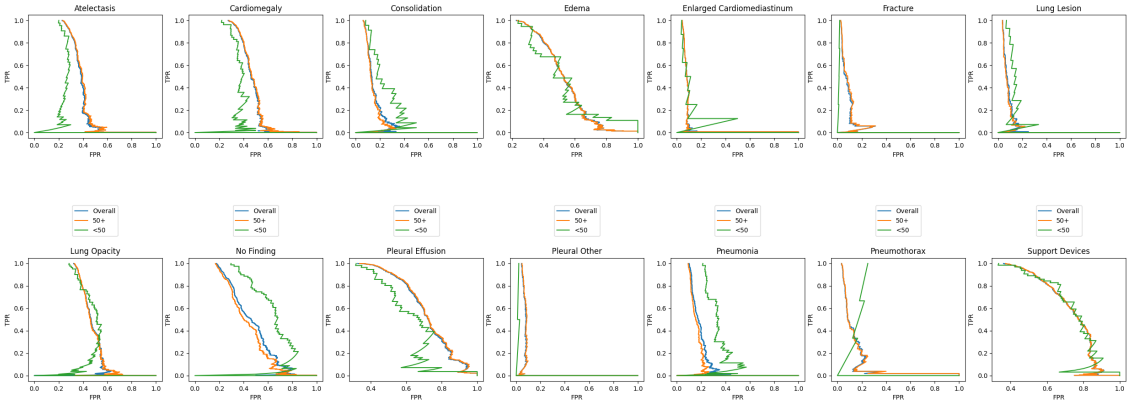

Fig. 22: PRC Plots by Age

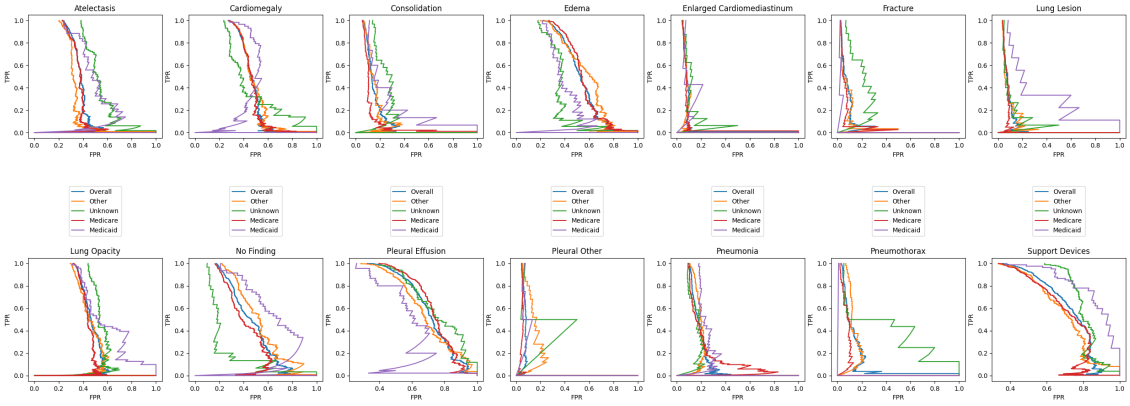

Fig. 23: PRC Plots by Insurance Type

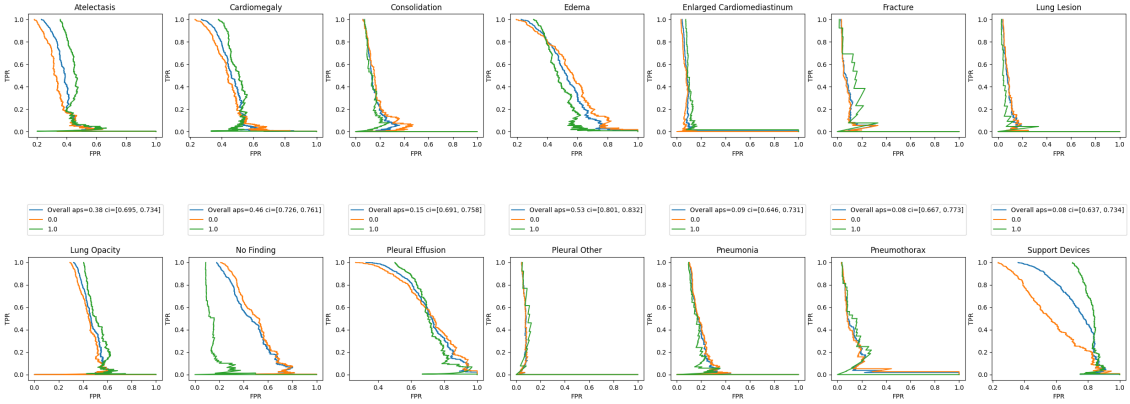

Fig. 24: PRC Plots by ICU Status

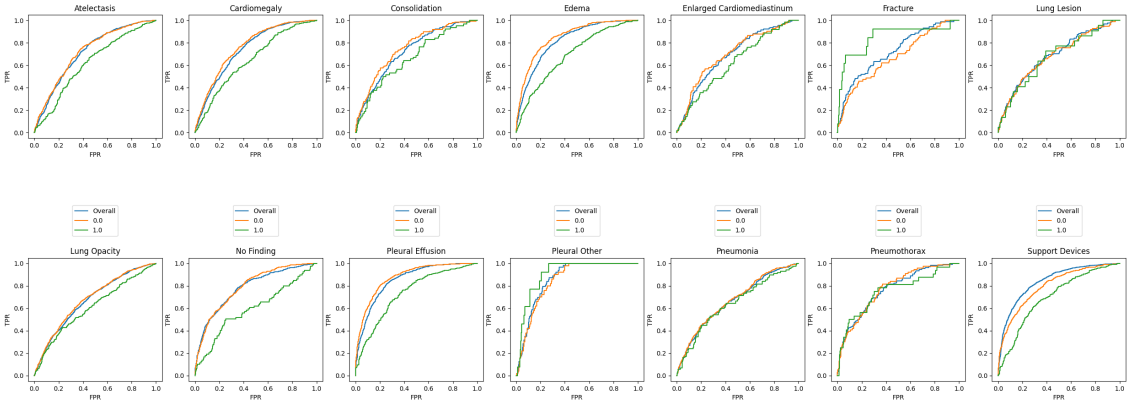

Fig. 25: ROC Plots by ICU Status

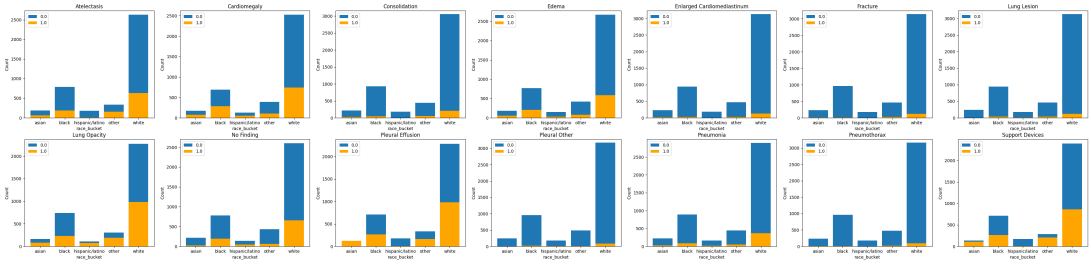

Fig. 26: A.1: Race Distribution Plots

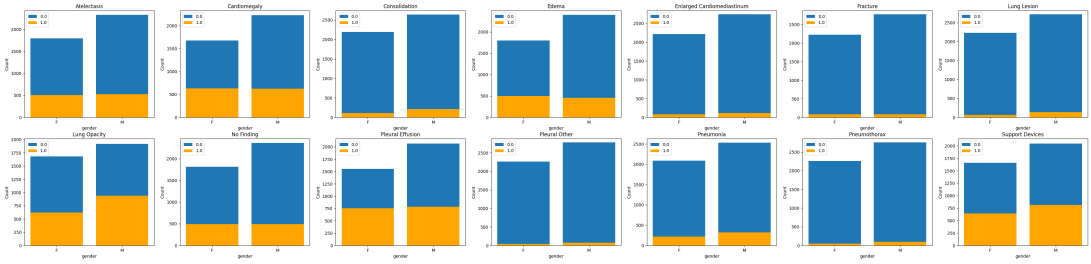

Fig. 27: A.2: Sex Distribution Plots

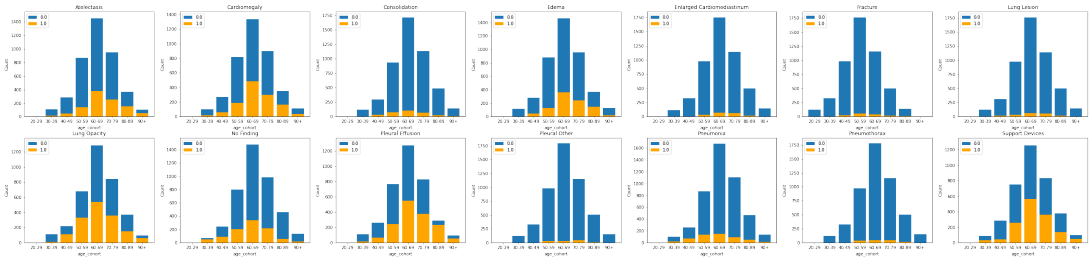

Fig. 28: A.3: Age Distribution Plots

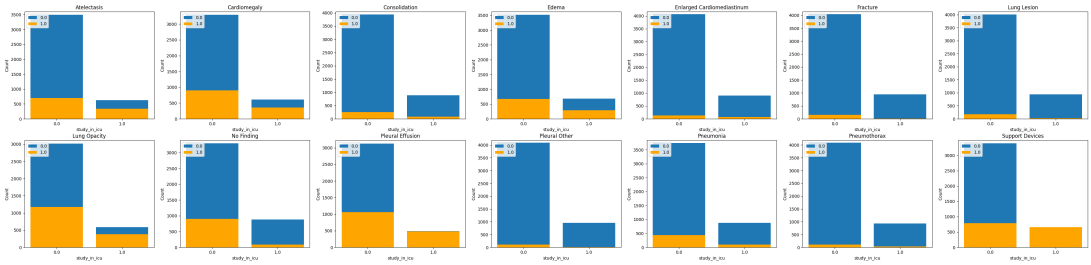

Fig. 29: A.4: ICU Distribution Plots

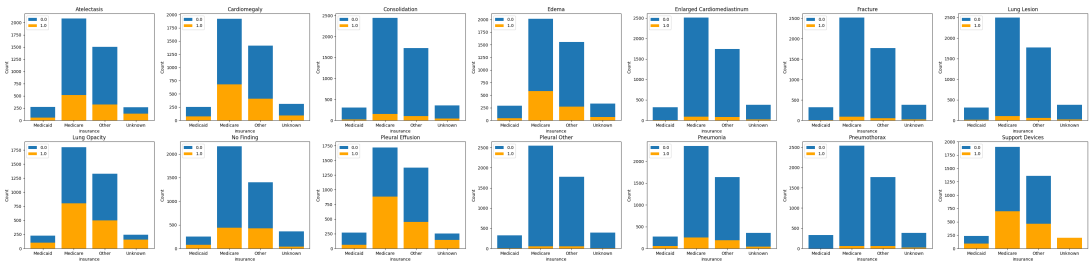

Fig. 30: A.5: Insurance Provider Distribution Plots

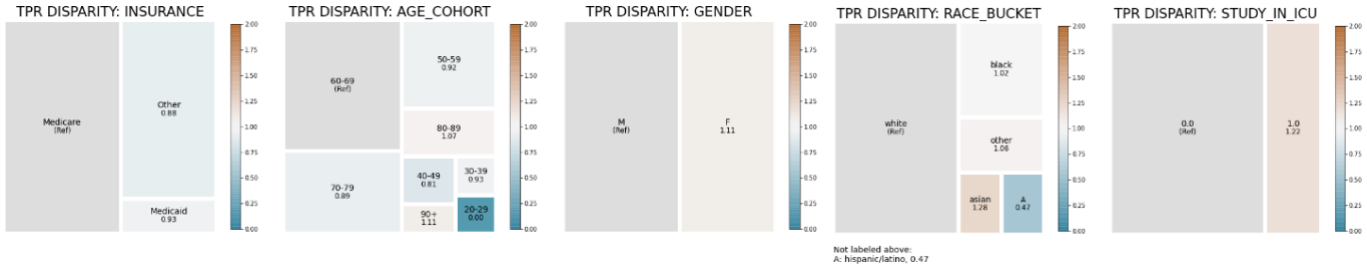

Fig. 31: A.6: Atelectasis TPR Disparity

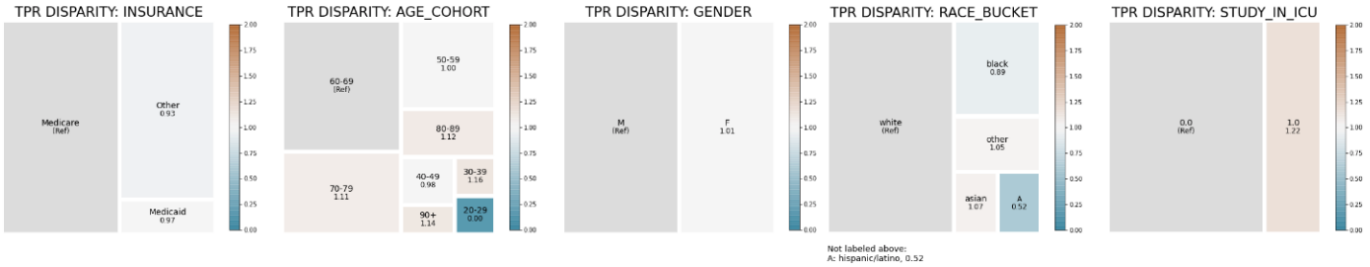

Fig. 32: A.7: Lung Opacity TPR Disparity

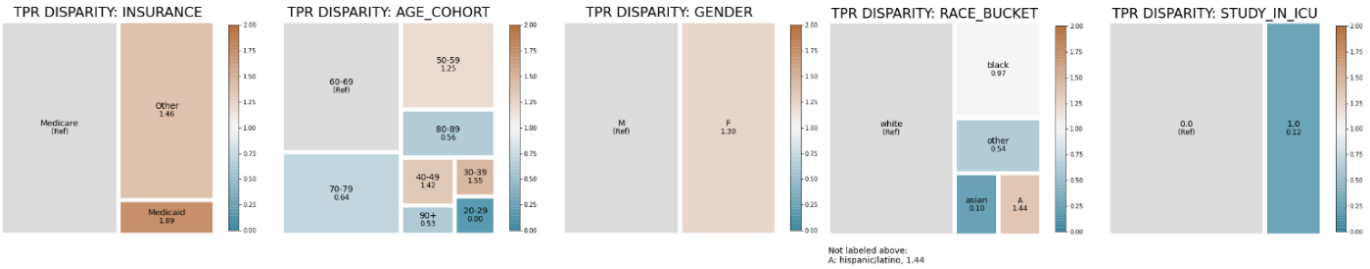

Fig. 33: A.8: No Finding TPR Disparity