Long-Tail analysis in Medical Dataset and Synthetic data generation

TEAM #15
HEEGYU EOM
JANGWON SON
JABIN KOO

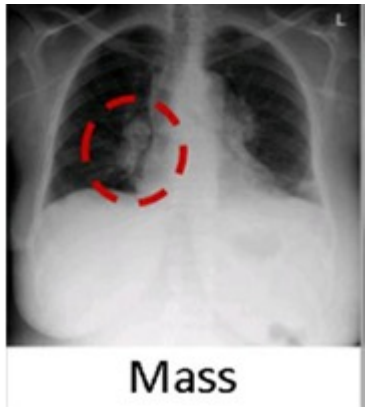
Medical Data Can be Long-Tailed

Medical imaging is one of the most activate and high-stakes applications of AI.

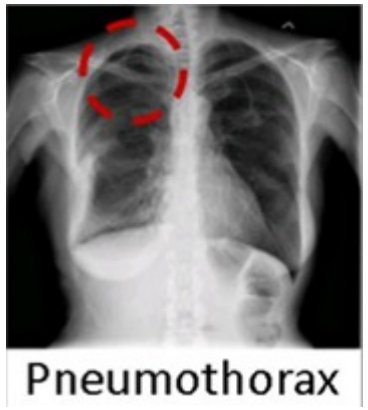
But diseases occur at different rates, so medical data is naturally **long-tailed**.

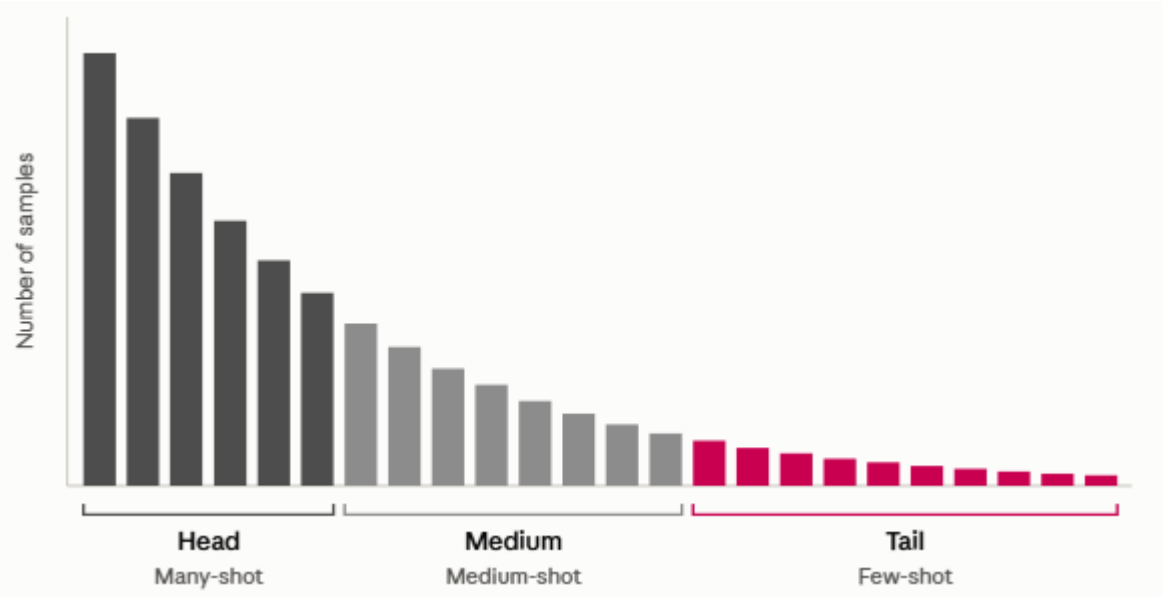
Models traind on such data often fail on rare diseases, where mistakes matter most.

Head class



Tail class

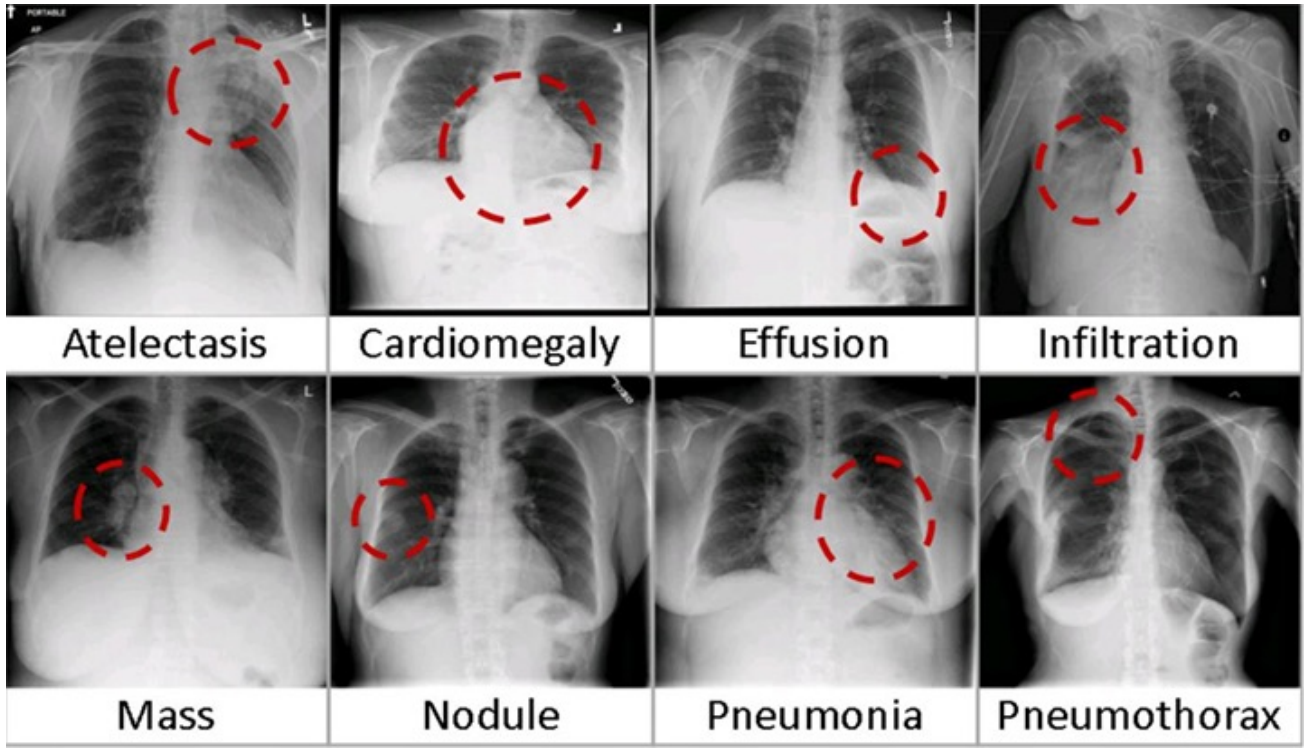




Project Goal: Solving Long-tail Problem in Medical Dataset

Our goal is **identifying** and **solving** **long-tail problems** in NIH-CXR-LT dataset.

We will **analyze** the dataset and use **synthetic data generation** and **data augmentation** methods to solve the problem.



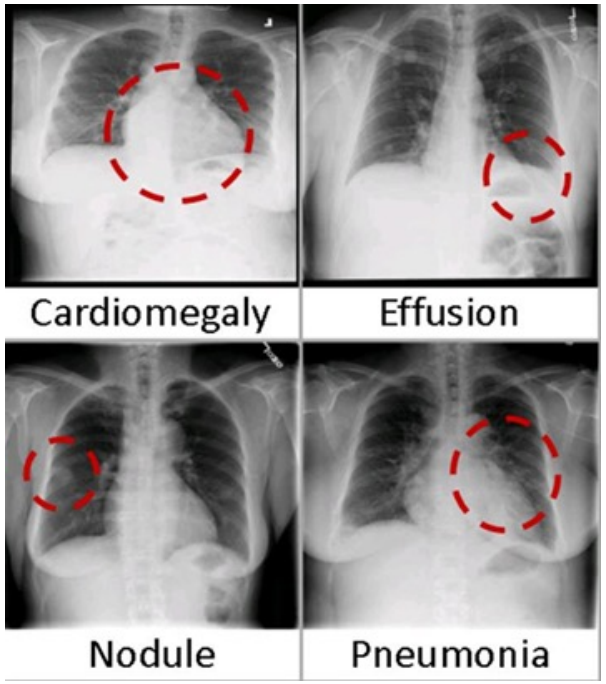
NIH-CXR-LT Dataset

NIH-CXR-LT is a large **public chest X-ray dataset** released by the NIH Clinical Center.

It is comprised of with 88,637 samples from 20 classes and includes information such as sex, age and view

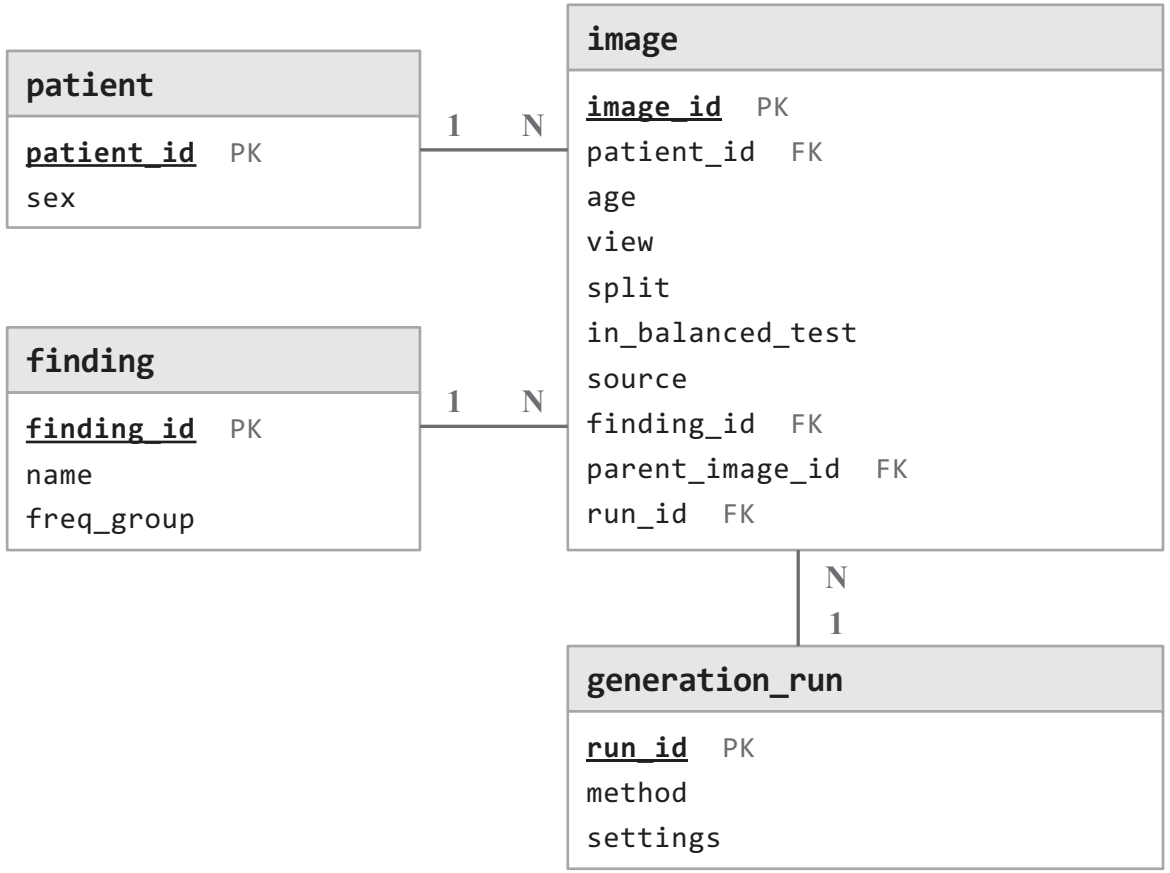
Split	Images	Patients	Images per class
Train	68,058	23,379	7 to 45,439
Balanced validation	300	112	15
Test	20,279	5,969	33 to 13,055
Balanced test	600	124	30

Table 1. NIH-CXR-LT splits



Data Pipeline and Database Design

Labels and splits come from the NIH-CXR-LT files, and **class, age, sex, view** and **patient ID** come from the ChestX-ray14 metadata. We join the two on the image ID and keep one row per image.



Finding the tail classes in the training split

```
SELECT f.name, COUNT(*) AS n_train
FROM image i
JOIN finding f ON f.finding_id = i.finding_id
WHERE i.split = 'train' AND i.source = 'real'
GROUP BY f.name
HAVING COUNT(*) < 100
ORDER BY n_train DESC;
```

name	n_train
Pneumoperitoneum	59
Hernia	49
Subcutaneous Emphysema	32
Pneumomediastinum	7

Class Imbalance in NIH-CXR dataset

Our analysis showed that there were **class imbalance**. Rare conditions that matter clinically has the fewest images.

In our training set, 7 head classes hold **93.4%** of the images, while 4 tail classes hold only **0.2%**.

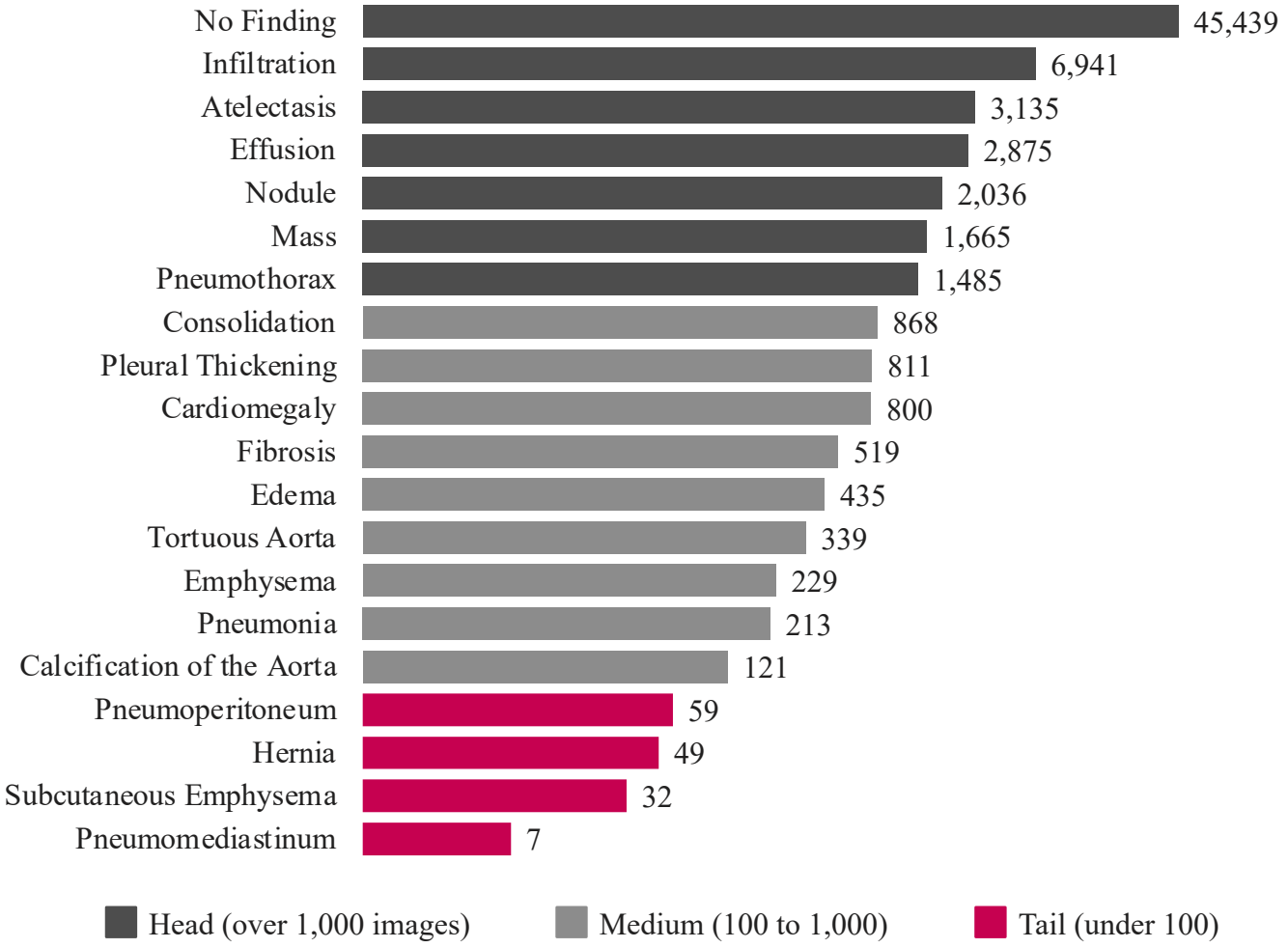
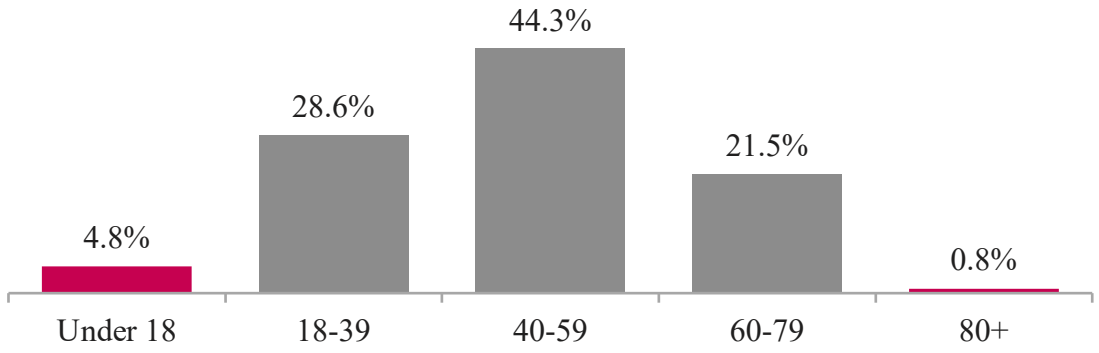


Figure 1. Training images per class in NIH-CXR-LT (log scale)

Skew Beyond Disease Class

Training images are also **uneven** by **age, view position** and **images per patient**. Sex is closer to balanced.

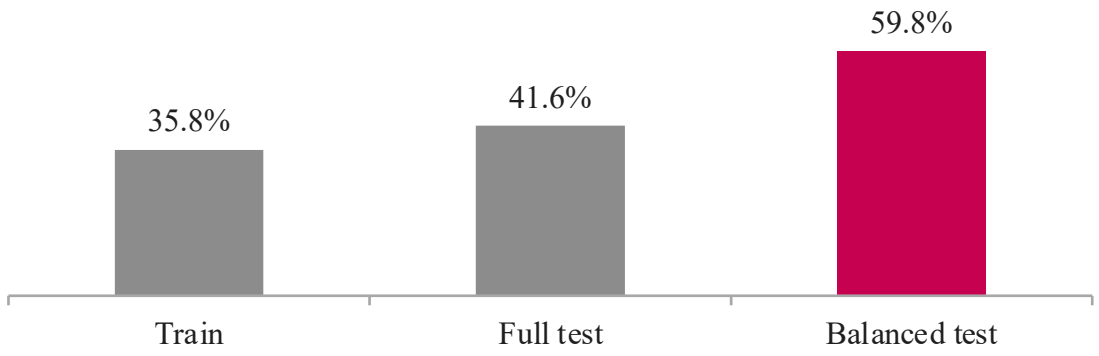
Patients under 18 give 4.8% of images, those 80 or older 0.8%



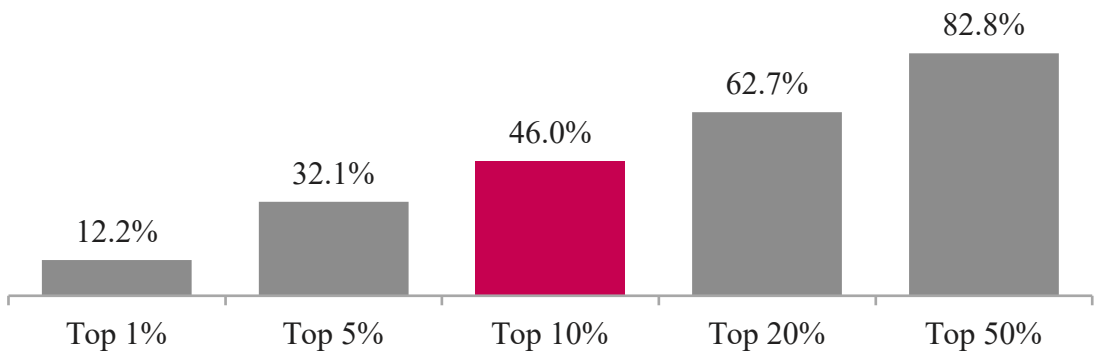
Male patients give 55.6% of images, female patients 44.4%



AP share is 35.8% in training and 59.8% in the balanced test set



The top 10% of patients give 46.0% of training images



Evaluation Performance on Tail Classes

We trained the model and evaluated them on balanced dataset.

Tail class has problem in both **AUROC** and **per-class accuracy.**

Especially, **tail class collapses** with accuracy this is because the model is trained to predict head.

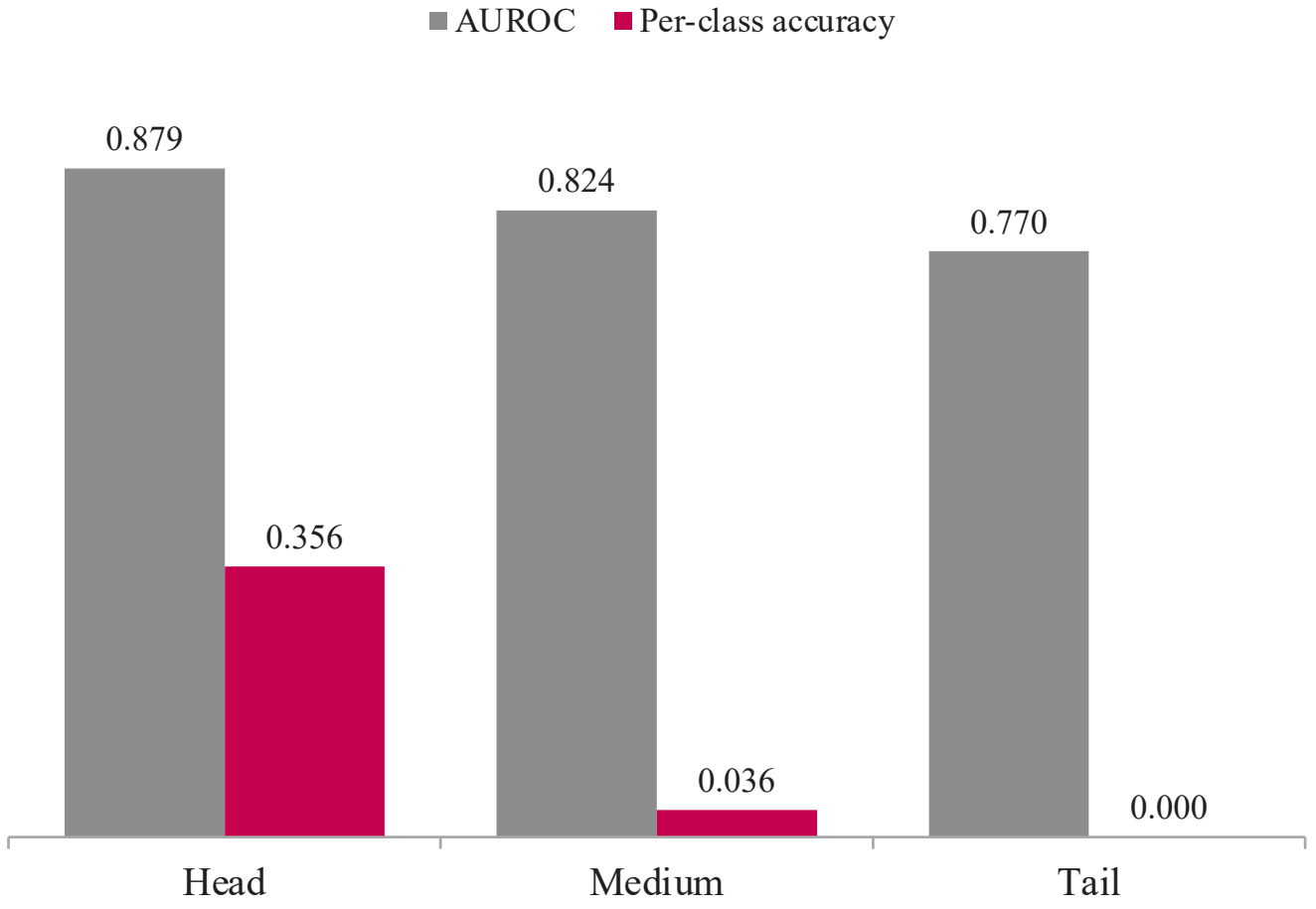
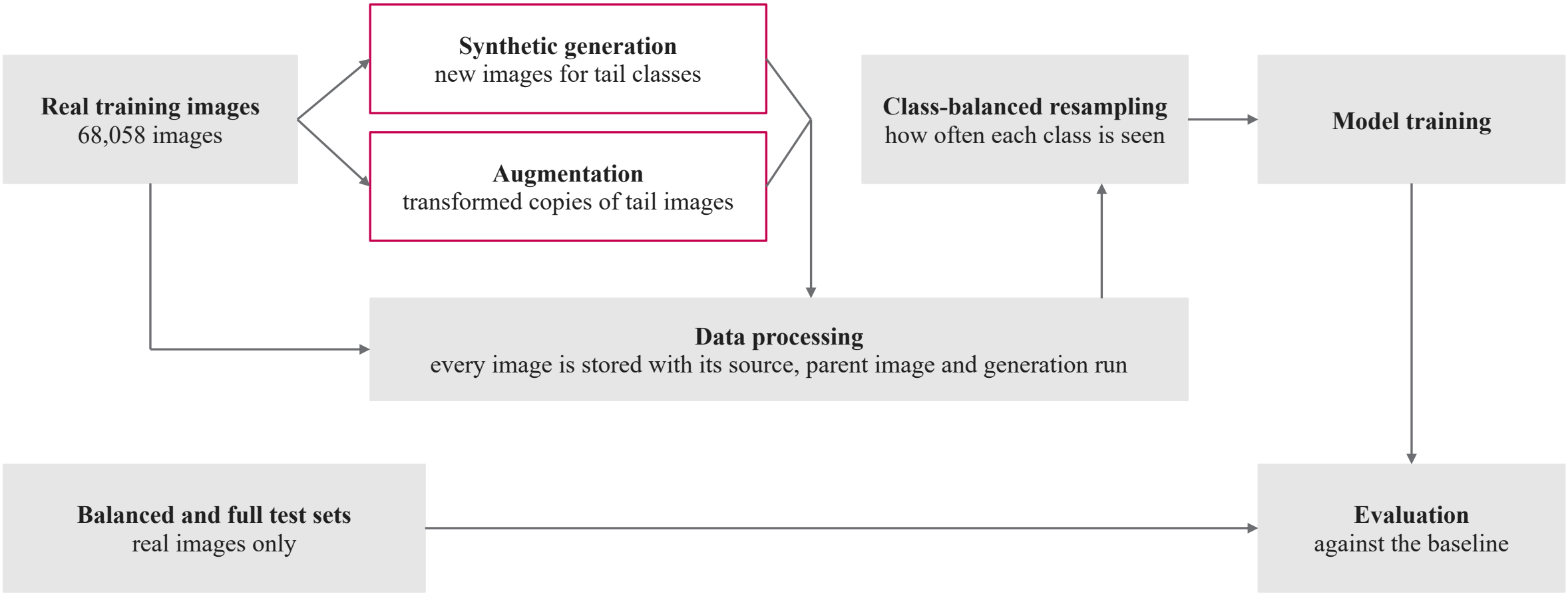


Figure 2. Our baseline on the balanced test set

Solution: Synthetic Data Generation and Data Augmentation

The tail fails because it has too few training examples, so we add examples on the data side methods.



Goal: Tail-Class Performance Improvement

We aim to **show analysis** and **solve** the different long-tail problem in NIH-CXR-LT dataset through **data generation** and **augmentation**.

Setting	Augmentation	Resampling	Synthetic data	Tail-class accuracy	
				Balanced test	Full test
Baseline				0.000	
+ Augmentation	O				
+ Resampling		O			
+ Synthetic data			O		
Combined	O	O	O		

Table 2. Planned comparison against the baseline

Thank you for listening!

Any questions?

CONTENTS

01 Problem Overview

02 Data Analysis

03 Baseline Diagnosis

04 Proposed Approach

05 Evaluation Plan

1

Problem Overview

2

Data Analysis

4

Proposed Approach

3

Baseline Diagnosis

5

Evaluation Plan