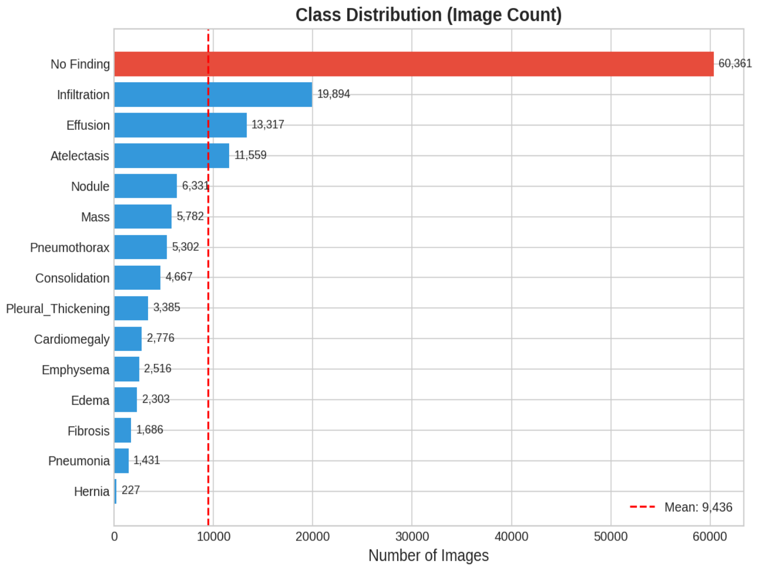
■ NIH Chest X-ray14 Dataset

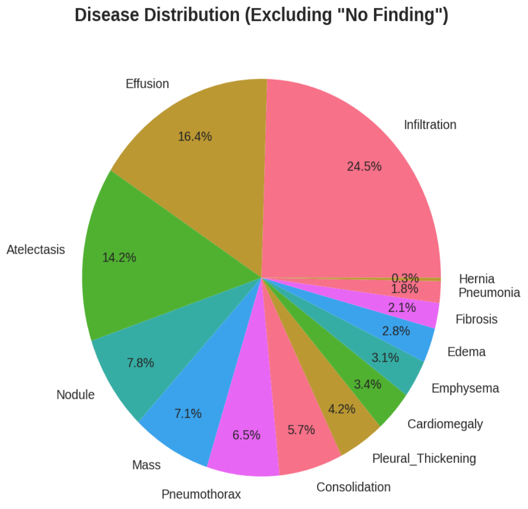
Exploratory Data Analysis Report

■ Dataset Overview

Metric	Value
Total Images	112,120
Unique Patients	30,805
Train Set	86,524 (77.2%)
Test Set	25,596 (22.8%)
Image Dimensions	~2518 × 2544 px
Multi-label Images	20,796 (18.5%)

■ Class Distribution





Disease Counts:

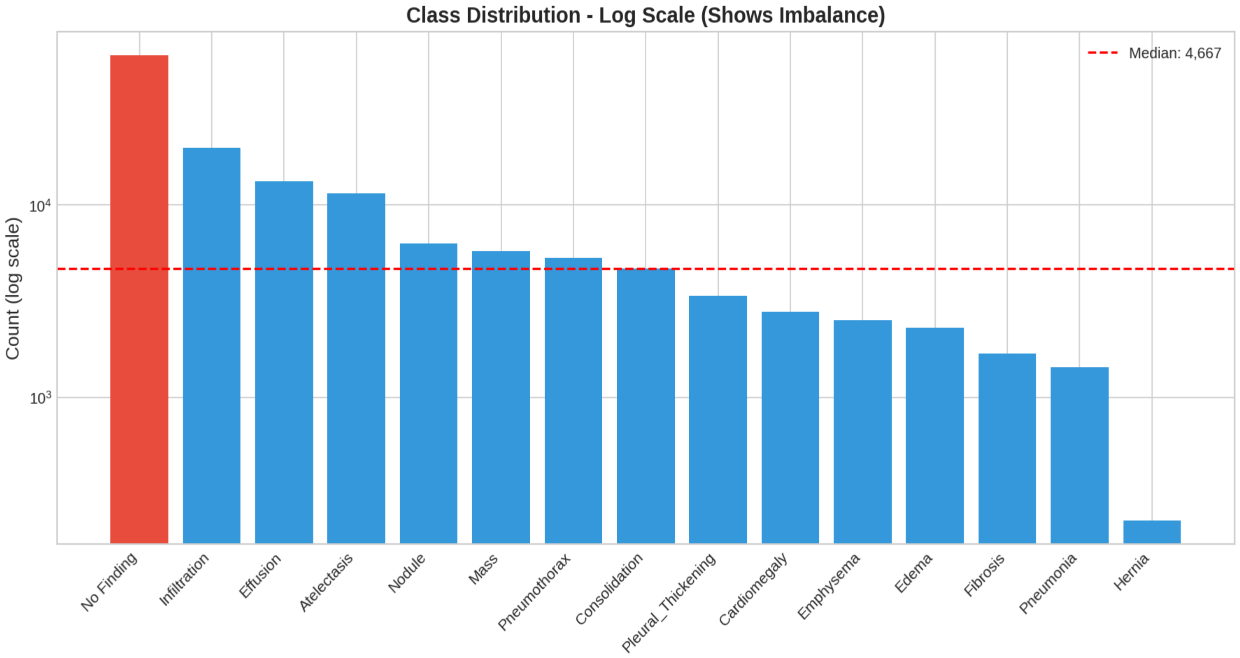
Class	Count	%
No Finding	60,361	53.8%

Infiltration	19,894	17.7%
Effusion	13,317	11.9%
Atelectasis	11,559	10.3%
Nodule	6,331	5.7%
Mass	5,782	5.2%
Pneumothorax	5,302	4.7%
Consolidation	4,667	4.2%
Pleural_Thick.	3,385	3.0%
Cardiomegaly	2,776	2.5%
Emphysema	2,516	2.2%
Edema	2,303	2.1%
Fibrosis	1,686	1.5%
Pneumonia	1,431	1.3%
Hernia	227	0.2%

■■ Class Imbalance Analysis

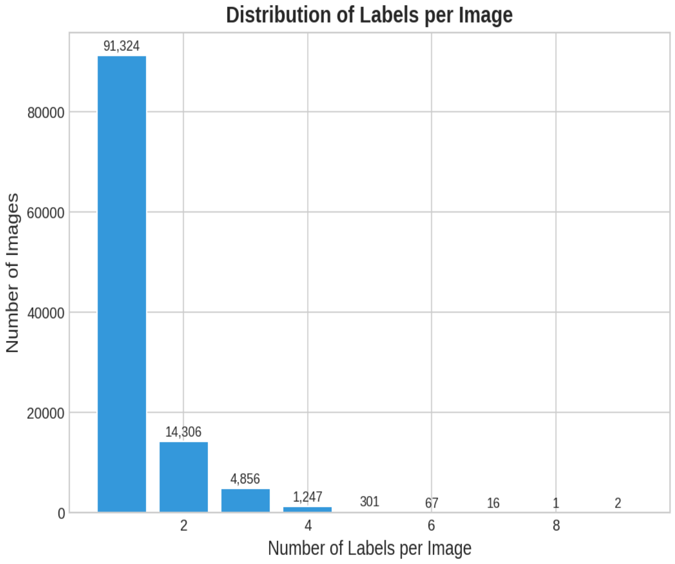
Imbalance Ratio: 266:1 (No Finding vs Hernia)

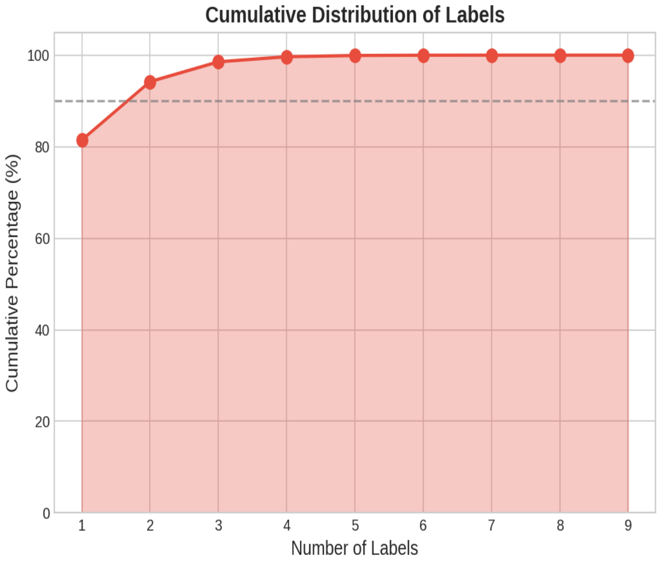
Recommendation: Use weighted loss or focal loss


Class Distribution - Log Scale (Shows Imbalance)

■■ Multi-Label Analysis

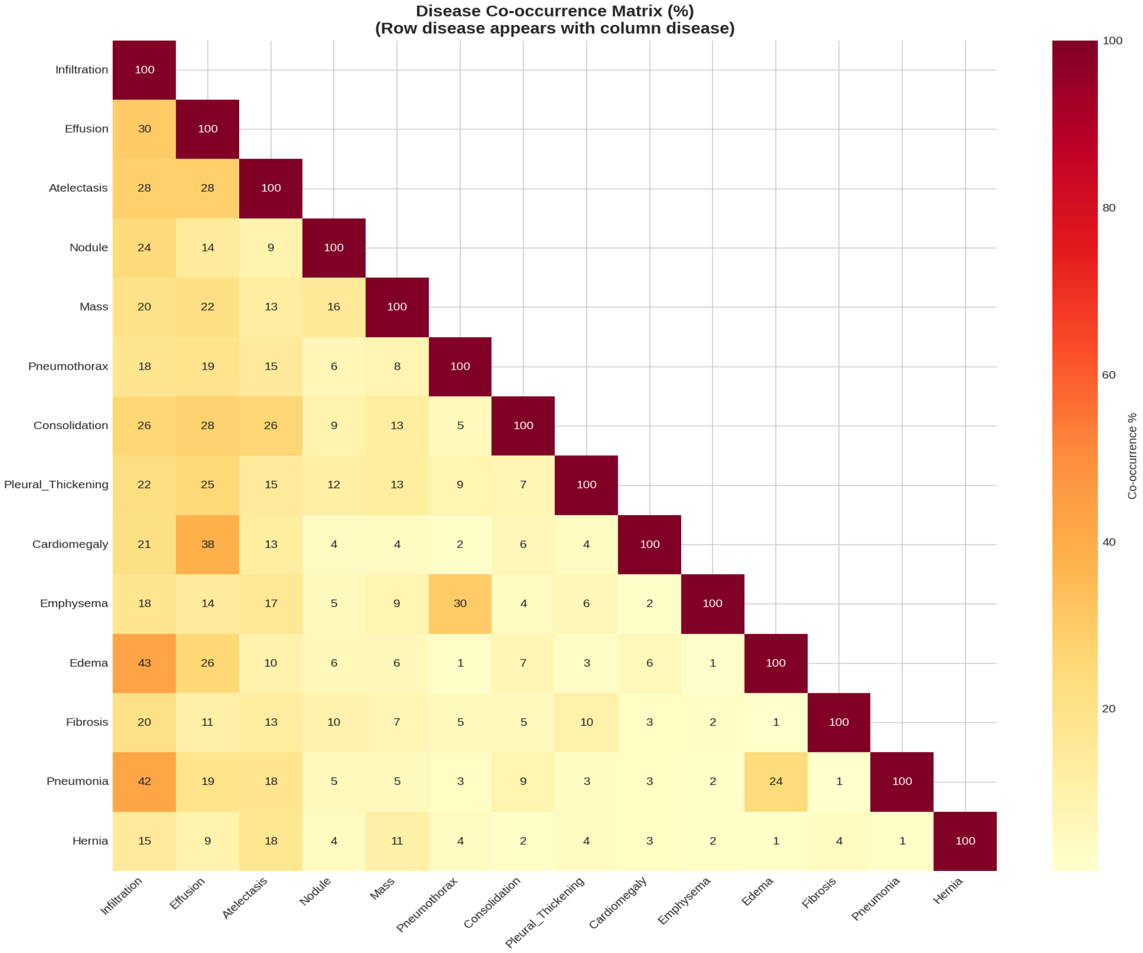
- **Mean labels per image:** 1.26
- **Max labels:** 9 diseases in single image
- **Single-label:** 91,324 (81.5%)
- **Multi-label:** 20,796 (18.5%)


Distribution of Labels per Image


Cumulative Distribution of Labels

■ Disease Co-occurrence

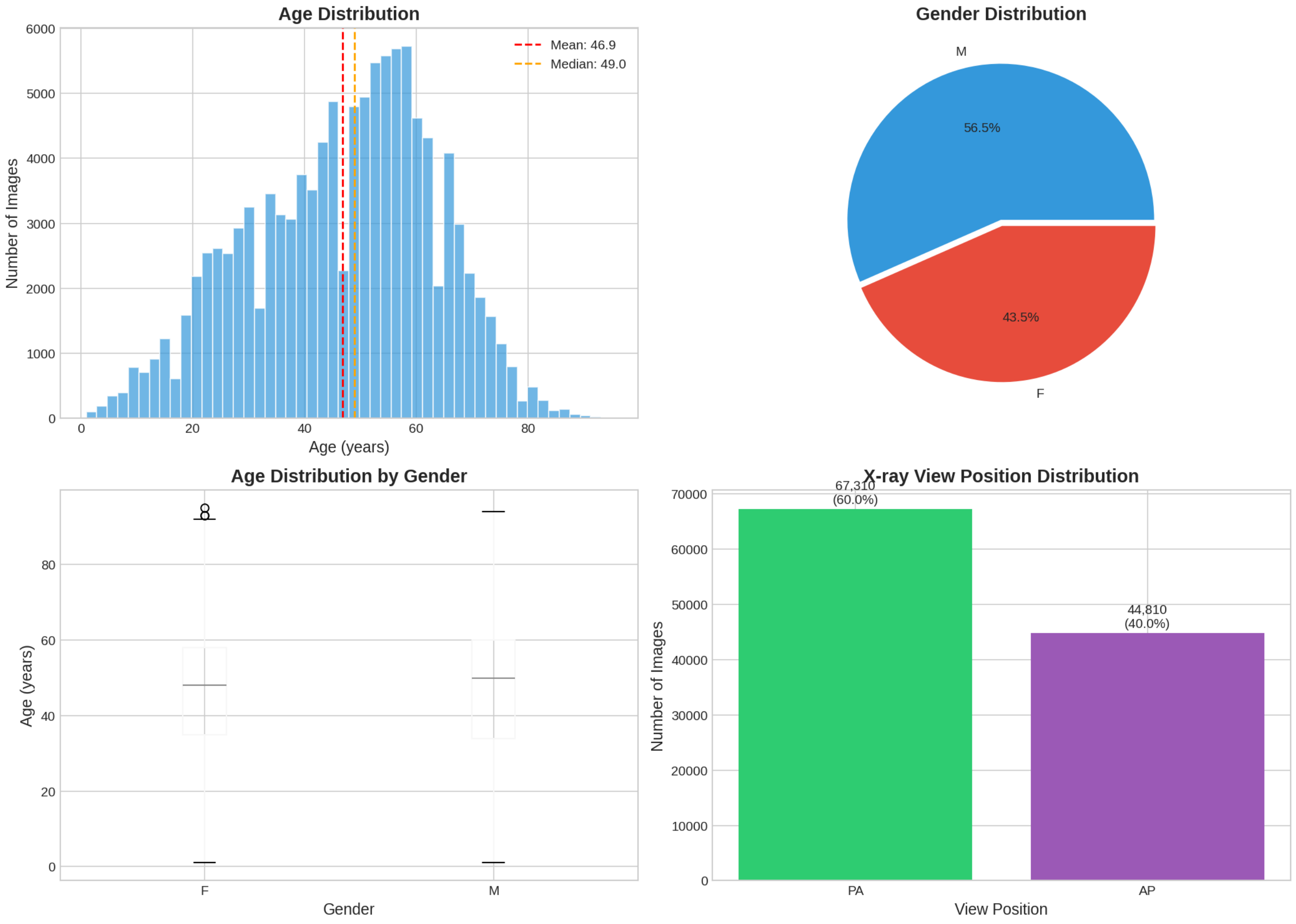
Top combinations: Infiltration+Effusion (4,000), Effusion+Atelectasis (3,275)



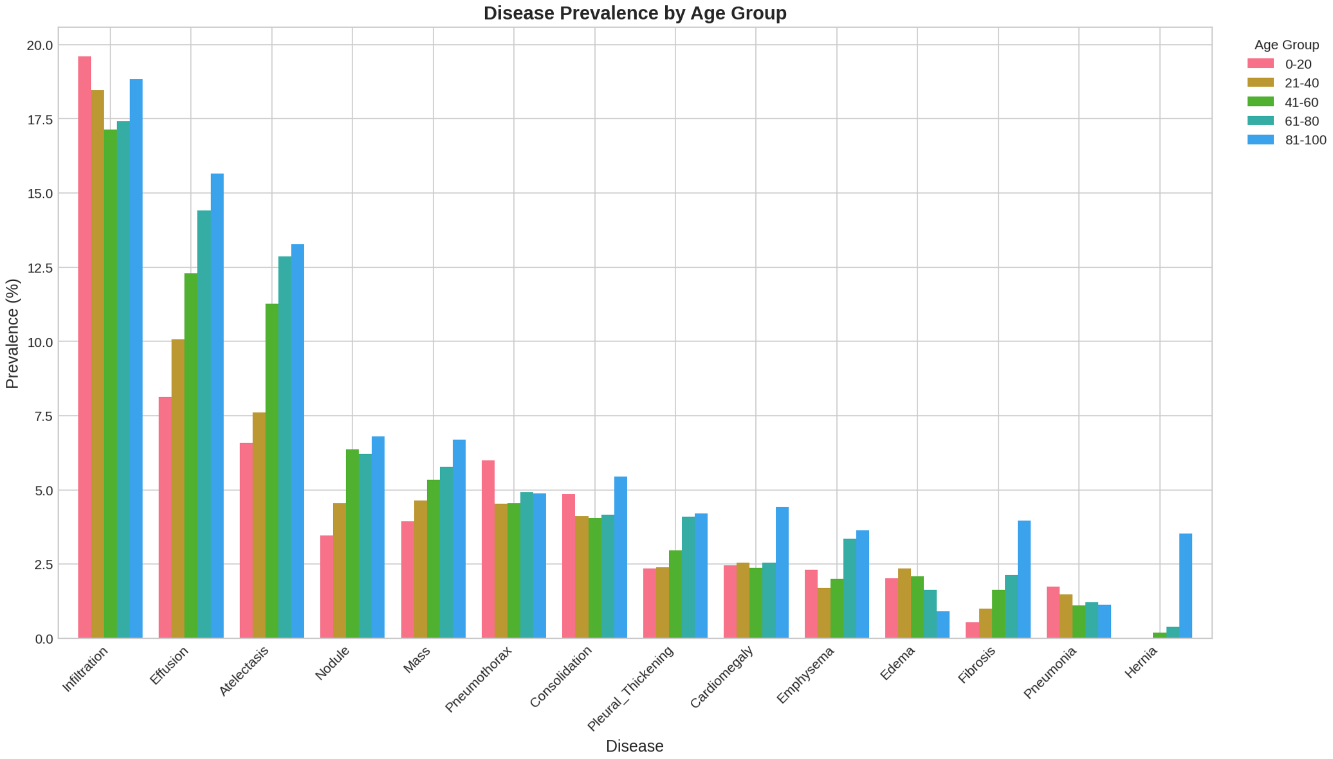
	Infiltration	Effusion	Atelectasis	Nodule	Mass	Pneumothorax	Consolidation	Pleural_Thickening	Cardiomegaly	Emphysema	Edema	Fibrosis	Pneumonia	Hernia
Infiltration	100													
Effusion	30	100												
Atelectasis	28	28	100											
Nodule	24	14	9	100										
Mass	20	22	13	16	100									
Pneumothorax	18	19	15	6	8	100								
Consolidation	26	28	26	9	13	5	100							
Pleural_Thickening	22	25	15	12	13	9	7	100						
Cardiomegaly	21	38	13	4	4	2	6	4	100					
Emphysema	18	14	17	5	9	30	4	6	2	100				
Edema	43	26	10	6	6	1	7	3	6	1	100			
Fibrosis	20	11	13	10	7	5	5	10	3	2	1	100		
Pneumonia	42	19	18	5	5	3	9	3	3	2	24	1	100	
Hernia	15	9	18	4	11	4	2	4	3	2	1	4	1	100

Patient Demographics

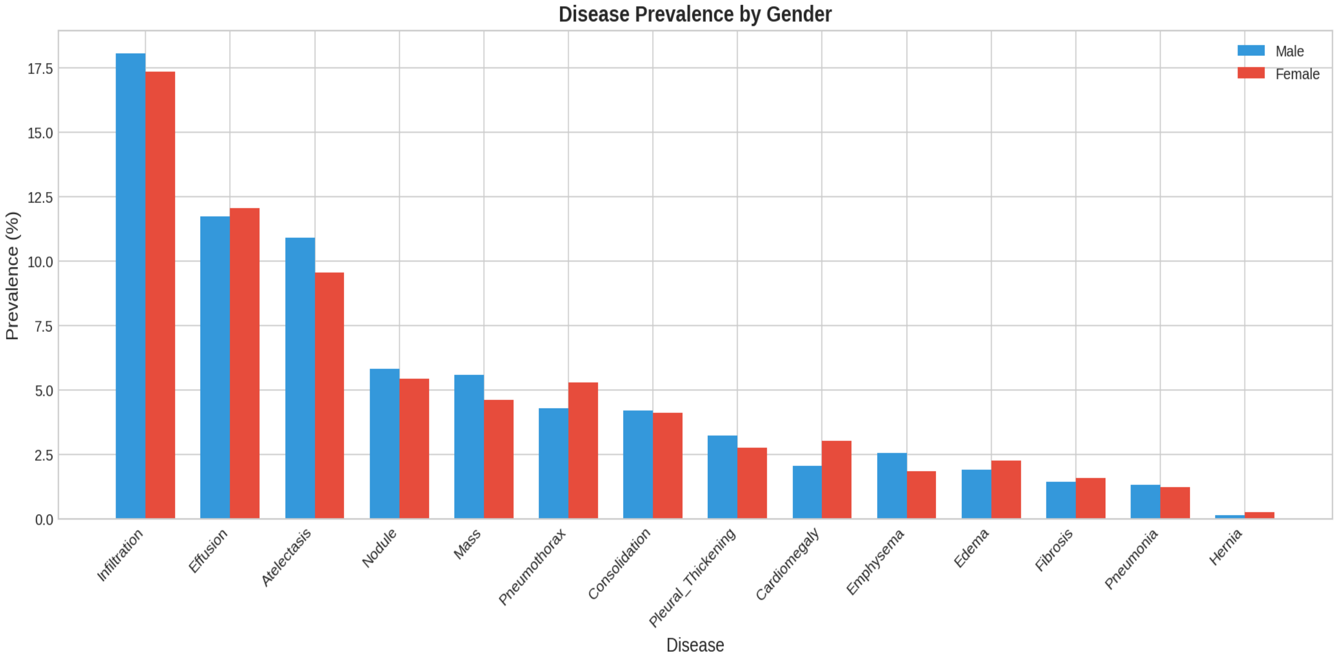
- **Age:** Mean 46.9 years (range 1-95)
- **Gender:** Male 56.5%, Female 43.5%
- **View:** PA 60%, AP 40%



Disease Prevalence by Age

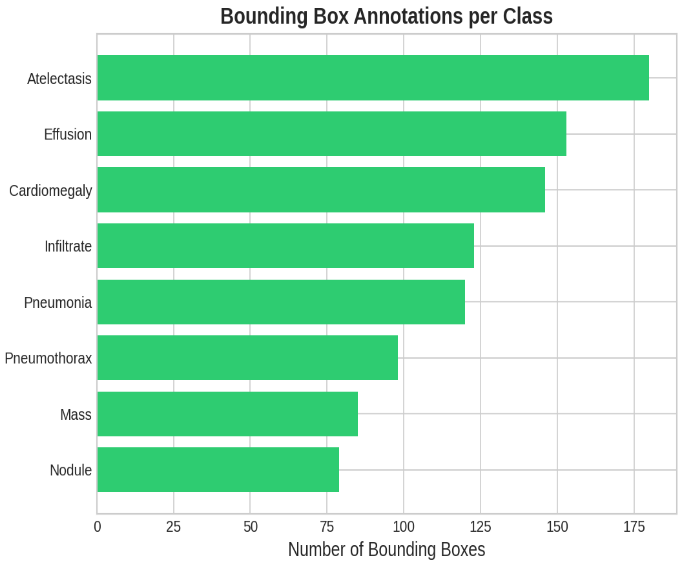


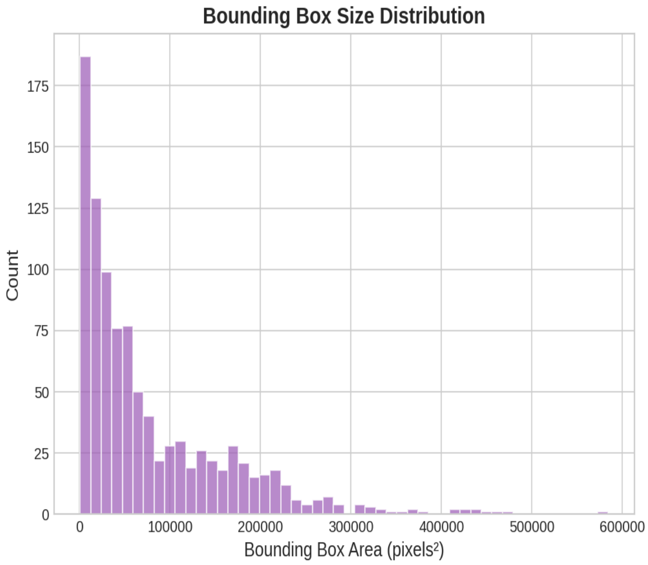
Disease Prevalence by Gender



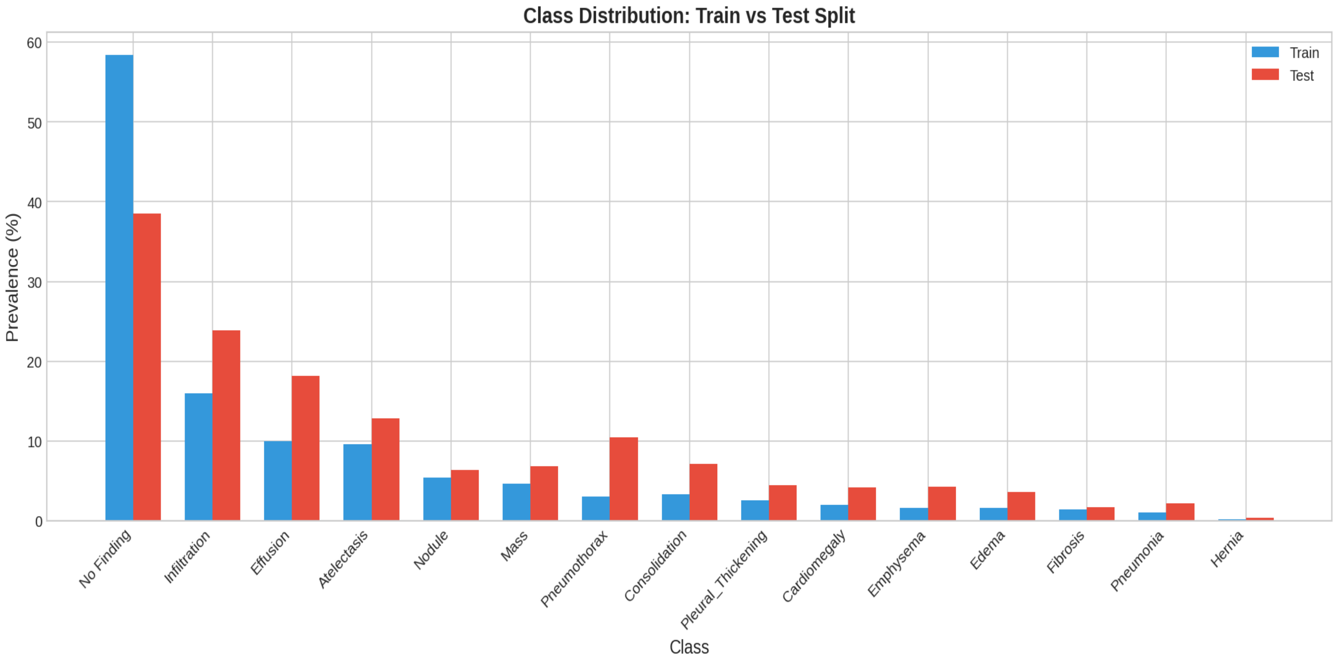
■ Bounding Box Analysis

■■ **Warning:** Only 984 bounding boxes (<1% coverage) - NOT suitable for YOLO!


Bounding Box Annotations per Class


Bounding Box Size Distribution

■ Train/Test Split


Class Distribution: Train vs Test Split

■ Key Insights & Recommendations

Finding	Recommendation
Task Type	Multi-label Classification
Loss Function	Weighted BCE or Focal Loss
Output	Sigmoid (NOT Softmax)
Metric	AUC-ROC
Model	DenseNet-121 or ConvNeXt
YOLO?	■ NO - Insufficient BBox data

Suggested Models:

1. DenseNet-121 (CheXNet architecture - state-of-the-art)

2. ConvNeXt (modern architecture)

3. EfficientNet (good accuracy/efficiency)

4. Vision Transformers (ViT, Swin)