CHEST X-RAY DISEASE CLASSIFICATION USING DEEP LEARNING

PRESENTATION

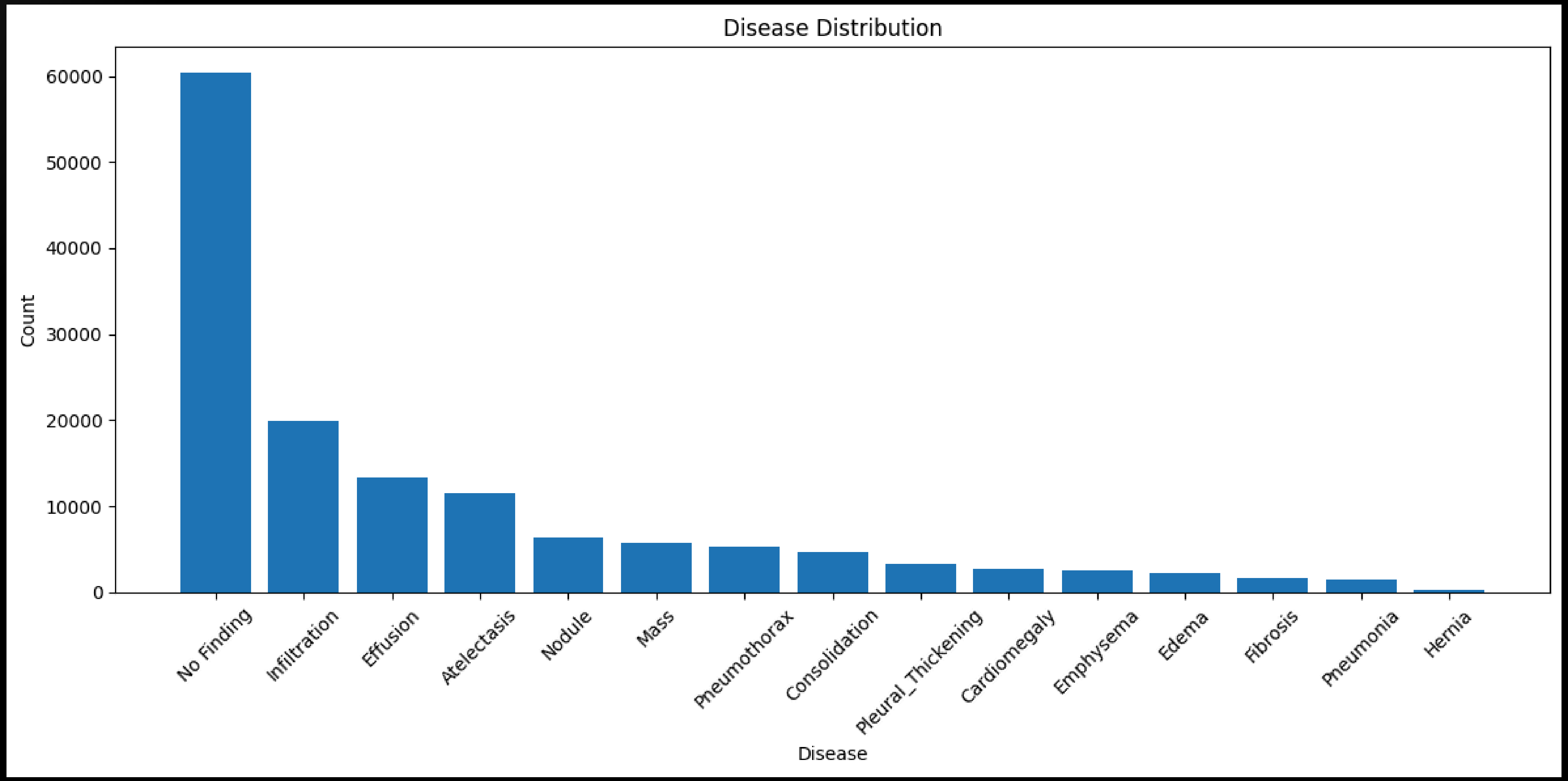
PRESENTED BY
ALİ SEYHAN
BAKİ TURHAN

ABOUT DATASET

- Dataset: NIH Chest X-rays
- Images: 112,120 frontal-view chest X-rays
- Patients: Over 30,000 unique patients
- Labels: 14 diseases + 1 “No Finding” category
- Multi-label Format: Each image can have zero or more disease labels.

- No Finding
- Infiltration
- Effusion
- Atelectasis
- Nodule
- Mass
- Pneumothorax
- Consolidation
- Pleural_Thickening
- Cardiomegaly
- Emphysema
- Edema
- Fibrosis
- Pneumonia
- Hernia

ABOUT DATASET



About 50% of the dataset is labeled “No Finding”.

There is also very little sample data from some diseases.

Therefore, we applied both class balancing and data reduction.

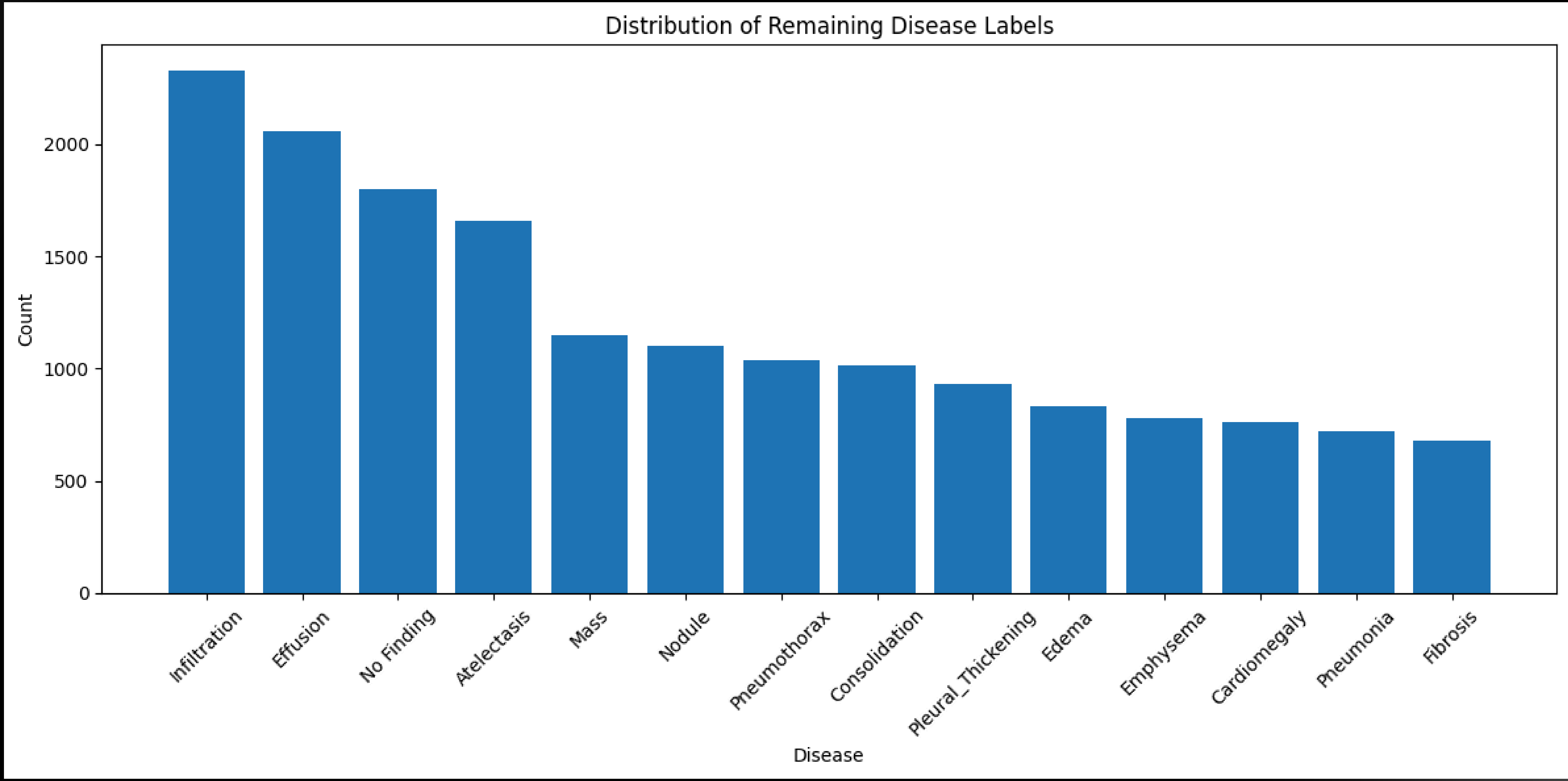
CLASS IMBALANCE AND PREPROCESSING

The dataset is highly imbalanced — some diseases have very few samples.

To address this issue:

- Hernia class was removed due to extreme underrepresentation.
- Undersampling applied to dominant classes such as "No Finding."
- Data Augmentation techniques were used to increase diversity.

ABOUT DATASET



MODEL ARCHITECTURES

DenseNet121

- The DenseNet121 model was loaded with weights pre-trained on ImageNet.
- The feature extractor layers were frozen for the first 5 epochs. From epoch all layers were made trainable for fine-tuning.
- The original classifier part was removed and replaced with the following structure:

1. Linear(in_features → 512)
2. ReLU
3. Dropout(0.5)
4. Linear(512 → num_classes)

- BCEWithLogitsLoss was used for multi-label classification.
- Loss function with pos_weight was defined to compensate for class imbalance.
- Early stopping was applied during training and validation.

VGG19

- The VGG19 model was also loaded with ImageNet pretrained weights.
- The first few convolution layers were frozen (transfer learning).
- Classifier layer was reconstructed as follows:

1. Linear(25088 → 4096) → ReLU → Dropout(0.5)
2. Linear(4096 → 1024) → ReLU → Dropout(0.5)
3. Linear(1024 → num_classes)

- The outputs were adjusted to be suitable for multi-label classification.
- BCEWithLogitsLoss and appropriate optimization strategies (Adam + weight decay) were used in training.

WHAT IS

SUBSET ACCURACY?

A strict measure of accuracy used in multi-label classification problems. If all labels of an instance are correctly predicted, it receives a score of 1, otherwise it is given a score of 0. Therefore, even a small error will cause the entire prediction to be considered incorrect. Low subset accuracy indicates that the model is struggling to completely predict all of the samples.

MACRO CLASS-WISE ACCURACY?

Accuracy for each class is calculated separately, then averaged. This metric provides balance across classes and ensures that rare classes contribute equally to the overall success. It provides a fairer assessment, especially in unbalanced data sets.

DENSENET121

```
📊 Classification Report:
                    precision    recall  f1-score   support

       Atelectasis       0.29      0.63      0.40       248
      Cardiomegaly       0.39      0.49      0.43        92
     Consolidation       0.24      0.66      0.35       155
             Edema       0.41      0.51      0.46       127
          Effusion       0.45      0.71      0.55       307
         Emphysema       0.65      0.51      0.57       115
          Fibrosis       0.22      0.37      0.28        97
      Infiltration       0.36      0.67      0.47       320
              Mass       0.25      0.54      0.34       178
        No Finding       0.44      0.65      0.52       277
            Nodule       0.22      0.50      0.30       159
Pleural_Thickening       0.17      0.68      0.27       140
         Pneumonia       0.24      0.41      0.30       107
      Pneumothorax       0.42      0.56      0.48       145

         micro avg       0.32      0.60      0.41      2467
         macro avg       0.34      0.56      0.41      2467
      weighted avg       0.35      0.60      0.43      2467
       samples avg       0.34      0.57      0.40      2467

🔬 AUC Scores per Class:
Atelectasis: 0.6868
Cardiomegaly: 0.8771
Consolidation: 0.7353
Edema: 0.8486
Effusion: 0.7967
Emphysema: 0.8743
Fibrosis: 0.7563
Infiltration: 0.6905
Mass: 0.7120
No Finding: 0.7918
Nodule: 0.6545
Pleural_Thickening: 0.6785
Pneumonia: 0.6866
Pneumothorax: 0.8391

✅ Subset Accuracy: 0.1231
✅ Macro Class-wise Accuracy: 0.7856

📈 Macro AUC Score: 0.7591
📈 Micro AUC Score: 0.7708
```

According to the classification report in the image, the overall accuracy of the model is low (subset accuracy: 0.1231),

However, significant results were obtained for certain diseases. In particular, high AUC scores are noteworthy in classes such as Cardiomegaly (AUC: 0.8771) and Emphysema (AUC: 0.8743).

The macro F1-score is 0.41 and the macro AUC score is 0.7591, indicating that the model's overall average level of success across classes is reasonable.

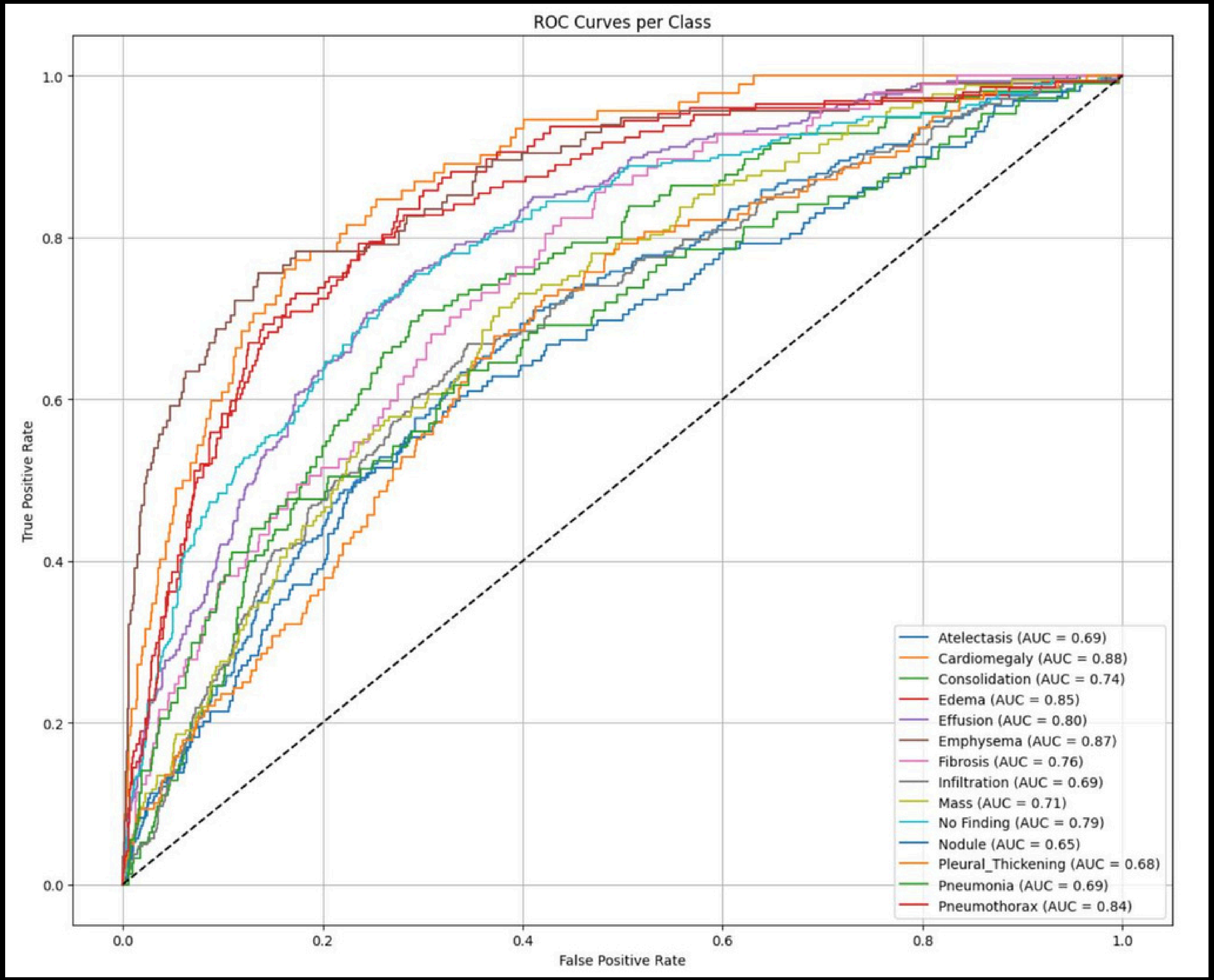
These results show that the model can discriminate some diseases better than others, but the imbalance between classes and the multi-label structure pose a challenge.

DENSENET121

ROC curves evaluate the classification performance of the model in terms of False Positive Rate and True Positive Rate. The closer the AUC value is to 1, the better the model discriminates the relevant class.

The highest AUC scores were observed for Cardiomegaly (0.88), Emphysema (0.87) and Edema (0.85). This shows the model's success in detecting these diseases. On the other hand, the AUC scores were lower in classes such as Pleural_Thickening (0.68) and Nodule (0.65), indicating that the model made more errors in these classes.

Overall, the ROC curves show that the model has strong discrimination ability in some disease types, but the imbalance between classes and overlapping symptoms limit the success in some classes.



DENSENET121

Highest Overall Performance

- DenseNet121 achieved the best results among all tested models in the multi-label chest X-ray classification task.

Strong Class-wise Performance

Performed especially well on:

- Edema
- Cardiomegaly
- No Finding

Robust to Class Imbalance

- Despite dataset imbalance, DenseNet121 delivered relatively balanced performance across all classes.

VGG19

📊 Classification Report:

	precision	recall	f1-score	support
Atelectasis	0.27	0.59	0.37	248
Cardiomegaly	0.28	0.32	0.30	92
Consolidation	0.21	0.53	0.30	155
Edema	0.32	0.49	0.39	127
Effusion	0.42	0.60	0.49	307
Emphysema	0.30	0.56	0.39	115
Fibrosis	0.17	0.42	0.24	97
Infiltration	0.31	0.67	0.42	320
Mass	0.20	0.56	0.30	178
No Finding	0.40	0.62	0.49	277
Nodule	0.18	0.52	0.27	159
Pleural_Thickening	0.18	0.50	0.27	140
Pneumonia	0.18	0.36	0.24	107
Pneumothorax	0.34	0.48	0.39	145
micro avg	0.27	0.55	0.36	2467
macro avg	0.27	0.52	0.35	2467
weighted avg	0.29	0.55	0.37	2467
samples avg	0.29	0.53	0.34	2467

🔬 AUC Scores per Class:
Atelectasis: 0.6461
Cardiomegaly: 0.7438
Consolidation: 0.6914
Edema: 0.8116
Effusion: 0.7525
Emphysema: 0.7631
Fibrosis: 0.7009
Infiltration: 0.6473
Mass: 0.6539
No Finding: 0.7431
Nodule: 0.6446
Pleural_Thickening: 0.6490
Pneumonia: 0.6475
Pneumothorax: 0.7810

✅ Subset Accuracy: 0.0809
✅ Macro Class-wise Accuracy: 0.7562

📈 Macro AUC Score: 0.7054

📈 Micro AUC Score: 0.7375

Although the overall accuracy (subset accuracy) is low at 0.0809, the macro class-wise accuracy (0.7562) shows that the model predicts well in certain classes.

The highest AUC scores were observed for Edema (0.8116), Pneumothorax (0.7810) and Emphysema (0.7631). In contrast, the discrimination power of the model is lower in classes such as Nodule (0.6446) and Pleural_Thickening (0.6490). Overall, the model has a micro AUC score of 0.7375 and a macro AUC score of 0.7054, indicating that on average the model shows a balanced discrimination ability across classes.

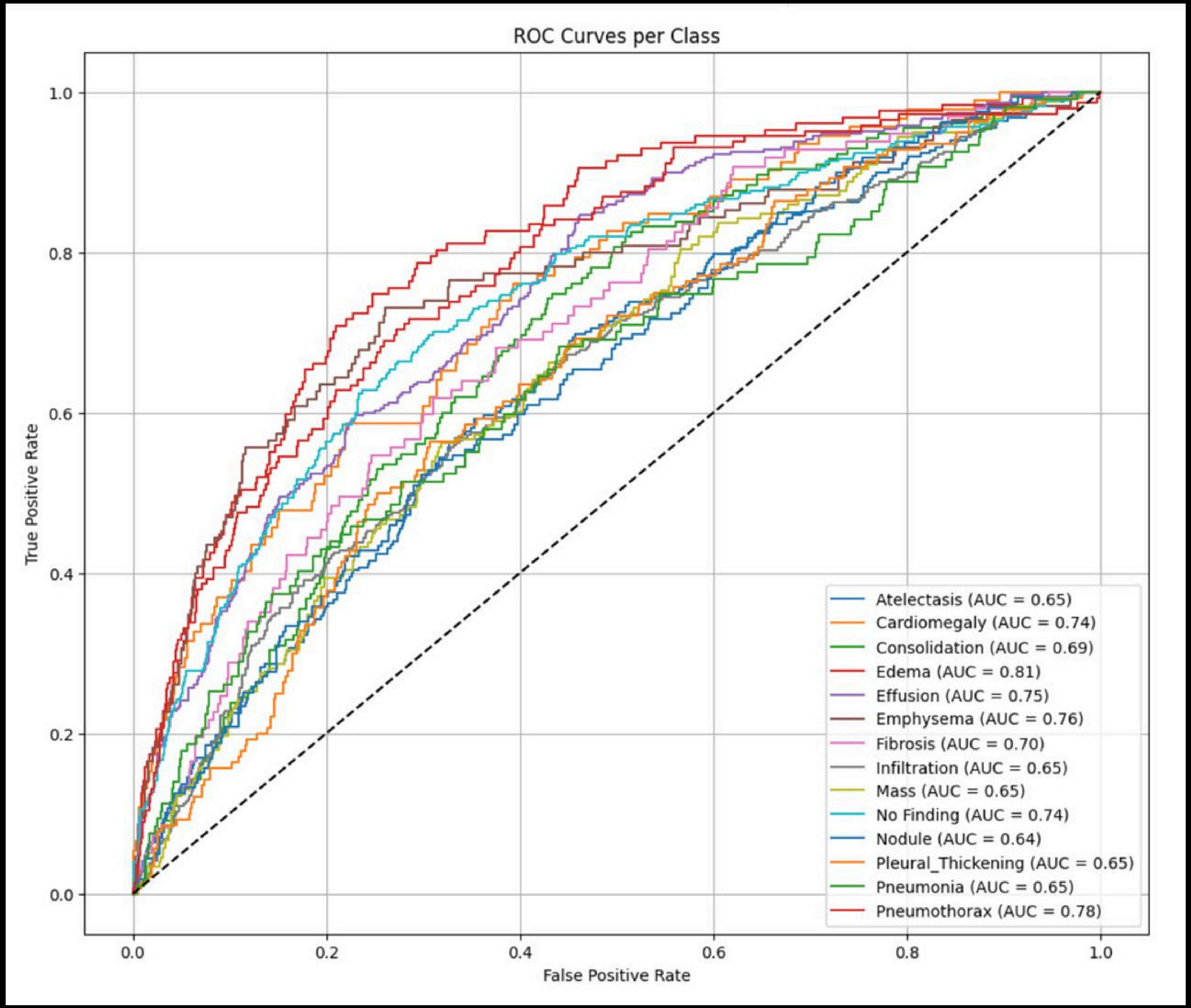
These results show that VGG19 can perform reasonably well in classification, but as with Densenet121, class imbalances and image similarities limit its success in some diseases.

VGG19

The model was able to discriminate diseases such as Edema (AUC = 0.81), Pneumothorax (AUC = 0.78) and Emphysema (AUC = 0.76) quite well, whereas its discrimination performance was poor in classes such as Nodule (AUC = 0.64), Atelectasis (AUC = 0.65) and Pleural_Thickening (AUC = 0.65).

The average AUC scores are around 0.70, indicating that VGG19 overall shows a balanced but limited success across classes.

These ROC curves suggest that VGG19 can exhibit high sensitivity, especially in some disease types, but that overlap between classes and data imbalances can affect the overall success.



VGG19

Lower Overall Performance

- VGG19 underperformed compared to DenseNet121 in the multi-label chest X-ray classification task.

Moderate Class-wise Performance

Performed relatively well on:

  - Edema
  - Emphysema
- Struggled with rare classes like:
  - Mass
  - Pleural Thickening

Challenges in Prediction

- Lower precision
- Inconsistent thresholding

Resulted in reduced confidence and reliability of predictions

THANK YOU

PRESENTED BY
ALİ SEYHAN
BAKİ TURHAN