# CheXFusion: Effective Fusion of Multi-View Features using Transformers for Long-Tailed Chest X-Ray Classification


**Dongkyun Kim**

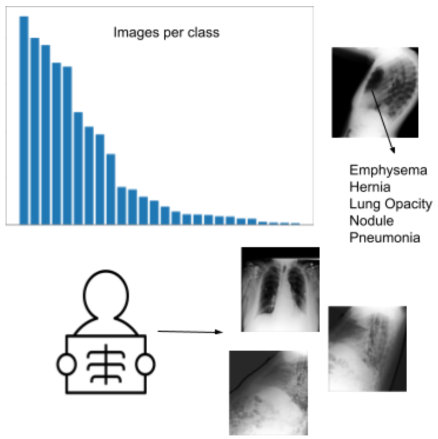
*Carnegie Mellon University, School of Computer Science*

Carnegie Mellon University
School of Computer Science


## Motivation

**Challenges Unique to Medical Image Classification**

- **Long-Tailed Distributions**: Medical conditions range from extremely common to very rare, and the dataset reflects this. Traditional ML models struggle with this imbalance, often misdiagnosing rare conditions.
- **Label Co-Occurrence**: Patients may have multiple conditions simultaneously, requiring a multi-label classification approach. Ignoring this can result in incomplete or inaccurate diagnoses.
- **Multiple Views**: Different imaging modalities and angles provide complementary information. Not exploiting these multiple views can lead to missed diagnostic cues.




**Gaps in Existing Solutions**

- Most existing approaches do not fully address these challenges, particularly in the context of chest X-rays.
- Previous work either focuses on single-view models or do not adequately handle class imbalance and label co-occurrence.

## Overview

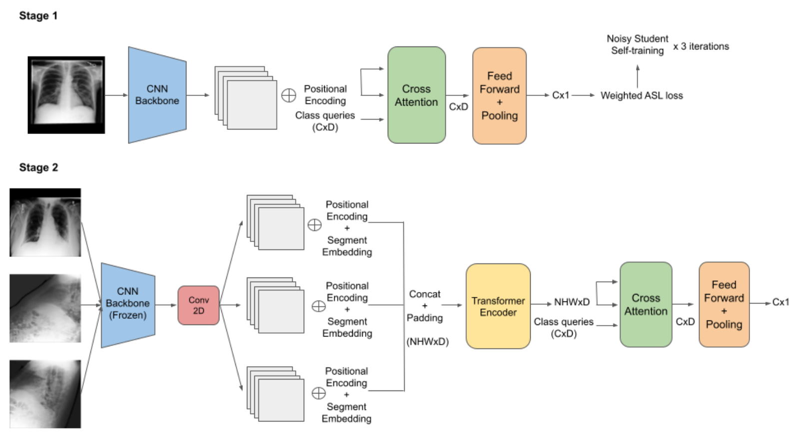
Our approach consists of **two stages**:

- **Stage 1**: Single-view CNN backbone trained as the feature extractor
- **Stage 2**: Transformer-based fusion module, CheXFusion, to integrate multi-view features.

**Key Takeaways**

- Our solution scores **first place** in both validation and test leaderboards of CXR-LT.
- CheXFusion can be applied as a plug-and-play method for other multi-label classification tasks.

## Methods




### 1. Backbone Pre-training

Objectives: A general feature extractor for each view in the subsequent fusion stage.

Key Innovations: Employ ML-Decoder as a classification head.

### 2. Transformer Fusion Model (CheXFusion)

Objectives: Integrate features from multi-view images effectively.

Key Innovations: Uses self-attention and cross-attention mechanisms for dynamic aggregation. Leverages positional encoding and segment embedding to handle permutation invariance.

### 3. Loss Function

Objectives: Tackle inter-class and intra-class imbalances in multi-label long-tailed classification.

Key Innovations: Weighted Asymmetric Loss

$$L = -\sum_{i=1}^{C} w_i((1-p_i)^{\gamma+} y_i \log(p_i) + p_{mi}^{\gamma-}(1-y_i)\log(1-p_{mi}))$$

### 4. Self-training

Objectives: Utilize additional unlabeled data

Key Contribution: Employs the "Noisy Student" approach, using a teacher model to produce pseudo-labels for unlabeled data

## Experiments

| wBCE | ASL [1] | Ml-decoder [23] | Hard pseudo | Soft pseudo | mAP | | | | AUC |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | total | head | medium | tail | total |
| | | | | | 0.311 | 0.601 | 0.231 | 0.122 | 0.816 |
| ✓ | | | | | 0.311 | 0.597 | 0.229 | 0.127 | 0.814 |
| | ✓ | | | | 0.313 | 0.603 | 0.231 | 0.126 | 0.815 |
| ✓ | ✓ | | | | 0.314 | 0.604 | 0.232 | 0.128 | 0.817 |
| ✓ | ✓ | ✓ | | | 0.322 | 0.609 | 0.234 | 0.146 | 0.821 |
| ✓ | ✓ | ✓ | ✓ | | 0.330 | 0.614 | 0.255 | 0.141 | 0.828 |
| ✓ | ✓ | ✓ | | ✓ | **0.336** | **0.612** | **0.270** | **0.143** | **0.832** |

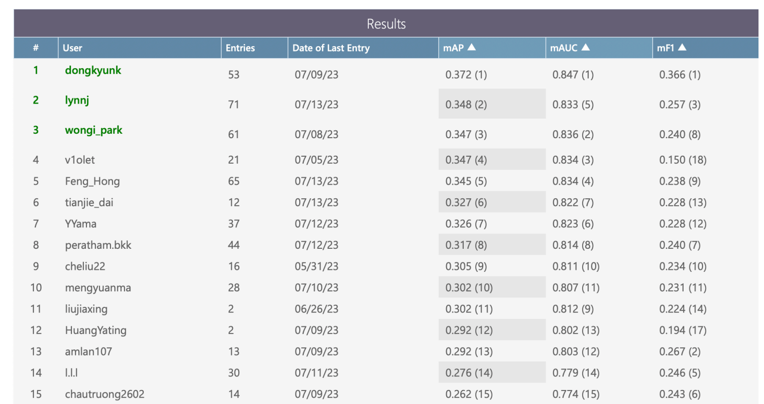
Table 2. Ablation studies on the various components of our proposed method.

| Backbone | Method | mAP | | | | AUC |
|---|---|---|---|---|---|---|
| | | total | head | medium | tail | total |
| ConvNeXt-S-1024 | Single-view | 0.340 | 0.616 | 0.270 | 0.152 | 0.833 |
| | Multi-view Weighted Average (4:6) | 0.355 | 0.626 | 0.284 | 0.170 | 0.841 |
| | Multi-view Weighted Average (5:5) | 0.357 | 0.627 | 0.288 | 0.172 | 0.842 |
| | Multi-view Weighted Average (6:4) | 0.359 | 0.628 | 0.296 | 0.167 | 0.844 |
| | Multi-view Weighted Average (7:3) | 0.362 | 0.629 | 0.306 | 0.167 | 0.847 |
| | Multi-view Concat | 0.357 | 0.622 | 0.293 | 0.173 | 0.839 |
| | ChexFusion (Ours) | **0.372** | **0.630** | **0.312** | **0.188** | **0.847** |

Table 1. Performance comparison of ChexFusion and the baseline methods on the CXR-LT validation set.

CheXFusion outperforms several baselines and achieves state-of-the-art performance on the CXR-LT dataset, indicating its potential for application in clinical settings

## Conclusion



Results

| # | User | Entries | Date of Last Entry | mAP ▲ | mAUC ▲ | mF1 ▲ |
|---|---|---|---|---|---|---|
| 1 | dongkyunk | 53 | 07/09/23 | 0.372 (1) | 0.847 (1) | 0.366 (1) |
| 2 | lynnj | 71 | 07/13/23 | 0.348 (2) | 0.833 (5) | 0.257 (3) |
| 3 | wongi_park | 61 | 07/08/23 | 0.347 (3) | 0.836 (2) | 0.240 (8) |
| 4 | v1olet | 21 | 07/05/23 | 0.347 (4) | 0.834 (3) | 0.150 (18) |
| 5 | Feng_Hong | 65 | 07/13/23 | 0.345 (5) | 0.834 (4) | 0.238 (9) |
| 6 | tianjie_dai | 12 | 07/13/23 | 0.327 (6) | 0.822 (7) | 0.228 (13) |
| 7 | YYama | 37 | 07/12/23 | 0.326 (7) | 0.823 (6) | 0.228 (12) |
| 8 | peratham.bkk | 44 | 07/12/23 | 0.317 (8) | 0.814 (8) | 0.240 (7) |
| 9 | cheliu22 | 16 | 05/31/23 | 0.305 (9) | 0.811 (10) | 0.234 (10) |
| 10 | mengyuanma | 28 | 07/10/23 | 0.302 (10) | 0.807 (11) | 0.231 (11) |
| 11 | liujiaxing | 2 | 06/26/23 | 0.302 (11) | 0.812 (9) | 0.224 (14) |
| 12 | HuangYating | 2 | 07/09/23 | 0.292 (12) | 0.802 (13) | 0.194 (17) |
| 13 | amlan107 | 13 | 07/09/23 | 0.292 (13) | 0.803 (12) | 0.267 (2) |
| 14 | l.l.l | 30 | 07/11/23 | 0.276 (14) | 0.779 (14) | 0.246 (5) |
| 15 | chautruong2602 | 14 | 07/09/23 | 0.262 (15) | 0.774 (15) | 0.243 (6) |

- We propose CheXFusion, a transformer-based fusion module that effectively integrates multi-view medical image features
- We conduct extensive experiments to verify the advantages of various data balancing techniques and self-training.
- Our solution achieves top performance in both the validation and test leaderboards of the CXR-LT shared task