An Optimized Ensemble Framework for Multi-Label Classification on Long-Tailed Chest X-ray Data

Jaehyup Jeong* Bosoung Jeoun* Yeonju Park*
KT Research & Development Center
KT Corporation
Seoul, Korea
{jaehyup.jeong, lynn.jeoun, park.yeonju}@kt.com

Bohyung Han†
ECE & IPAI
Seoul National University
Seoul, Korea
bhhan@snu.ac.kr

Abstract

Chest X-rays (CXR) are essential in the diagnosis of lung disease, but CXR image classification is challenging because patients often have multiple diseases simultaneously. This requires multi-label classification to identify multiple abnormalities within a single image, which is complicated by different disease patterns and overlapping pathologies. In addition, CXR image classification faces the problem of long-tail distribution, with few common and mostly rare diseases, which can lead to biased predictions, especially for rare classes. There have been limited attempts to address these challenges in the medical domain, and applying general domain approaches to medical data may not be straightforward due to the unique characteristics of medical data. This paper presents an optimized ensemble framework to solve multi-label long-tailed classification on the MIMIC-CXR-LT dataset, which is the main objective of the ICCV CVAMD 2023 workshop competition, CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays. Various experiments have been conducted, from architecture design to data augmentation, to identify the most suitable components. The proposed framework improves the performance of the long-tail distribution classification problem on class-imbalanced multi-label medical images and is placed in the top ranks in the CXR-LT competition.

1. Introduction

Chest X-ray plays a crucial role in diagnosing various lung diseases. Recent advances in deep learning have shown great promise in medical image analysis, assisting radiologists in the accurate detection of diseases [1, 2, 3, 4]. Nevertheless, the chest X-ray image classification task poses several practical challenges including class co-occurrences, long-tailed distribution, and so on.

Diagnosing chest X-rays involves a multi-label classification problem [5, 6, 7], as patients often present with multiple disease findings simultaneously. Unlike traditional classification tasks, where each image is assigned to a single class, multi-label classification requires models to capture multiple abnormalities within a single chest X-ray image. The coexistence of diverse disease patterns in a single image adds complexity to the classification process, requiring the recognition of intricate interrelationships among different pathologies. The overlapping nature of diseases further complicates accurate disease detection. Furthermore, chest X-ray image classification faces the long-tail distribution problem [11], where only a small subset of diseases is frequently observed while a majority of diseases are hardly observed. The imbalance in disease distribution poses challenges in model training as rare diseases often incur biased predictions leading to reduced performance on the less prevalent classes [10].

Few attempts have addressed the challenges in multi-label classification and long-tail distribution handling in the medical domain. While some approaches have been proposed for general domains [21, 23, 18], they may not be directly applicable due to the unique characteristics of medical data.

This paper presents an optimized ensemble framework for multi-label long-tailed classification on chest X-rays. To effectively handle the multi-label classification task, our approach employs the class-specific residual attention (CSRA) classifier head [24], which captures separate features for every class. Additionally, to tackle the long-tail problem, it leverages an ensemble technique, which combines predictions from the head-specific (HEAD), tail-specific (TAIL), and head-and-tail (ALL) models, yielding more balanced predictions and improved performance. We explored various model architectures, loss functions, optimizers, and data augmentations for our task. To fur-

*Equal contribution
†Corresponding author

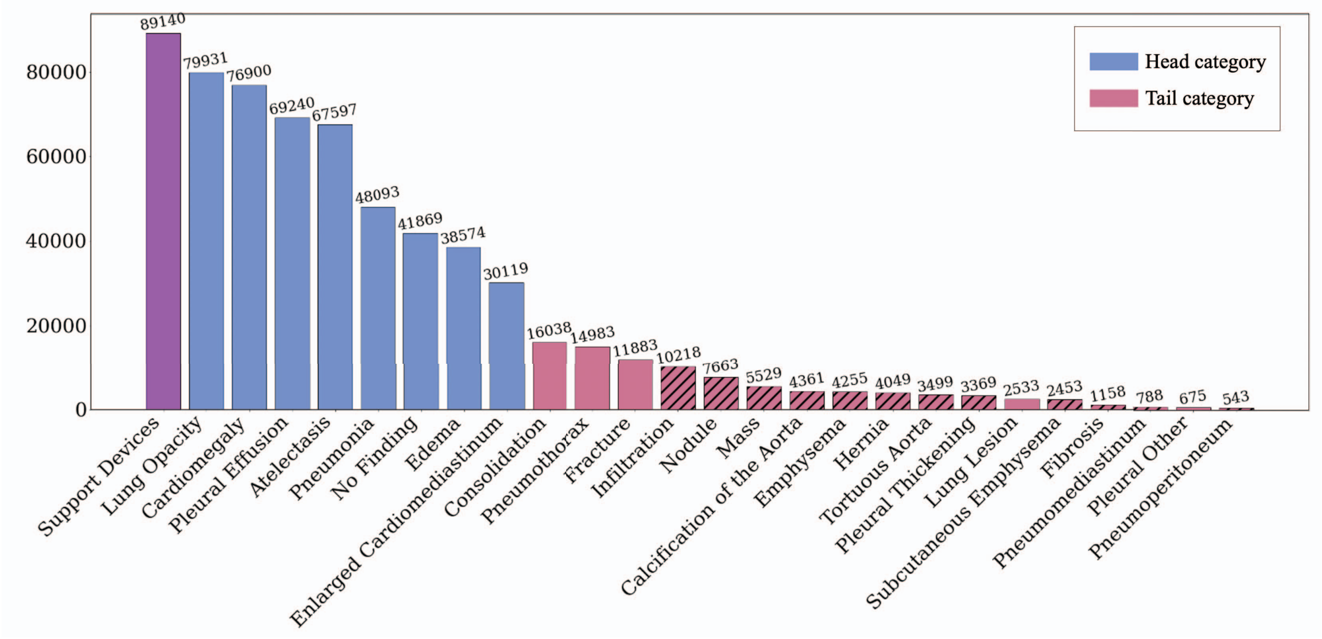

Figure 1. Long-tailed distribution of a total of 26 pulmonary disease labels for MIMIC-CXR-LT with both original and newly added classes. The values for each bar indicates the number of class frequency. The textured bars represent the added classes for long-tailed distribution. The "Support Devices" class (purple) is included in both head and tail categories.

ther enhance the performance, CXR pre-training on the NIH ChestXRay14 [20] dataset was conducted, enabling the model to leverage domain-specific information. We performed extensive experiments on the MIMIC-CXR-LT dataset provided by ICCV CVAMD 2023 workshop competition *"CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays (CXR-LT)*[1]*"* and validated the effectiveness of our approach.

To summarize, the main contributions of this work are as follows:

- We successfully address the multi-label classification by incorporating the CSRA classifier head into our model.
- We effectively address the long-tail distribution problem by leveraging an ensemble approach via a combining prediction of the HEAD, TAIL, and ALL models.
- We explore various model architectures, loss functions, optimizers, data augmentation techniques, and pre-trained models, and evaluate their effectiveness on the chest X-ray image classification dataset.

The rest of this paper is organized as follows. Section 2 defines our task and discusses the dataset. Our approach is described in Section 3, and experimental results are presented in Section 4. We conclude this work and discuss future work in Section 5.

2. Multi-Label Long-Tailed Classification on Chest X-Rays

This section first defines our target task in this work and then discusses the specifications of the dataset and evaluation metrics.

2.1. Task Definition

The main objective is to develop a classification model that accurately identifies and classifies a wide range of clinical findings, including lung opacity, infiltration, pleural effusion, and other thorax diseases, in chest X-ray images. This task requires multi-label classification as more than one abnormality can be observed simultaneously on a single chest X-ray image. In the absence of findings, the model should predict "No Finding", which can co-occur with the "Support Devices" class but not with any other ones.

In addition, the distribution of abnormalities follows a long-tail pattern, where a small number of common abnormalities have abundant instances available for analysis, while the majority of abnormalities are rarely observed. For example, as depicted in Figure 1, the most prevalent clinical finding, *lung opacity*, has 79,931 instances, whereas the least frequent finding, *pneumoperitoneum*, has only 543 instances.

2.2. Dataset

Our work is based on an expanded version of MIMIC-CXR-JPG [12] dataset, MIMIC-CXR-LT, which is com-

[1] `https://codalab.lisn.upsaclay.fr/competitions/12599`

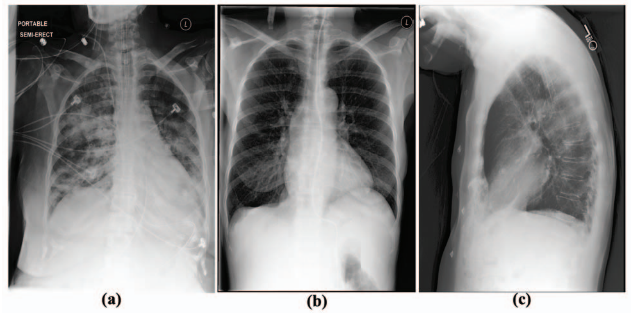

Figure 2. The examples of Chest X-rays in MIMIC-CXR-LT with view positions of (a) Anterior-Posterior (AP), (b) Posterior-Anterior (PA), and (c) Lateral.

Train	Validation	Development	Test	Total
224,894	40,000	36,769	75,492	377,110

Table 1. **MIMIC-CXR-LT Data split.** The number of CXRs for each data split provided from ICCV CVAMD 2023 workshop competition.

posed of a total of 26 pathology classes and 377,110 CXR images with three different view positions as shown in Figure 2. The MIMIC-CXR-JPG originally has more than 200,000 CXRs in 14 classes along with corresponding clinical reports. Based on the approach of producing extra labels from studies of long-tailed [11] and multi-label classification [15], the labels of the new 12 classes are obtained by parsing matching reports. Consequently, the expanded dataset, MIMIC-CXR-LT, exhibits a strong long-tail distribution with multiple labels.

The MIMIC-CXR-LT dataset, provided by the ICCV CVAMD 2023 workshop competition, is officially split into training, development, and test sets, consisting of 264,849, 36,769, and 75,492 samples, respectively. The development and test sets are provided without labels, and the evaluation with the test set was allowed up to five times after the development phase ended. To evaluate benchmark performance during training, we tentatively split the given training set into two groups to generate an additional validation set as shown in Table 1. The data for this validation set was randomly chosen while maintaining a long-tailed class distribution similar to the whole dataset. Lastly, all view positions—Anterior-Posterior (AP), Posterior-Anterior (PA), and Lateral—were employed without the exclusion of specific views.

2.3. Evaluation Metrics

The evaluation of our approach was performed through three metrics: 1) mean macro Average Precision (mAP), 2) mean Area Under the Reciever Operating Characteristic Curve (mAUC), and 3) mean F1 score (mF1). Though mAUC is commonly used to evaluate the performance of classification tasks, it might be inappropriate due to the severe class imbalance in MIMIC-CXR-LT. Therefore, mAP was adopted as the primary evaluation metric for long-tailed multi-label classification task; it calculates the results through decision threshold and exhibits strong robustness to highly class-imbalanced data. Note that mAUC and mF1 were considered to estimate the validity of our approach's performance thoroughly.

Feature extractor	Classifier	Val. mAP
MLP-Mixer-Base	Linear	0.2067
PoolFormer-M48	Linear	0.2757
EfficientNet-B5	Linear	0.2739
ConvNeXt-Base	Linear	0.3025
ConvNeXt-Small	CSRA	0.3034
ConvNeXt-Base	CSRA	**0.3198**
ConvNeXt-Large	CSRA	0.3065

Table 2. **Architecture.** Comparison of feature extractor and classifier architectures. The upper block compares the performance of feature extractors with a linear classifier, and the lower block compares the performance of ConvNeXt variants with CSRA classifier.

3. Our Approach

This section presents our approach for solving multi-label long-tailed classification in detail. To begin with, the experiments for selecting a suitable model architecture were implemented by exploring existing pre-trained models. We identified an appropriate loss function and image input size using the validation set by referring to the earlier experimental results and tuned hyperparameters for optimization on the development set. To tackle the long-tailed problem, an ensemble approach was employed by integrating the best-optimized HEAD, TAIL, and ALL models. Further experiments for searching domain-specific pre-trained models were explored. In the end, our final proposed optimized ensemble framework, illustrated in Figure 3, consists of two phases: 1) pre-training phase, and 2) ensemble phase.

3.1. Label Grouping

To handle this skewed distribution, the thorax disease classes are divided into two groups: HEAD for classes with over 20,000 instances, and TAIL for the remaining classes. Additionally, the "Support Devices" class represents the presence of specific medical devices or equipment used for patient care and monitoring. Unlike other disease classes, it has the potential to appear simultaneously with the "No Finding" class. These distinguishing characteristics lead to its inclusion in both the head and tail categories.

3.2. Architecture

Table 2 shows a comparison between four different models including MLP-Mixer [19], PoolFormer [22], Efficient-

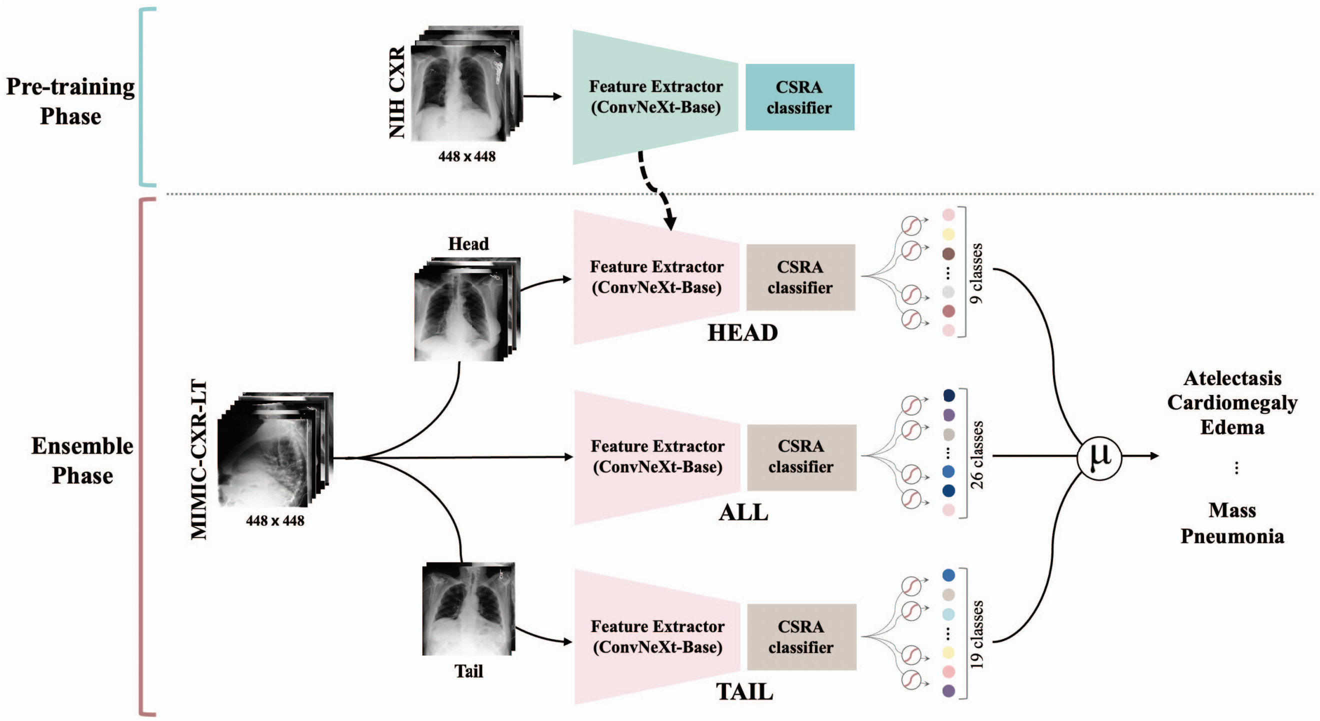

Figure 3. Optimized ensemble framework for long-tailed multi-label medical image classification using ConvNeXt-Base model with ImageNet-21K pre-trained model. The dashed arrow indicates the initialization of the feature extractors for MIMIC-CXR-LT. µ denotes the "Average" ensemble strategy.

Net [17], and ConvNeXt [13] pre-trained on ImageNet-21K [16] with a linear classifier. Among these models, ConvNeXt demonstrated superior performance. One of the notable strengths of CSRA [24] is its ability to effectively capture different spatial regions occupied by objects from different classes. To further leverage the benefits of CSRA and enhance our model's performance, the CSRA classifier was integrated with ConvNeXt across various model sizes. Such an integration facilitates the utilization of unique spatial information and class-specific features, resulting in more precise and robust predictions. After a thorough exploration and evaluation, we selected ConvNeXt-Base as the feature extractor and CSRA as the classifier, as the combination led to the best performance for our task.

3.3. Loss Function

In our pursuit of finding the most effective loss function for addressing the multi-label and long-tail problems in CXR images, various loss functions were explored thoroughly. They include the binary cross-entropy (BCE) loss, which is commonly used for multi-label classification tasks, the weighted binary cross-entropy (WBCE) loss, which is designed to handle class imbalance, and specific loss functions tailored to tackle the challenges of the long-tail problem, such as distribution balanced (DB) loss, label smoothing (LS) loss, and focal loss. As shown in Table 3, the BCE loss exhibited the best performance compared to the other loss functions, and we adopted the BCE loss for our task.

Loss function	Val. mAP
BCE	**0.3198**
WBCE	0.2661
DB	0.2875
LS	0.3031
Focal	0.3014

Table 3. **Loss Function.** Comparison of loss functions derived from binary cross-entropy.

3.4. Input Size Selection

We determined the most appropriate input size through experiments, where three different input sizes were considered: 224×224, 448×448, and 640×640. To set the size of the images, two different techniques were applied: *resize* and *random resized crop*. According to Table 4, the input size of 448×448 with *random resized crop* during the training phase lead to the best performance. Note that during inference, *center crop* was employed instead of *random resized crop*.

Input size	Val. mAP
Rz (224×224)	0.2984
RRC (224×224)	0.3400
Rz (448×448)	0.3198
RRC (448×448)	**0.3454**
Rz (640×640)	0.3053
RRC (640×640)	0.3307

Table 4. **Input size Selection.** Comparison of different input sizes by employing *resize* (Rz) or *random resized crop* (RRC) during the training phase.

Optimizer	Learning rate	(β_1, β_2)	Val. mAP	Dev. mAP
AdamW	4e-3	(0.9, 0.999)	0.3454	0.3093
Lion	4e-4	(0.9, 0.99)	0.3549	0.3322
Lion	1e-5	(0.9, 0.99)	0.3691	0.3364
Lion	4e-5	(0.9, 0.99)	0.3633	**0.3419**
Lion	4e-6	(0.9, 0.99)	0.3652	0.3374
Lion	4e-5	(0.95, 0.98)	0.3693	0.3324

Table 5. **Optimizer.** Comparison between the AdamW and Lion optimizers. The hyperparameters for the Lion optimizer have been tuned and optimized.

3.5. Optimizer

The AdamW optimizer [14] sets learning rates adaptively, which are adjusted based on the gradient magnitude. Such a strategy makes it suitable for a wide range of problems and architectures. On the other hand, the Lion optimizer [8] is more memory-efficient than AdamW as it only keeps the track of the momentum. To compare these two optimizers, the performance was calculated not only on the validation set but also on the development set. As presented in Table 5, the Lion optimizer outperformed the AdamW optimizer in most cases by about 3% points in the development set. Hence, we focused on tuning the hyperparameters of the Lion optimizer. As a result, the best performance was achieved with a learning rate of 4×10^{-5}, where β_1 and β_2 were set to 0.9 and 0.99, respectively.

3.6. Ensemble Strategy

To mitigate the class imbalance issue, separate models for each category were trained, namely the HEAD and TAIL models. Initially, we merged these two specialized models by concatenating their prediction scores for each class and averaging the prediction scores of "Support Devices". However, this strategy was not particularly good compared to ALL model, which was trained with both head and tail categories together, as shown in Table 6. The reason lies in the nature of multi-label classification; the HEAD and TAIL models tend to focus on their respective categories, resulting in a failure to capture the intricate relationships between these two groups.

Ensemble	Models	Strategy	Dev. mAP
	ALL	-	0.3419
✓	HEAD, TAIL	Merge	0.3332
✓	ALL, HEAD, TAIL	Max	0.3420
✓	ALL, HEAD, TAIL	Average	**0.3426**

Table 6. **Ensemble strategy.** Comparison of different ensemble strategies.

Data augmentation	Dev. mAP
F, RRC	**0.3459**
F, RRC, R	0.3450
F, RRC, GC	0.3442
F, RRC, R, GC	**0.3485**

Table 7. **Data augmentation.** Comparison of different combinations of data augmentation techniques for training.

To overcome this drawback and improve overall performance, we developed two additional ensemble strategies that integrate the ALL, HEAD, and TAIL models. The "Max" strategy selects the highest prediction value for each class among the three models, and the "Average" strategy computes the average of the prediction scores of overlapping classes in these three models. These strategies allowed us to leverage the unique characteristics of each category while simultaneously capturing the interrelationship between different classes.

According to Table 6, the "Average" strategy is the best among the three strategies, thereby being adopted as the final ensemble strategy. Averaging predictions from multiple models mitigates variance introduced by individual models' errors, ensuring stability and consistency of predictions. It also balances the importance of predictions across classes, enhancing ensemble robustness by compensating for noisy model predictions and improving overall reliability. The prediction value for class c of the ensemble model is defined as

$$e_c := \begin{cases} \frac{1}{3}(a_c + h_c + t_c), & c \in \mathcal{S} \\ \frac{1}{2}(a_c + h_c), & c \in \mathcal{H} - \mathcal{S} \\ \frac{1}{2}(a_c + t_c), & c \in \mathcal{T} - \mathcal{S} \end{cases}, \qquad (1)$$

where a_c, h_c, and t_c denote the prediction values of class c given by ALL, HEAD, and TAIL models. $\mathcal{S}$ is a set consisting of a single element "Support Devices", and $\mathcal{H}$ and $\mathcal{T}$ are sets of head and tail categories, respectively.

3.7. Pre-training on NIH CXR

Drawing insights from [9], which demonstrated that when transfer learning to large X-ray targets, the models pre-trained on a large-scale natural image dataset such

Submission	Data augmentation	Models	Pre-trained	Test mAP	Test mAUC	Test mF1
Test_S1	F, RRC	ALL	ImageNet-21K	0.3506	0.8361	0.2622
Test_S2	F, RRC, R, GC	ALL	ImageNet-21K	0.3513	0.8366	0.2608
Test_E1	F, RRC	ALL, HEAD, TAIL	ImageNet-21K	0.3520	0.8364	0.2630
Test_E2	F, RRC, R, GC	ALL, HEAD, TAIL	ImageNet-21K	0.3535	**0.8370**	0.2623
Test_Final	F, RRC, R, GC	ALL, HEAD, TAIL	NIH CXR	**0.3537**	0.8361	**0.2656**

Table 8. **Summary of experimental details and results of our final submissions.** The test mAP, mAUC, and mF1 scores for our final submissions to the CXR-LT competition, based on different experimental setting combinations.

Category	Class	Test_S1	Test_S2	Test_E1	Test_E2	Test_Final
	Support Devices	0.9079	0.9086	0.9088	0.9091	0.9064
HEAD	Lung Opacity	0.5966	0.5977	0.5962	0.5971	0.5971
	Cardiomegaly	0.6479	0.6511	0.6503	0.6529	0.6523
	Pleural Effusion	0.8276	0.8274	0.8280	0.8288	0.8288
	Atelectasis	0.6061	0.6070	0.6083	0.6087	0.6087
	Pneumonia	0.3060	0.3064	0.3050	0.3058	0.3054
	No Finding	0.4734	0.4727	0.4756	0.4769	0.4782
	Edema	0.5511	0.5511	0.5527	0.5534	0.5530
	Enlarged Cardiomediastinum	0.1864	0.1858	0.1849	0.1857	0.1844
TAIL	Consolidation	0.2274	0.2297	0.2270	0.2286	0.2278
	Pneumothorax	0.5285	0.5310	0.5313	0.5349	0.5330
	Fracture	0.2549	0.2585	0.2634	0.2655	0.2624
	Infiltration	0.0593	0.0585	0.0586	0.0579	0.0573
	Nodule	0.1859	0.1865	0.1875	0.1889	0.1920
	Mass	0.2201	0.2250	0.2187	0.2247	0.2244
	Calcification of the Aorta	0.1386	0.1403	0.1405	0.1412	0.1403
	Emphysema	0.1901	0.1932	0.1899	0.1920	0.1926
	Hernia	0.5634	0.5802	0.5763	0.5866	0.5853
	Tortuous Aorta	0.0619	0.0615	0.0631	0.0627	0.0613
	Pleural Thickening	0.1051	0.1102	0.1043	0.1070	0.1083
	Lung Lesion	0.0403	0.0422	0.0413	0.0422	0.0419
	Subcutaneous Emphysema	0.5527	0.5434	0.5560	0.5512	0.5557
	Fibrosis	0.1674	0.1486	0.1621	0.1528	0.1628
	Pneumomediastinum	0.3715	0.3715	0.3724	0.3844	0.3843
	Pleural Other	0.0412	0.0347	0.0401	0.0343	0.0372
	Pneumoperitoneum	0.3044	0.3104	0.3094	0.3190	0.3165
	Test mAP	0.3506	0.3513	0.3520	0.3535	**0.3537**

Table 9. **Test mAP scores of our submissions from each class.** The classes are listed in descending order from head to tail category. “Support Devices” is on the top since it is the most occurrences and belongs in both categories.

as ImageNet-21K achieve comparable or superior performance to the networks pre-trained on a large-scale medical X-ray dataset, we aim to enhance performance through pre-training on the NIH ChestXRay14 (NIH CXR) [20] dataset. The 14 lung disease classes in NIH CXR are identical to some of the classes in MIMIC-CXR-LT, and for training convenience, we used only six of them, based on descending class frequency values. The 86,524 images for training and 25,596 images for testing were used on the chosen architecture.

4. Experiments

This section describes the implementation details of our model including experimental setup and data augmentation strategy and analyzes our results.

4.1. Implementation details

Experimental setup We train with both the train and validation sets, 264,849 images in total, to find the best model. Based on the observations discussed in the previous section, the pretrained model was trained using the binary cross-entropy loss and the Lion optimizer, with an input size of 448×448. Additionally, for learning an ensemble model, the “Average” strategy was adopted. All models were trained for 30 epochs on a single NVIDIA A100 80GB PCIe GPU.

Data augmentation To investigate the effectiveness of data augmentation, we tested several data augmentation strategies, which include *horizontal flip* (F), *random resized crop* (RRC), *rotation* (R), and *gamma correction* (GC). According to Table 7, the use of F, RRC, or F, RRC, R, GC achieves the best performances on the development set. Our model adopted these two combinations for data augmentation for the final submission.

4.2. Results

We made a total of five submissions for the final evaluation under the CXR-LT competition policy. Comprehensive details and corresponding scores for each submission are provided in Table 8. While Test_S1 and Test_S2 were executed using single models with each specific data augmentation method as previously described, the ensemble approach was employed for Test_E1, Test_E2, and Test_Final. Notably, Test_Final, our proposed framework, was pre-trained using the NIH CXR dataset, unlike other submissions pretrained on ImageNet-21K. The experimental results demonstrate that Test_Final achieves the best performance in our target evaluation metrics in general. Although the mAUC of Test_Final is not the highest, it is still comparable to the other submissions.

Table 9 lists the mAP scores for individual classes in the head and tail categories.s. Class-wise AP from the ensemble model sets, including Test_E1, Test_E2, and Test_Final, tends to improve compared to the sets with a single model, including Test_S1 and Test_S2. Although Test_E2 shows the highest class-wise AP scores for most classes, Test_Final attains the best mAP by addressing the challenges of the long-tail distribution and achieving a well-balanced class-wise AP.

5. Conclusion and Future work

We explored an optimized ensemble framework for multi-label long-tailed classification on chest X-rays in this paper. To address this challenging task, we examined crucial components including model architectures, loss functions, optimizers, and data augmentation strategies. Moreover, an ensemble approach was adopted to tackle long-tailed distribution. Our final submission pretrained on NIH CXR demonstrated competitive performance on the MIMIC-CXR-LT dataset.

One limitation of our approach is that our model relies on the NIH CXR dataset with full supervision. Unsupervised pretraining methods could allow us to learn better representative features from large-scale datasets, potentially leading to further performance enhancements and improved generalization capabilities.

References

[1] Hanan S. Alghamdi, Ghada Amoudi, Salma Elhag, Kawther Saeedi, and Jomanah Nasser. Deep learning approaches for detecting covid-19 from chest x-ray images: A survey. *IEEE Access*, 9:20235–20254, 2021.

[2] Sanhita Basu, Sushmita Mitra, and Nilanjan Saha. Deep learning for screening covid-19 using chest x-ray images. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 2521–2527. IEEE, 2020.

[3] Abhijit Bhattacharyya, Divyanshu Bhaik, Sunil Kumar, Prayas Thakur, Rahul Sharma, and Ram Bilas Pachori. A deep learning based approach for automatic detection of covid-19 cases using chest x-ray images. *Biomedical Signal Processing and Control*, 71:103182, 2022.

[4] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021.

[5] Bingzhi Chen, Jinxing Li, Guangming Lu, Hongbing Yu, and David Zhang. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE journal of biomedical and health informatics*, 24(8):2292–2302, 2020.

[6] Bingzhi Chen, Zheng Zhang, Yingjian Li, Guangming Lu, and David Zhang. Multi-label chest x-ray image classification via semantic similarity graph embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2455–2468, 2022.

[7] Haomin Chen, Shun Miao, Daguang Xu, Gregory D Hager, and Adam P Harrison. Deep hierarchical multi-label classification of chest x-ray images. In *International conference on medical imaging with deep learning*, pages 109–120. PMLR, 2019.

[8] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023.

[9] Mehdi Cherti and Jenia Jitsev. Effect of pre-training scale on intra-and inter-domain, full and few-shot transfer learning for natural and x-ray chest images. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2022.

[10] Qingji Guan, Zhuangzhuang Li, Jiayu Zhang, Yaping Huang, and Yao Zhao. Joint representation and classifier learning for long-tailed image classification. *Image and Vision Computing*, 137:104759, 2023.

[11] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C Shen, George Shih, Ronald M Summers, Yifan Peng, and Zhangyang Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 22–32. Springer, 2022.

[12] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of

labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[15] Dana Moukheiber, Saurabh Mahindre, Lama Moukheiber, Mira Moukheiber, Song Wang, Chunwei Ma, George Shih, Yifan Peng, and Mingchen Gao. Few-shot learning geometric ensemble for multi-label classification of chest x-rays. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 112–122. Springer, 2022.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[17] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[18] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021.

[19] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[20] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[21] Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. Does head label help for long-tailed multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14103–14111, 2021.

[22] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.

[23] Mengqi Yuan, Jinke Xu, and Zhongnian Li. Long tail multi-label learning. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 28–31, 2019.

[24] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 184–193, 2021.