Advanced Augmentation and Ensemble Approaches for Classifying Long-Tailed Multi-Label Chest X-Rays

Trong-Hieu Nguyen-Mau, Tuan-Luc Huynh, Thanh-Danh Le,
Hai-Dang Nguyen, and Minh-Triet Tran*
University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
Viet Nam National University, Ho Chi Minh City, Vietnam
20120081@student.hcmus.edu.vn, {htluc19,ltdanh19}@apcs.fitus.edu.vn,
nhdang@selab.hcmus.edu.vn, tmtriet@fit.hcmus.edu.vn

Abstract

Chest radiography is a common medical diagnostic procedure, often resulting in a long-tailed distribution of clinical findings. This challenges standard deep learning methods, which tend to favor more common classes and might miss less frequent but equally important ”tail” classes. Chest X-ray diagnoses represent a multi-label problem due to the potential for multiple simultaneous diseases in patients. In this paper, we propose straightforward yet highly effective techniques to address the long-tailed imbalance in chest X-ray datasets. We specifically utilize EfficientNetV2 and ConvNeXt as our primary architectures, allowing the image sizes to influence architectural decisions. To counter dataset imbalance, we employ various basic and advanced augmentations. Mosaic augmentation is applied, and we alter the method of obtaining the label to manage this multi-label classification problem. We leverage the Binary Focal Cross-Entropy loss function and deploy several ensemble strategies to boost performance. These include Stratified K-Fold cross-validation and Test Time Augmentation. Our proposed method demonstrated its effectiveness during the Development and Testing phases of the CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays competition. Our approach yields substantial results with an mAP of 0.354, securing a position within the top five.

1. Introduction

Diagnostic medical examinations often display long-tailed distribution patterns, notably in chest radiography. While some diseases appear frequently, most cases are rare, resulting in a class imbalance [11]. Although several strategies exist to handle this, recent focus has been on applying these methods to address the complexity of long-tailed medical image recognition [39, 21]. Deep long-tailed learning, one of the most formidable problems in visual recognition, aspires to train efficient deep models from a plethora of images following a long-tailed class distribution [38].

When diagnosing chest X-rays (CXRs), the problem becomes multi-label as patients often show multiple disease findings simultaneously. Intriguingly, only a few studies have included label co-occurrence knowledge in the learning process [5, 4]. Considering the long-tailed class distribution in chest X-ray datasets, incorporating label co-occurrence information may offer valuable insights for addressing imbalanced and infrequent disease categories in this complex medical imaging task.

The widespread use of large-scale image classification benchmarks in the medical sector, typically consisting of single-label images with balanced label distributions, forms a disparity between conventional deep learning methods and the challenges associated with long-tailed, multi-label tasks like disease diagnosis in CXRs. Consequently, traditional approaches need help tackling the inherent complexities of class imbalance and label co-occurrence prevalent in such tasks [11].

Addressing this issue, Wang *et al.* developed a specialized benchmark specifically for long-tailed, multi-label medical image classification [2, 1]. Notably, previous attempts to introduce supplementary classes to MIMIC-CXR-JPG were made by Holste *et al.* [11] and Moukheiber *et al.* [22], who investigated long-tailed learning techniques and ensemble methods for few-shot learning on CXRs.

Our study diverges from the coarse-to-fine grouping approach in multi-label long-tailed chest X-ray classification. Instead, it directly classifies even the small categories without merging similar classes. The proposed strategy aims to provide a more detailed and precise classification scheme by maintaining separate classes for each disease, enhancing the differentiation between distinct categories, especially for rare diseases.

Our solution is in the Top 5 of the Development and Testing phases of the CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays competition [1] with the mAP score of 0.354, compared to the distribution of 0.293 ± 0.063 in the challenge. In summary, our contributions mainly include the following:

- We deliberately select architectural backbones, leveraging EfficientNetV2 [29] and ConvNeXt [18], to achieve impressive performance on datasets with a long-tailed distribution.
- Upon examining image sizes, we discover that larger images deliver superior performance. Therefore, we aim to exploit the largest possible resolution of images that a backbone network can support to preserve as much useful information as possible.
- In addressing the issue of imbalanced datasets, we employ an array of augmentations ranging from basic to advanced. Notably, we incorporate Mosaic augmentation [31] into this multi-label classification problem, significantly enhancing performance.
- We employ the Binary Focal Cross-Entropy loss function to handle long-tailed datasets.
- We use several ensemble strategies, including Stratified K-Fold cross-validation and Test Time Augmentation, to boost performance.

The content of this paper is organized as follows. In Section 2, we briefly review existing methods related to our work. Then we present our proposed method in Section 3. Experimental results are discussed in Section 4. Finally, we conclude our work and present open problems in Section 5.

2. Related Work

2.1. Multi-label Classification

In multi-label classification for chest X-ray images using deep learning approaches, several influential works have advanced the field. CheXNet [23] made a significant contribution by employing a DenseNet [12] trained on the ChestX-ray14 dataset, achieving impressive AUC results. Subsequent research has focused on enhancing model performance and addressing specific challenges, such as Ma *et al.* [20], who introduced a squeeze-and-excitation (SE) block and global/local attention modules to capture disease-specific features while also adopting a two-stage training method for handling class imbalance. Recently, [25] investigated state-of-the-art classifiers, revealing extensive bias patterns with potential vulnerabilities in real-world deployments, and proposed using multi-source datasets to mitigate such issues during data collection.

Moreover, transformer-based architectures have gained prominence, with promising results achieved by pre-training Vision Transformers (ViTs) [8] on a large chest X-ray dataset using Masked Autoencoders (MAE) [9] for reconstructing missing pixels, demonstrating comparable or superior performance to state-of-the-art CNNs in multi-label thorax disease classification. Additionally, an impressive approach involved a multi-label classification model based on the Swin Transformer backbone [17], incorporating a Multi-Layer Perceptron (MLP) head architecture. This model achieved state-of-the-art performance on the Chest X-ray14 dataset, with an average AUC score of 0.810.

2.2. Long-Tailed Data Distribution

Efficiently addressing long-tailed data distributions has been an active area of research, with various methodologies like OLTR [19] proposed to handle imbalanced class frequencies. Zhang *et al.* [36] introduce MBNM, a Multi-Branch Network based on Memory Features, for long-tailed medical image recognition in computer-aided diagnosis, utilizing three branches, including a tail learning branch with a feature memory module, to improve classification performance for rare diseases in imbalanced medical datasets. Additionally, efforts to address Partial Long-Tailed Multi-Label Classification (PLT-MLC)[37] explore re-weighting strategies in Knowledge-Contrastive Learning (KCL) [30] to mitigate false negatives. Yang *et al.* [32] use contrastive learning, category prototypes, and a prototype recalibration strategy in a single-stage pipeline, achieving superior performance on medical image datasets by enhancing feature representation and addressing imbalanced data distribution effectively.

2.3. Class-balanced Losses

Class imbalance presents a challenge across various classification tasks, compelling researchers to explore effective techniques to mitigate its impact. One notable approach is the introduction of focal loss by Lin *et al.* [16], tailored explicitly for dense object detection. Focal loss dynamically re-weights the contribution of hard-to-classify examples, leading to significant performance improvements, especially for underrepresented minority classes.

Another noteworthy method, proposed by Shu *et al.* [26], centers around a mapping function that explicitly uses sample weights. This technique effectively addresses class imbalance issues, yielding substantial performance boosts for deep neural networks. Researchers have also explored novel propositions like that of Li *et al.* [15], who introduced a parametric cross-entropy loss function with individualized data augmentation. This integrated approach not only enhances the efficacy of handling class imbalance but also exhibits versatility in diverse classification scenarios.

Chen *et al.* [6] propose class-center triplet loss, which addresses imbalanced training data in medical image diagnosis. The framework effectively separates class distributions and promotes compact class representations.

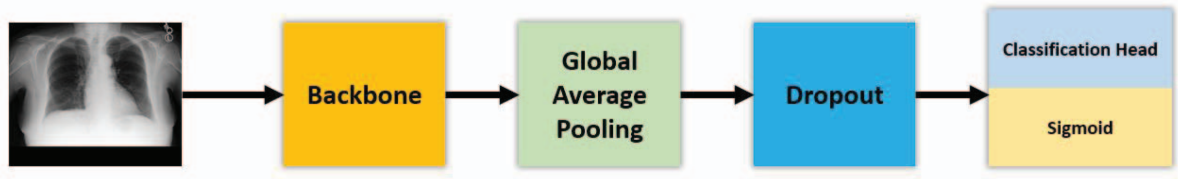


Figure 1: Overall architecture of our proposed method, a multi-label long-tailed classifier for chest X-rays.

3. Proposed Method

In this section, we elaborate on our proposed methods for effectively training a model when the training samples follow a long-tailed distribution. The proposed strategy in this study avoids the coarse-to-fine approach, which groups similar classes, and instead aims to directly classify even the small categories in multi-label long-tailed chest X-ray classification. The approach maintains a more detailed classification scheme by treating each disease as a separate class, ensuring accurate differentiation between distinct categories, including rare diseases. The study introduces highly effective techniques that achieved competitive results in the CXR-LT competition's Development and Testing phases. These techniques leverage EfficientNetV2 and ConvNeXt architectures, utilize advanced augmentations, implement Mosaic augmentation for multi-label classification, and deploy ensemble strategies like Stratified K-Fold cross-validation and Test Time Augmentation.

3.1. Architecture

A straightforward approach is utilized for training the multi-label classifier. We construct models that adhere to the architecture described in Figure 1 to learn representations from naturally distributed data. The difference lies in the choice of backbone architectures. Empirically, we have chosen two different models for the backbone: EfficientNetV2 [29] and ConvNeXt [18]. We thereby briefly introduce these two backbones.

EfficientNetV2, an extension of the renowned EfficientNet model [28], stands out with competitive performance despite its compact size. Its faster training speed and superior parameter efficiency, achieved through training-aware neural architecture search and scaling, make it an ideal choice for efficiently handling computationally expensive large input image sizes. This core idea has been pivotal in achieving our final competitive results. Notably, the efficiency of EfficientNetV2 proves particularly valuable when dealing with the substantial CXRs (chest X-rays) dataset [11], demanding significant time and resources for training. Moreover, EfficientNetV2 has excelled in image classification tasks [29]. It has recently been utilized for multi-label classification with high model scores [24], further reinforcing its potential as a feature extractor backbone for multi-label classification.

Since the inception of the Vision Transformer (ViT) [8] and Swin Transformer [17], transformer-based architectures have emerged as the dominant paradigm, propelling numerous models to achieve state-of-the-art results in various computer vision tasks, notably, by re-evaluating design spaces and pushing the capabilities of pure CNNs, incrementally enhancing a standard ResNet [10] towards the vision transformer design. This research introduced a series of pure CNN models known as ConvNeXt [18]. With ConvNeXt's compelling performance across various computer vision tasks, we recognize its potential for generalization in multi-label settings, particularly in addressing the challenges of long-tailed distributions. Our experiment demonstrates the efficacy of ConvNeXt in handling such settings, reaffirming its promise as a compelling choice for multi-label long-tailed medical image classification, such as chest X-ray images [11].

The multi-label classification process starts with the backbone extracting features from the input images. Global Average Pooling is then applied to summarize the spatial information, reducing the data to a fixed-size vector. Dropout [27] is also implemented as a network-level regularization technique, randomly deactivating neurons during training to prevent overfitting and improve generalization. Finally, the processed data is sent to the classification head, where the sigmoid function determines the independent probabilities for each class.

3.2. Large Area and Sufficient Information

In our experiments, we notice a significant variation in the aspect ratio between image width and height, ranging from approximately 0.26 to 2.72 (with a standard deviation of about 0.15). This diversity in aspect ratio could introduce inefficiencies during model training, as image sizes must be standardized during the preprocessing stage. Figure 2 provides a visual representation of the image size distribution within the training set.

During our investigation, we observe that the original images are relatively large, requiring resizing to smaller dimensions for training and inference. However, this downsizing adversely affects performance. Notably, as we increase the image size, a significant improvement in performance is evident.

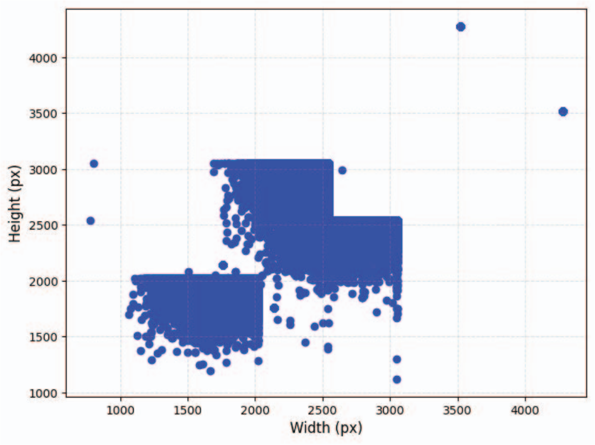


Figure 2: Scatter plot illustrating the distribution of image width (x-axis) and height (y-axis) in the training set.

Downsizing images may lead to the loss of crucial features, particularly in long-tailed classes with fewer samples, making it challenging for models to discern patterns and features effectively. Consequently, we conclude that utilizing the largest feasible image size (within hardware resource constraints) substantially enhances performance.

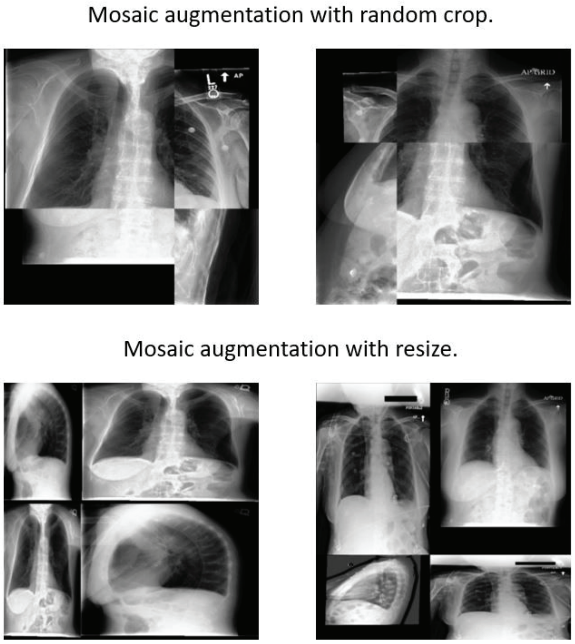
3.3. Data Augmentation

Data augmentation is crucial in model training, enhancing the model's generalization and minimizing overfitting. This benefits imbalanced datasets, improving the model's performance on underrepresented classes. Our augmentation pipeline includes various techniques such as random scaling, horizontal flipping, rotation, and random contrast, saturation, and brightness adjustments. Additionally, we integrate advanced augmentation methods like CutOut [7], Mixup [34], and Cutmix [33].

Inspired by Mosaic augmentation [31, 3], we adapt it for classification tasks. This technique combines four images to create a single composite image, enabling the model to identify objects across diverse contexts, and promoting resilience to specific context dependencies. This improves the model's performance, especially in long-tailed datasets, by enhancing learning for rare classes and diversifying small sample classes in various contexts.

We introduce two variations of the Mosaic method, both using four images from the training set. The first variation involves random cropping, providing dynamic and generalized contexts, while the second uses full resizing, preserving the maximum amount of label-related information. These two types of Mosaic augmentation [31, 3] are visually illustrated in Figure 3.

A significant challenge is determining how to handle labels following the application of Mosaic augmentation, as in [31, 3]. A typical approach might simply average the labels of all four input images. However, our observations indicate that this is ineffective. Instead, we propose an alternative method that is both straightforward and efficient.


Mosaic augmentation with random crop.

Mosaic augmentation with resize.

Figure 3: Visualizations of two types of Mosaic augmentation applied for chest X-rays images.

Let $D = (x_1, y_1), \ldots, (x_N, y_N)$ represent the dataset, where N denotes the number of training samples. Each sample-label pair (x_k, y_k), with $k \in \{1, \ldots, N\}$, consists of an input sample x_k and its corresponding multi-label vector y_k. Here, $y_k = [y_k^1, \ldots, y_k^C] \in \{0, 1\}^C$, with C being the number of classes. The vector y_k contains binary values, where y_k^c indicates the presence (1) or absence (0) of class c in the sample x_k. Assuming the Mosaic augmentation takes four sample-label pairs as input: $(x_a, y_a), (x_b, y_b), (x_c, y_c), (x_d, y_d)$, the corresponding new sample-label pair is calculated as follows:

$$x_{\text{result}} = \text{MOSAIC}(x_a, x_b, x_c, x_d)$$
$$y_{\text{result}} = \text{CLIP\_VALUE}(y_a + y_b + y_c + y_d)$$

CLIP_VALUE function is responsible for constraining all input values in $[0, 1]$. This ensures that the output label generated after the Mosaic augmentation maintains the presence of classes in the input images. By adhering to this requirement, the model is motivated to learn patterns and features from either part or the entirety of the input image, enabling it to discern the corresponding class accurately. As a result of this constraint, we observe a notable improvement in the model's performance. This modification in label calculation has been applied to other augmentation techniques, such as Mixup [34] and Cutmix [33], further enhancing the model's ability to generalize and achieve robust performance in diverse scenarios.

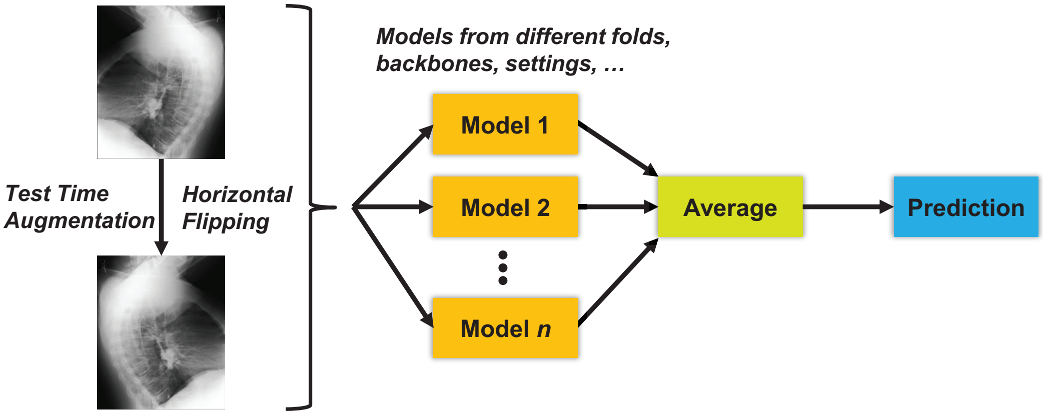


Figure 4: Ensemble strategy in our method.

3.4. Loss Function

In our approach, we leverage the Binary Focal Cross-Entropy loss function, which has proven effective in dealing with imbalanced datasets commonly encountered in computer vision tasks. While Binary Cross-Entropy is suitable for binary and multi-label classification, it treats all samples equally in the loss computation, making it less favorable for datasets with long-tailed distributions.

Focal loss [16], on the other hand, addresses this issue by introducing a modulating factor that down-weights the contribution of easily classified examples to the overall loss. By doing so, the model focuses more on hard negatives, which are implicitly amplified by this operation. The formulation of the Focal loss is given as follows:

$$\mathrm{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

where p_t represents the predicted probability for the target class, α_t is the class-specific balancing parameter, and γ is the focusing parameter. This aspect is particularly beneficial when dealing with imbalanced datasets, as it prevents the model from favoring the majority class and instead encourages learning from less common, challenging instances. We use $\alpha_t = 0.25$ and the default $\gamma = 2.0$ as described in the original paper.

To further address the issue of class imbalance in binary classification, we apply a weight balancing technique. This technique ensures that our model gives equal attention to both classes, regardless of their prevalence in the dataset. Additionally, we implement label smoothing with a parameter of 10^{-2}, thereby improving the model's capacity to generalize and enabling it to make accurate predictions on unseen data even when the class distribution is skewed.

3.5. Ensemble Approaches

As shown in Figure 4, our ensemble strategy utilizes Stratified K-Fold cross-validation and Test Time Augmentation, presented in the following two sections.

3.5.1 Stratified K-Fold Cross-Validation

To address the challenges posed by the long-tailed distribution in our dataset and the multi-label classification task, we employ Stratified K-Fold cross-validation with K set to 5. This approach involves dividing the dataset into five subsets or 'folds', ensuring that each fold maintains the same data distribution, including representation of all classes, especially the scarce ones. Such a balanced representation is unachievable through conventional random shuffling.

The benefits of this approach are manifold. Firstly, it significantly minimizes underfitting by using most data for training. Secondly, it also curbs overfitting, as a substantial portion of the data is used for the validation set. Moreover, this method is advantageous in scenarios with unbalanced datasets, as multiple folds help maintain a representative sample of the original data. Analyzing the model performance for each fold offers insights that allow us to fine-tune the model further and can even be employed for hyperparameter tuning. Ultimately, we can achieve reliable, generalizable, and highly accurate predictions by averaging the ensemble of models from all the folds.

3.5.2 Test Time Augmentation

Test Time Augmentation (TTA) is a strategy for applying various transformations to a test image. These transformed images are then fed into the trained model. By averaging the results, a more confident prediction is achieved.

In our experiments, we confine our transformations to horizontal flipping. This modification does not dramatically change the input data but provides a unique perspective on it. The technique aids in enhancing the model's predictive accuracy by offering a more robust estimate while preserving contextual information. It acknowledges potential variations in the test data that could assist the model in making more precise predictions, thus improving its generalizability.

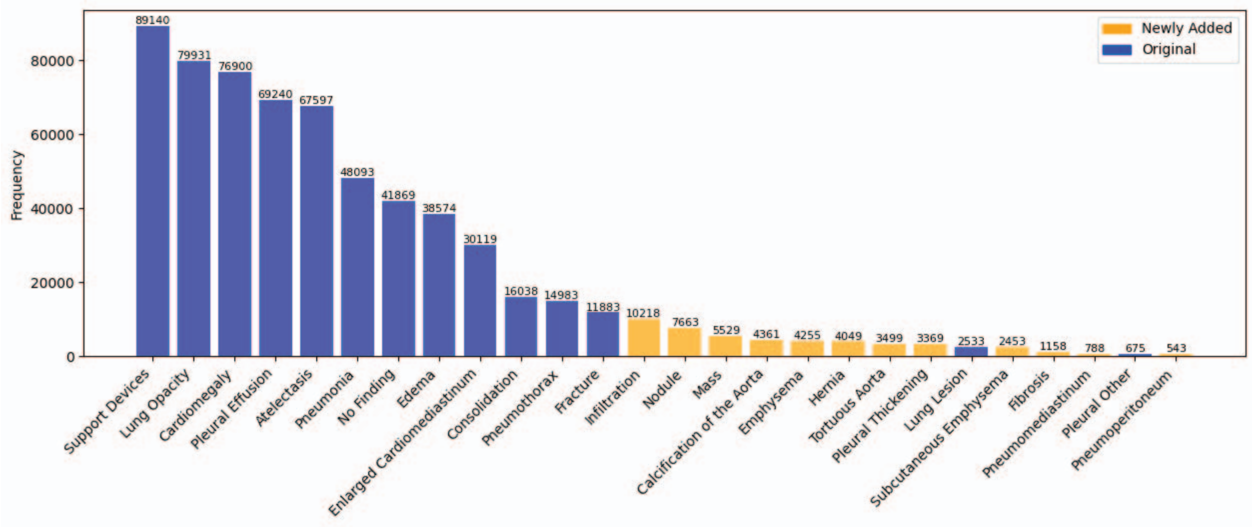


Figure 5: Class distribution in the training set of the expanded MIMIC-CXR-JPG dataset. Yellow bars indicate newly added classes, many of which are scarce in occurrence.

4. Experiments

4.1. Data Overview

MIMIC-CXR-JPG [13] is an extensive publicly available database with labeled chest radiographs [1]. The dataset was created by converting DICOM-format chest radiographs from the original MIMIC-CXR into JPEG format, facilitating easier use for researchers without specialized knowledge in the medical domain. In total, MIMIC-CXR-JPG comprises 377,110 chest X-rays associated with 227,835 imaging studies.

The expanded dataset in this shared task introduces 12 new pathologies based on radiology reports from the original dataset. This expansion results in a total of 26 classes, creating a long-tailed class distribution, as shown in Figure 5. This work follows the procedure of Holste *et al.*[11] and Moukheiber *et al.*[22] to establish a benchmark for long-tailed, multi-label medical image classification. The dataset is divided into three subsets: training, development, and testing, with a split ratio of 7:1:2, respectively.

The dataset is organized into levels of patients, studies, and images. This hierarchical structure enables researchers to access and analyze medical records from multiple patients, where each patient has one or more studies, and each study has one or more images. Importantly, all images within the same study have the same label, which allows for the aggregation of information from the images to arrive at a final decision.

Phase	#Images
Train	264,849
Development	36,769
Test	75,492

Table 1: Number of images for each phase of CXR-LT competition.

4.2. Implementation Details

All our models are implemented using Tensorflow's Keras. Depending on the configuration, we resize input data to either 224×224, 512×512, or 768×768, subject to the mentioned augmentation procedure, and scaled to a range between 0 and 1.

We employ the Adam optimization algorithm [14], setting the momentum parameters at $\beta_1 = 0.9$ and $\beta_2 = 0.999$, with a learning rate of 10^{-4}. Each model undergoes training for 20 epochs on four NVIDIA Tesla V100 32GB GPUs.

4.3. Evaluation Metrics

The shared task shows that the difficulty in evaluating multi-label classification can stem from a significant class imbalance in the data, which requires proper matrices to assess the result appropriately [11] appropriately.

The mean Average Precision (mAP) is a suitable metric to meet the requirements as it can measure performance across decision thresholds and be resilient to class imbalance. This metric provides a "macro-averaged" mAP across the 26 classes, offering a comprehensive view of the model's performance across all categories.

[1] https://physionet.org/content/mimic-cxr-jpg/2.0.0/

Index	Backbone	Image Size	K-Fold	TTA	Development			Testing		
					mAP	mAUC	mF1	mAP	mAUC	mF1
(1)	EfficientNetV2S	224			0.300	0.805	0.127	-	-	-
(2)	EfficientNetV2S	512			0.322	0.819	0.147	-	-	-
(3)	EfficientNetV2S	768			0.324	0.819	0.146	-	-	-
(4)	EfficientNetV2S	768		x	0.329	0.823	**0.166**	-	-	-
(5)	EfficientNetV2S	768	x	x	**0.347**	**0.834**	0.150	-	-	-
(6)	ConvNeXtTiny	224	x	x	0.314	0.815	0.123	-	-	-
(7)	ConvNeXtSmall	512	x	x	-	-	-	0.343	0.830	0.147
(8)	(5) + (6)	-	-	-	0.344	0.832	0.127	0.347	0.836	**0.153**
(9)	(5) + (7)	-	-	-	-	-	-	**0.354**	**0.838**	0.148
(10)	(5) + (6) + (7)	-	-	-	-	-	-	0.351	0.837	0.137

Table 2: Benchmark results for the Development and Testing phases of the CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays competition [1]. The **best** results are highlighted.

Another metric often used in research for this dataset is the Area Under the Receiver Operating Characteristic Curve (AUC) [22, 35]. However, AUC can be heavily inflated under such conditions, leading to overestimating the model's performance. Therefore, its role in this context is limited and does not contribute significantly to addressing the root problem compared to mAP. In addition, mean AUC (mAUC) and mean F1 score (mF1) are included as supplementary evaluation metrics for other assessments. Still, they are not considered the main metrics for evaluating model performance in imbalanced data. The mAUC provides an average measure of the model's ability to distinguish between classes, while the mF1 score, calculated using a decision threshold of 0.5 for each class, provides a balance between precision and recall.

4.4. Quantitative Results

In the Development phase, we begin with the baseline model (1) using the EfficientNetV2S backbone and an image size of 224 × 224. We then progressively increase the image size to 512 × 512, resulting in significant improvements of around 2% for mAP and mF1, and 1.4% for mAUC. However, as the image size grows, the performance gains begin to saturate, and we settle on an image size of 768 × 768. Hence, experiments (1), (2), and (3) consistently demonstrate improved model performance as the image size scales from 224 × 224 to 768 × 768. Experiment (4) incorporates Test Time Augmentation (TTA), further enhancing metrics by 0.5%, 0.4%, and 2% in terms of mAP, mAUC, and mF1, respectively.

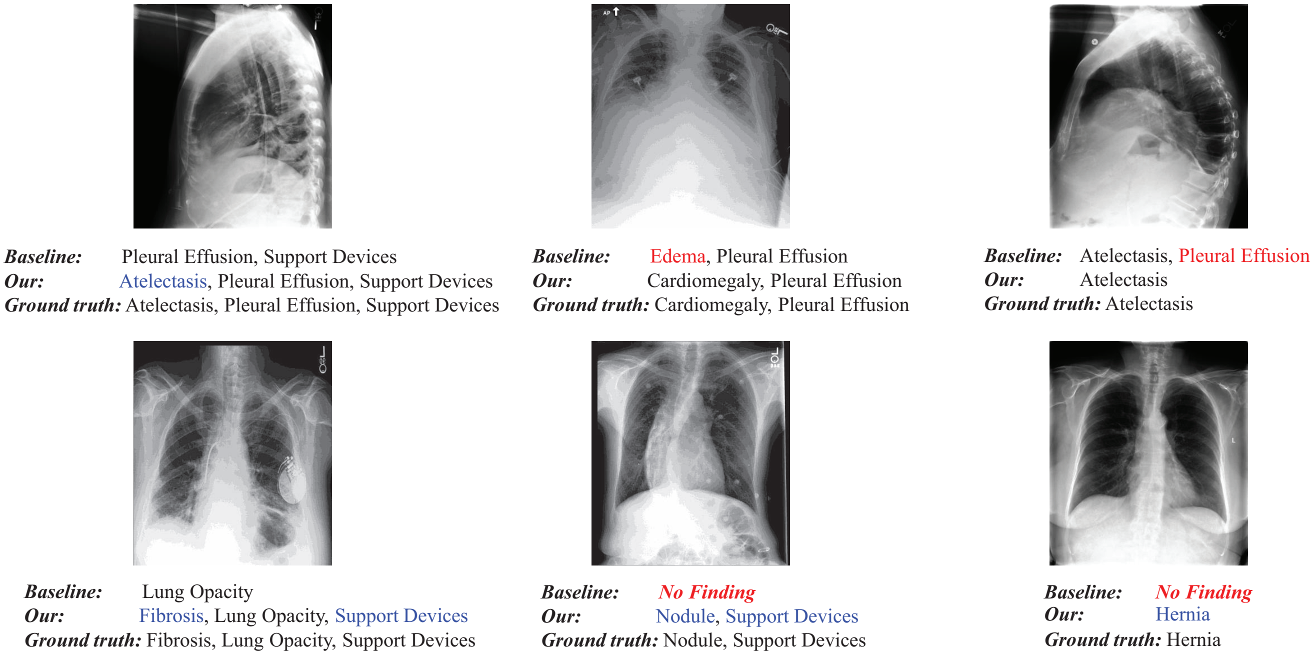
Implementing the Stratified K-Fold strategy results in a notable mAP improvement from 0.329 to 0.347 during the Development phase. Although there is a 1.6% drop in mF1, the effectiveness of our strategy becomes evident as it favors two of the three metrics.

In a similar approach to the baseline model, we experiment with the ConvNextTiny backbone, utilizing the stratified K-fold and TTA strategies. This approach is competitive compared to the EfficientNetV2S counterpart, achieving similar performance levels in terms of mAP and mAUC. However, we cannot provide the Development phase results of subsequent experiments on the ConvNext backbones, as this phase has expired.

In the Testing phase, we explore various ensemble methods to optimize performance. Among these ensembles, the most effective one, labeled as (8), combines experiments (5) and (7), achieving an impressive mAUC of 0.838, an mF1 of 0.148, and an mAP of 0.354. However, an exciting observation arises from the experiment (10), an ensemble of three models where two models share the same ConvNeXt architecture. Experiment (10) experiences a significant drop in mF1, suggesting that ensembles composed of similar models might slightly penalize performance due to architectural bias. This architectural bias could be the reason behind the lower performance in mF1 for ConvNeXt. To counteract this effect, introducing more models with different backbones, such as EfficientNetV2S, might help average the bias and enhance overall ensemble performance. This approach could mitigate the architectural bias and contribute to more balanced predictions, especially for rare classes, in the multi-label long-tailed classification of chest X-rays during the Testing phase.

These results underscore the efficacy of our approach, highlighting the crucial role of utilizing large images in medical imaging tasks, such as chest X-rays. By incorporating additional techniques like Stratified K-Fold Cross-validation and Test-Time Augmentation (TTA), we achieve competitive performance in the multi-label long-tailed classification of chest X-rays.

Out of the 17 teams competing in the CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays competition, our method achieves a competitive mAP score of 0.354, surpassing the test distribution mean (0.293) with a low standard deviation of 0.063 by a substantial margin.



Baseline: Pleural Effusion, Support Devices
Our: Atelectasis, Pleural Effusion, Support Devices
Ground truth: Atelectasis, Pleural Effusion, Support Devices

Baseline: Edema, Pleural Effusion
Our: Cardiomegaly, Pleural Effusion
Ground truth: Cardiomegaly, Pleural Effusion

Baseline: Atelectasis, Pleural Effusion
Our: Atelectasis
Ground truth: Atelectasis

Baseline: Lung Opacity
Our: Fibrosis, Lung Opacity, Support Devices
Ground truth: Fibrosis, Lung Opacity, Support Devices

Baseline: *No Finding*
Our: Nodule, Support Devices
Ground truth: Nodule, Support Devices

Baseline: *No Finding*
Our: Hernia
Ground truth: Hernia

Figure 6: Visualizing inference results of the baseline and our methods on Training set.

4.5. Qualitative Results

Figure 6 provides visualizations of the results obtained on the Training set using both the baseline and our final method, shedding light on the distinct performance characteristics of the two approaches. Missing or wrong baseline predictions are in red. More accurate prediction results of our method over baseline one are in blue.

Notably, the baseline tends to consistently predict the majority classes, such as Pleural Effusion and Support Devices, despite utilizing Focal Loss to mitigate the impact of imbalanced data. This tendency can be attributed to the challenges posed by small image sizes during preprocessing, which may result in the loss of crucial information unique to rare classes.

In contrast, our proposed method demonstrates enhanced accuracy in predicting labels that closely align with the ground truth. In the second row's leftmost case, our approach provides a comprehensive multi-label prediction for the chest X-ray image, capturing multiple pertinent disease indicators. Meanwhile, the baseline only outputs a single label, potentially overlooking crucial diagnostic insights.

Additionally, in the two cases on the right in the second row, the baseline classifies them as No Finding, which might be expected given the presence of two rare classes, namely Nodule and Hernia, in the ground truth. In contrast, our method showcases its robustness in providing correct labels even for these rare classes, thus improving the model's capability to handle long-tailed data distributions effectively.

5. Conclusion

Our study focuses on addressing the challenge of long-tailed data distribution in chest X-ray datasets, particularly in multi-label classification tasks for chest X-ray diagnoses. We leverage advanced models such as EfficientNetV2 and ConvNeXt to overcome this imbalance while incorporating various techniques to balance the data. Notably, image augmentations, including Mosaic augmentation tailored at the label level, are crucial in enhancing performance for this problem. Furthermore, adopting the Binary Focal Cross-Entropy loss function contributes to handling long-tailed datasets effectively. We integrate several ensemble strategies to boost model performance, such as Stratified K-Fold cross-validation and Test Time Augmentation.

The efficacy of our approach is validated in the CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays competition, where our team achieves a top-five ranking with the mAP score of 0.354 with reference to the distribution of 0.293 ± 0.063 in the challenge. Our contributions have the potential to significantly enhance diagnostic procedures in healthcare and offer valuable insights into addressing long-tailed imbalances in medical image datasets.

Future work includes exploring transfer learning from medical imaging domains, fine-tuning augmentations, and self-supervised learning. Emphasizing uncertainty estimation, model interpretability, and engaging medical experts would enhance performance and clinical relevance.

Acknowledgment: This research is funded by Viet Nam National University Ho Chi Minh City (VNU-HCM) under grant number DS2020-42-01.

References

[1] CXR-LT: Multi-label long-tailed classification on chest x-rays. `https://doi.org/10.13026/721s-vs37`,`https://codalab.lisn.upsaclay.fr/competitions/12599`.

[2] ICCV CVAMD 2023 Shared Task on Multi-Label Long-Tailed Classification on Chest X-Rays — bionlplab.github.io. `https://bionlplab.github.io/2023_ICCV_CVAMD/`. [Accessed 24-Jul-2023].

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[4] 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.

[5] 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.

[6] Kanghao Chen, Weixian Lei, Shen Zhao, Wei-Shi Zheng, and Ruixuan Wang. Pcct: Progressive class-center triplet loss for imbalanced medical image classification. *IEEE Journal of Biomedical and Health Informatics*, 27(4):2026–2036, 2023.

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[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] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[13] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems*, 34:3163–3177, 2021.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[18] 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.

[19] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

[20] Yanbo Ma, Qiuhao Zhou, Xuesong Chen, Haihua Lu, and Yong Zhao. Multi-attention network for thoracic disease classification and localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1378–1382. IEEE, 2019.

[21] Yassine Marrakchi, Osama Makansi, and Thomas Brox. Fighting class imbalance with contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 466–476. Springer, 2021.

[22] 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.

[23] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[24] Manuel Alejandro Rodríguez, Hasan AlMarzouqi, and Panos Liatsis. Multi-label retinal disease classification using transformers. *IEEE Journal of Biomedical and Health Informatics*, 2022.

[25] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion:

Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[26] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

[29] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[30] Tianqi Wang, Lei Chen, Xiaodan Zhu, Younghun Lee, and Jing Gao. Weighted contrastive learning with false negative control to help long-tailed product classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 574–580, 2023.

[31] Zhiwei Wei, Chenzhen Duan, Xinghao Song, Ye Tian, and Hongpeng Wang. Amrnet: Chips augmentation in aerial images object detection. *arXiv preprint arXiv:2009.07168*, 2020.

[32] Zhixiong Yang, Junwen Pan, Yanzhan Yang, Xiaozhou Shi, Hong-Yu Zhou, Zhicheng Zhang, and Cheng Bian. Proco: Prototype-aware contrastive learning for long-tailed medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–182. Springer, 2022.

[33] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[34] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[35] Ruru Zhang, Haihong E, Lifei Yuan, Jiawen He, Hongxing Zhang, Shengjuan Zhang, Yanhui Wang, Meina Song, and Lifei Wang. MBNM: Multi-branch network based on memory features for long-tailed medical image recognition. *Computer Methods and Programs in Biomedicine*, 212:106448, Nov. 2021.

[36] Ruru Zhang, E Haihong, Lifei Yuan, Jiawen He, Hongxing Zhang, Shengjuan Zhang, Yanhui Wang, Meina Song, and Lifei Wang. Mbnm: multi-branch network based on memory features for long-tailed medical image recognition. *Computer Methods and Programs in Biomedicine*, 212:106448, 2021.

[37] Wenqiao Zhang, Changshuo Liu, Lingze Zeng, Beng Chin Ooi, Siliang Tang, and Yueting Zhuang. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. *arXiv preprint arXiv:2304.10539*, 2023.

[38] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[39] Jiaxin Zhuang, Jiabin Cai, Ruixuan Wang, Jianguo Zhang, and Weishi Zheng. Care: Class attention to regions of lesion for classification on imbalanced data. In *International Conference on Medical Imaging with Deep Learning*, pages 588–597. PMLR, 2019.