Chest X-Ray Feature Pyramid Sum Model with Diseased Area Data Augmentation Method

Changhyun Kim*
SK Telecom
Seoul, South-Korea
changhyk@sktmedai.com

Giyeol Kim†
Gacheon University
Gyeonggido, South-Korea
rlduf422@gachon.ac.kr

Sooyoung Yang‡
ChungAng University
Seoul, South-Korea
jimmy1016@cau.ac.kr

Hyunsu Kim§
SungKyunKwan University
Gyeonggido, South-Korea
hyunsu517@g.skku.edu

Sangyool Lee¶
SK Telecom
Seoul, South-Korea
sangyoollee@sktmedai.com

Hansu Cho‖
SK Telecom
Seoul, South-Korea
hansu.cho@sktmedai.com

Abstract

Deep learning has shown considerable promise in medical image analysis, but significant challenges remain. These stem from the inherent complexities of medical images, such as varying sizes of lesions within the same image and the potential coexistence of multiple diseases. To address these issues, we propose a novel model combining TResNet with Feature Pyramid Network (FPN). This model adeptly handles multi-label classification, demonstrating robust performance across a range of lesion sizes. Furthermore, most medical images follow a long-tail distribution, presenting class imbalance problems, where the occurrence of one lesion often correlates with the presence of others. Considering these correlations, we introduced a strategy for dealing with the class imbalance issue by augmenting minority classes using bounding box information of the disease. Our proposed approach offers a novel solution for handling the unique challenges in deep learning-based medical image analysis, paving the way for more precise interpretations of complex medical images. The performance of mAP in 26 disease classes has been improved from 32.76% to 33.37% in a single model, and 35.11% in ensemble model.

1. Introduction

Tackling multi-label classification problems in the medical image domain holds paramount importance, as it enables accurate and comprehensive detection of multiple diseases within a single image. The ability to address multi-label classification in medical imaging has significant implications for improving diagnostic accuracy, treatment planning, and patient outcomes, making it a crucial area of scholarly pursuit.

Various research efforts have been undertaken to address the challenges posed by multi-label classification problems. Among the well-known approaches for tackling multi-label classification tasks, the binary relevance (BR) [5]transformation stands out. This method involves training independent classifiers for each label. However, it suffers from the limitation of overlooking label correlations as it handles each label independently [43]. Subsequently, to address this limitation, alternative techniques that explicitly consider the correlation among labels have been proposed, such as classifier chain or graph neural network-based approaches [44, 12, 17].

In multi-label classification tasks, another important aspect to consider in real-world applications is the presence of data with long-tailed class distributions [30, 33, 36, 25]. Long-tailed distribution refers to a situation where a few classes, known as the "head" classes, dominate the majority of the data, while many classes, known as the "tail" classes, have only a small amount of data. The class imbalance observed in long-tailed classification leads the models to be biased towards the abundant head classes, resulting in significantly lower performance for the tail classes [51, 8, 50, 7, 21].

Addressing the issue of long-tailed class imbalance has been the focus of numerous studies in recent years [8, 50, 28, 58, 59]. One common approach in long-tailed learning is the re-weighting method, where different training loss

*First and Corresponding author
†Second author contributed equally
‡Second author contributed equally
§Second author contributed equally
¶Third author
‖Third author

weights are assigned to each class to adjust the loss. This re-weighting minimizes the bias caused by class imbalance. Representative methods include Class-balanced Loss [14], Balanced Softmax [46], Focal Loss [32], and Vector Scaling Loss [39]. All of which have shown improvements in tail class performance.

Another widely used paradigm is information augmentation, which involves providing additional information during model training to enhance performance in long-tailed scenarios. This encompasses transfer learning and data augmentation methods. One previous study utilizing transfer learning involved training the model for representation learning on all long-tailed samples and then fine-tuning it on a more class-balanced subset [15]. This approach gradually transferred the learned features to the tail classes, ensuring balanced performance across all classes. On the other hand, data augmentation involves applying predefined transformations to the training data to increase both the quantity and quality of the data [40]. A related study, MiSLAS [61], in deep long-tailed learning, validated that data mixup as a form of data augmentation addressed the model's overconfidence issue, resulting in performance enhancement.

To address the challenge of long-tailed multi-label classification, we created synthetic data to augment the existing dataset. Leveraging this augmented data, we trained a novel model called Feature Pyramid Sum Model (FPSM). Instead of combining different scale feature map outputs such as bounding boxes or classification indexes, we firstly create the results of each of these feature maps. The difference is that the SUM process that combines them exists once more to give an ensemble effect. By adding this process, it is more suitable for multi-label classification than object detection. Through FPSM, our novel approach successfully captures features of diseases of various sizes and shapes.

Furthermore, proposed Diseased Area Data Augmention Method (DADAM) enhances the robustness of our methodology against class imbalance.The data augmentation technique primarily focused on classes that constitute the tail, given their propensity to induce class imbalance. Our data augmentation approach employs two distinct methods. The first involves utilizing existing normal images from the dataset as background images, from which disease patches are extracted from disease images and overlaid. The second method incorporates images containing classes that correspond to the tail as the background, onto which the disease patches are then superimposed. A notable consideration during the implementation of these methods was the interrelationship amongst the diseases. By primarily augmenting classes that constitute the tail through these techniques, not only is class imbalance mitigated, but the inherent feature of medical images, multi-labeling, is also accounted for, with due regard to the interrelationship between diseases.

2. Related Works

2.1. TResNet

The TResNet [47] model originated from ResNet50, aimed to enhance model performance while preserving GPU efficiency. TResNet introduces various modifications to achieve improved model performance and increased GPU throughput.

Firstly, TResNet replaces stem unit of ResNet50 [23] with the SpaceToDepth stem [48], which rearranges spatial data blocks into depth, enabling more efficient data processing. This minimized information loss while enhancing GPU throughput. Secondly, TResNet combines elements from ResNet34's BasicBlock layer and ResNet50's Bottleneck layer [23]. The BasicBlock layer, comprised of two conv3$\times$3 layers, offers a larger receptive field, while the Bottleneck layer, consisting of two conv1$\times$1 and one conv3$\times$3 layers, achieves higher accuracy at the expense of increased GPU usage. As a result, TResNet strategically places BasicBlock layers in the initial two stages and Bottleneck layers in the last two stages, effectively improving GPU throughput and model performance. Thirdly, TResNet replaces BatchNorm + ReLU with the In-place activated batchnorm (Inplace-ABN) layer. Inplace-ABN combines BatchNorm and activation into a single in-place operation, effectively reducing the memory requirements during deep network training. Additionally, TResNet employs Leaky-ReLU as the activation function for the Inplace-ABN, leading to increased GPU inference speed and accuracy. Finally, TResNet introduces optimized squeeze-and-excitation [26] layers (optimized SE layers) in the first three stages and adopts anti-alias downsampling for the downsampling layer [57]. These adjustments result in a reduction in GPU throughput but significantly enhance model performance.

By incorporating these modifications derived from ResNet50, TResNet achieved state-of-the-art accuracy in single-label datasets besides ImageNet, and multi-label classification task, at the time of its publication.

2.2. Feature Pyramid Networks

The Feature Pyramid Network [31] (FPN) is a devised method primarily aimed at object detection. The main objective of this approach is to recognize various-sized objects present within an image. Previously, several attempts, such as featurized image pyramid and pyramidal feature hierarchy, were made to detect objects of different scales. The featurized image pyramid method involves resizing the input image to different scales and feeding them into the model, which yields promising results in capturing objects of varying sizes. However, it suffers from slow inference speed and excessive memory usage.

On the other hand, pyramidal feature hierarchy extracts feature maps at predefined convolutional layers in the net-

work, utilizing multi-scale feature maps to achieve high performance. Nevertheless, it faces the issue of a semantic gap arising from differences in feature map resolutions.

In contrast, the Feature Pyramid Network incorporates a top-down pathway and lateral connections, enabling the utilization of transformed feature maps. This unique approach enhances the ability to detect smaller objects more effectively.

This capability allows FPN to effectively handle objects of different scales, proving valuable in a wide range of computer vision applications, including image recognition and object detection tasks.

2.3. Data Augmentation

Data augmentation refers to a set of methods designed to enhance the volume of data by creating additional data instances derived from the original ones. This approach is particularly valuable in the medical image domain, where challenges in data scarcity and class imbalance hinder the performance of deep learning models. Data augmentation encompasses a range of transformations, starting from basic geometric alterations such as flipping and rotating within a single image to more sophisticated methods like using Generative Adversarial Networks (GANs) [19] to create entirely new data. Recently, innovative approaches like Mixup [56], CutMix [55] and Copy-Paste [20], which involve mixing different data samples, have gained prominence in the research community. Of particular interest is the CutMix technique, which involves replacing regions of an original image with patches from other images. This method bears strong resemblance to the augmentation techniques we have employed. Leveraging data augmentation, researchers can diversify the dataset without the need to collect new data directly, resulting in an expanded and more varied dataset.

2.4. Automated Machine Learning

Automated machine learning aims to reduce development costs while providing high-performance models. Two prominent techniques in this field are Neural Architecture Search (NAS) and Hyperparameter Optimization (HPO). NAS is an algorithm for exploring neural network model architectures, and it encompasses various methods depending on how the search space, search strategy, and performance estimation strategy are defined [18]. The search space represents the domain in which the algorithm explores, encompassing aspects such as the number of layers and convolution methods. The search strategy determines how the optimal architecture is discovered within the search space and is designed to balance exploration and exploitation. Lastly, the performance estimation strategy involves predicting the performance of candidate architectures extracted through the search strategy. This predicted performance guides NAS in iterating the process of creating new architectures.

Hyperparameter Optimization (HPO) is the process of fine-tuning various variables, such as learning rate, batch size, and loss function, among others, during the model training process to find the optimal combination.

HPO encompasses a range of methods, including Grid Search, Random Search [4], and Bayesian Optimization [49], as fundamental approaches. Grid Search explores all possible combinations exhaustively to identify the optimal configuration, whereas Random Search explores random combinations of hyperparameters. In contrast, Bayesian Optimization leverages previous results to suggest promising hyperparameter combinations, making the search for the optimal solution more efficient. Each method has its own strengths and weaknesses, and selecting the appropriate approach depends on the specific context and requirements of the optimization task.

2.5. Ensemble Methods

Ensemble techniques refer to machine learning methods that combine predictions from various individual models known as base learners, resulting in a more accurate and powerful predictive model. Given that each individual model possesses its distinct strengths and weaknesses, this amalgamation of diversities contributes significantly to the overall enhancement of predictive performance.

The most prominent model ensemble techniques encompass model averaging [53, 27] and bagging [6]. To begin with, model averaging can be categorized into unweighted model averaging [53] and weighted model averaging [27]. Unweighted model averaging is typically employed when ensembling similar or identical base learners, whereas weighted model averaging is used when ensembling base learners with different structures. Bagging [6] is a two-step process involving bootstrapping and aggregation. Bootstrapping divides the original dataset into subsets called bagging samples, and each base learner is trained on different subset samples. Consequently, each base learner produces independent observations, which are combined at aggregation step using methods like voting. By combining model outputs through the aforementioned ensemble techniques, the ensemble model can achieve superior performance compared to individual models [53, 27, 6].

3. Proposed method

3.1. Feature Pyramid Sum Model (FPSM)

The proposed feature pyramid multi-label classification model can be conceptually applied to both CNN models and transformer models, which is not a pre-processing method models that changes or amplifies the size or characteristics of an input image, and also not a post-processing method such as non-maximum-suppression for output. It is a performance-enhancing technique that can be applied

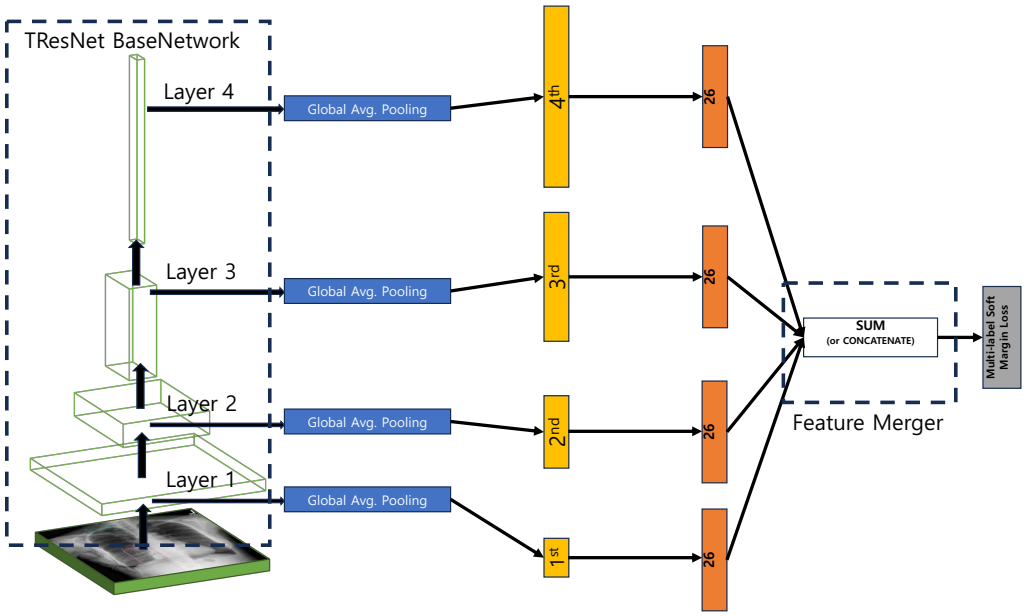
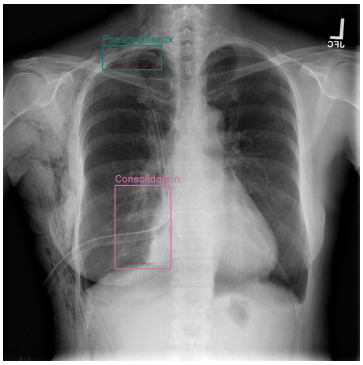
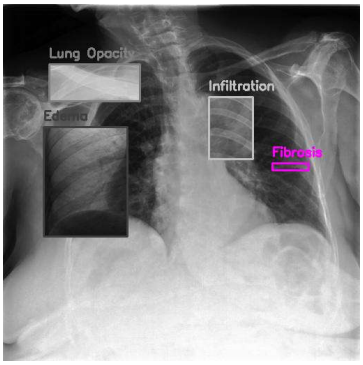


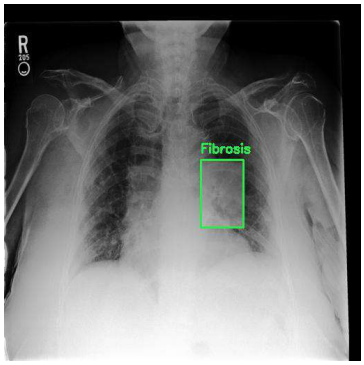
Figure 1: Chest X-Ray Feature Pyramid Sum Model (CXR-FPSM)



(a) Original



(b) augmented with normal image



(c) augmented with disease image

Figure 2: Illustrative images using Diseased Area Data Augmentation Method (DADAM). Disease regions are identified using bounding box information provided by the MIMIC-CXR-JPG, ChexDet, NIH Chest X-ray, VinDR-CXR datasets. The images include: (a) Shows the location of the disease in a normal image. (b) Utilizes the normal image as the background and overlays patches of diseases that are highly correlated. (c) Uses the image of the disease as the background and attaches patches of diseases that have a high correlation with the given disease.

directly to the model itself.

Description of each block As shown in Figure 1, TResNet [47] was used for base-network. Table 1 explains three quantities in the feature pyramid model: feature map size, channels, and feature dimension. The number of output channels of each feature pyramid layer is equal to 76, 152, 1216, and 2432 from stage 1 to stage 4, respectively. When a feature is created in this way, the next step is to convert the feature map into a 1-dimensional channel vector through a global average pooling layer. The next step allows you to add binary embedding. This step is an optional process, and binary embedding can be fused to existing features with a length of two, which is the number of datasets. It goes through the fully connected layer again and makes it into a feature with a length of 26. The purpose is to make it possible to recognize affected areas of various sizes by extracting four types of multi-scale features made of the same length. To achieve this purpose, firstly we calculate the multi-label soft margin loss on each separate feature map. Secondly, we

Layer	Feature map Width × Height	Channels	Feature Dimension
Input Image	448 × 448	3	N/A
Stage 1	112 × 112	76	76 (+2)
Stage 2	56 × 56	152	152 (+2)
Stage 3	28 × 28	1216	1216 (+2)
Stage 4	14 × 14	2432	2432 (+2)

Table 1: TResNet Feature Pyramid Feature map size, Channels and Feature dimension

sum those four losses into one scalar value in one training step.

3.2. Diseased Area Data Augmentation Method (DADAM)

The Long-tail problem refers to the phenomenon in classification tasks where the distribution of each class varies significantly [37]. This can lead to a decrease in classification performance, particularly in the tail classes vthat have limited samples compared to the dominant classes [2]. Class imbalance is closely related to the Long-tail problem, as it exacerbates the performance degradation caused by the uneven class distribution [60]. The imbalance can result in biased models that tend to favor the majority classes, making it challenging to effectively classify minority classes. The Long-tail problem can significantly impact the performance of classification models. The scarcity of samples in the tail classes makes it difficult for the model to learn their distinguishing features accurately [22]. The dominant classes receive more attention during the training process, leading to a biased decision boundary and lower accuracy for the tail classes.

In medical datasets, diagnosing diseases through medical imaging often involves distinguishing subtle differences between abnormal conditions and normal images [34]. This poses unique challenges for data augmentation techniques commonly used in computer vision tasks [54]. Traditional augmentation methods such as cropping or color transformations may inadvertently remove or diminish the features related to the disease region, as the differences between normal and abnormal images can be subtle [41].

Medical image datasets often consist of multi-label images, where an image can have multiple diseases or abnormalities simultaneously [10]. This is due to the nature of medical conditions where patients can have comorbidities or multiple pathologies [1]. Therefore, classifying medical images requires handling multiple labels simultaneously. In multi-label medical image classification, there is often a correlation among different diseases present in an image [45]. Certain diseases may co-occur or exhibit dependencies, which can provide valuable contextual information for accurate diagnosis [9]. Figure 3 illustrates the correlations between diseases in both the original dataset and the augmented dataset. By calculating the inter-disease correlations in the original dataset and reflecting these findings in the data augmentation process, we are able to generate an augmented dataset that retains a similar disease correlation structure as the original. This methodical approach ensures the creation of a meaningful and representative dataset for further model training and evaluation. Incorporating the interrelationship among diseases can enhance the model's understanding of complex patterns and improve classification performance.

To address the Long-tail problem and leverage the correlation among diseases, we propose two augmentation strategies for medical image classification.

CutMix of diseased area patches on the normal image Given the objective of mitigating the long-tail problem, the augmentation of tail classes is crucial. In this process, a disease class from the tail is randomly selected, following which a disease patch is extracted using bounding box information from the corresponding diseased image. While appending this patch onto a normal image, the relative positions in the diseased and normal images are kept as identical as possible. Given that the dimensions and other attributes of the images can vary widely, exact alignment proves challenging. However, an approximate alignment suffices in preserving the disease-specific features. Furthermore, in the case of multi-label datasets, several diseases may coexist in a single image. Therefore, patches from other diseases, preferably those strongly correlated with the initial disease, should also be appended onto the normal images to maintain the multi-label nature. It is essential to ensure that the appended patches do not overlap to prevent the loss of disease-specific features.

Mixup of diseased area patch on the disease image In cases where obtaining disease images or accurate bounding box information for tail classes is challenging, we propose an alternative approach. We randomly select a disease from the tail classes and use an image containing that disease as the background. To ensure that the most correlated disease is represented, we exclude the other diseases present in the background image and extract a disease patch using the bounding box information. The patch is then blended into the background image, considering a transparency value of 0.6 to effectively differentiate the disease features. Given that the background image may contain multiple diseases, blending the patch ensures that relevant disease information is preserved.

Our proposed strategies, patch fusion with normal images and disease patch blending with background images,

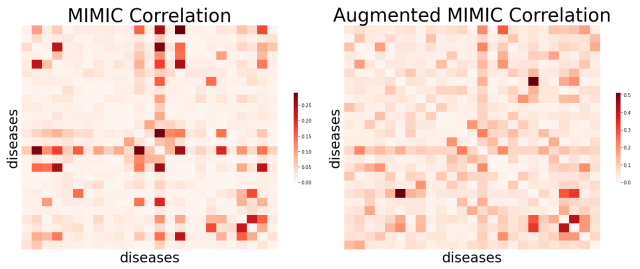


Figure 3: An illustrative comparison of the Pearson correlation coefficients for the two datasets used in model training. The figure presents the correlations in the 210k dataset and the 320k dataset respectively, providing insights into the structure and relationships within each dataset.

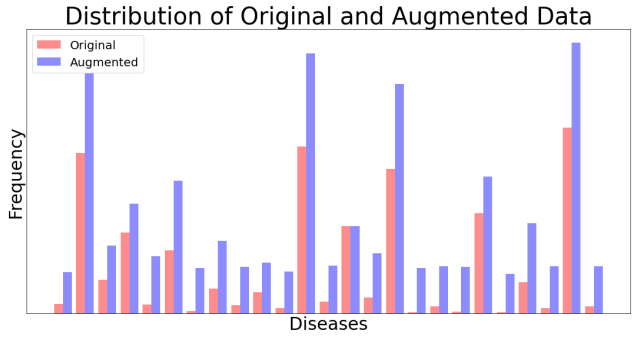


Figure 4: A comparison of the distributions between the 210k and 320k datasets utilized for model training. Rows represent the 26 various diseases, while columns indicate the number of data points for each disease.

aim to mitigate the Long-tail problem and improve the accuracy of medical image classification models.

Figure 2 illustrates examples of chest X-ray images where the disease-relevant regions are depicted with bounding boxes, as well as examples of augmented images generated using our two proposed augmentation techniques. It is worth noting that multiple diseases can be contained within a single image, and the size of these diseases can vary considerably. To account for this, we propose the use of a Feature Pyramid Network, capable of recognizing features of diverse sizes, to enhance the accuracy of medical image classification. This strategy, with its ability to consider a broad range of disease feature sizes, shows promise in increasing the robustness and precision of disease detection in medical imaging.

4. Experiments

4.1. Datasets

To train the proposed model, we utilized an expanded version of MIMIC-CXR-JPG [29], a prominent benchmark dataset for automated thorax disease classification. Each CXR study in the dataset was tagged with 12 additional disease findings derived from the corresponding radiology reports [24]. The subsequent long-tailed dataset includes a total of 377,110 CXRs, with each one labeled with a minimum of one of the 26 possible clinical findings(inclusive of a ”No Finding” class).

We also employed the VinDR-CXR [38] dataset, which comprises 18,000 images of diseases, each labeled with one of 28 multi-labels(including ”No Finding” class). This dataset embodies 22 critical findings (local labels) and 6 diagnoses (global labels), and includes bounding box information for 983 images.

Another dataset used in our research was the NIH Chest X-ray [52] Dataset, consisting of 112,120 X-ray images from 30,805 unique patients, with fourteen image labels of diseases text-mined from the associated radiological reports using natural language processing, where each image may have multiple labels. This dataset features approximately 1,000 bounding box annotations.

The ChexDet [35] dataset, which incorporates 3,578 images from NIH ChestX-14 was also utilized. This dataset classifies images into 13 common disease categories and contains bounding box information for 2,478 images. For the augmentation of data using our proposed disease patch application method, we utilized the MIMIC-CXR-JPG dataset and three other public datasets: VinDR-CXR, NIH Chest X-ray, and ChestX-Det, each comprising chest X-ray images with bounding box information on diseases.

We utilized a total of 264,851 images from the MIMIC-CXR-JPG dataset to train our model. The dataset was split into training and validation subsets at a ratio of 8:2. From the designated training subset, we augmented images using two proposed methods. First, by affixing disease patches from diseased images onto normal images, we generated an additional 50k images. Second, we produced 50k more images by attaching disease patches to diseased images used as the background.

To train our proposed model, we established two separate datasets based on the augmented images. The first dataset was comprised of the original 210k images, derived from the training split of the MIMIC-CXR-JPG dataset. The second dataset was an extended version of the first, integrating the original 210k images with an additional 100k augmented images, thereby totaling to a count of 330k images. The distribution of the 210k and 320k datasets is depicted in Figure 4. In an effort to reduce the disparity between classes, the 320k dataset has been primarily augmented focusing on the tail classes. By doing so, we aim to mitigate the long-tail problem, thereby achieving a more balanced class distribution for enhanced model performance. These datasets were curated meticulously to ensure a robust training mechanism for our proposed model.

4.2. Feature Pyramid Sum Method

Tables 2, 3, and Tables 1, 2 in Appendix A are the results of training on the 210k Train dataset by randomly dividing the entire 270k MIMIC dataset into Train:Validation=8:2. As shown in Table 2, the feature combination set of the Feature Pyramid Sum Model optimized for Chest X-Ray data has the highest value of F1-score=27.68% and Precision=20.90% in the 4^{th}+3^{rd}+2^{nd}+1^{st} feature set (using all features). The 4^{th}+3^{rd}+2^{nd} feature set (using all features except first feature), it has the highest values at mAP=33.10%, Recall=76.98%, and AUC=82.83%. Thus, we experimentally proved that using FPSM feature sets that include multi-scale features of various organ sizes in parallel are better feature sets for multi-labeling problems compared to using only one feature (TResNet baseline feature). Ultimately, the problem that this model aims to solve is to answer which diseases are present simultaneously in a single image. Therefore, it can be understood that different feature sets are more suitable depending on the size and shape of the organs according to various disease-specific organ sizes using various FoV (Field of View) size feature maps. To optimize multiple scale features for recognizing multiple diseases at the same time, loss functions can be separated by size. As shown in Table 2 and 3 (use_concatenate=0), the method of summing multiple features is better than the method of concatenating multiple features (Tables 1 and 2 use_concatenate=1 in Appendix A) representing less than 1% performance decrease in all performance metrics except precision. The precision metric is 7% higher in sum merger than in concatenate merger. To summarize the two performance benefits of the FPSM described above, 1) 3 or 4 multi-layer features considering various affected part sizes are simultaneously used to improve performance, and 2) the multi-layer feature is used to calculate each feature loss, and the method of selecting features as the sum of individual losses improves performance even more than a model that concatenates these individual features and trains them with a single loss. In particular, in the case of precision, the difference is improved compared to the sum method in the case of concatenate. In order to generalize this, we conducted additional experiments on MURA MSK [42] dataset and found that the above two performance improvements were occurred in the same way.

4.3. Model Ensemble

By averaging the final output probability values of the following three models, we achieved the highest performance of mAP=35.11% on the validation set by ensemble, which has mAP=32.8% in the test phase. The three models are as follows. First, we selected the model with an mAP value of 33.10% using the 4^{th}+3^{rd}+2^{nd} feature set, which was the best performing model in Table 2. Second, we selected the model with an mAP value of 33.04% using the 4^{th}+3^{rd}+2^{nd} feature set, which was the best performing model in Table 3. Finally, we started fine tunning with a model having an mAP value of 33.10%. We fine-tuned it with 320k image dataset augmented by DADAM, and reached a maximum performance of 33.37% after one epoch as a single model. By averaging the final output probability values of these two table models and one above fine-tunned model, we can obtain a final performance of mAP=32.80% in the test phase among our submitted models.

Table 4 shows the 26-class performance metrics of this ensemble model. Figure 5 shows that the AUC trend of the 26 individual classes between test phase set and the self-validation set is almost similar. In other words, the distribution of the validation dataset we used can reasonably be

Used Feature Pyramid	F1	mAP	Recall	Precision	AUC
4^{th}	27.13	32.77	75.79	20.46	82.65
4^{th} + 2^{nd}	27.24	32.79	76.20	20.53	82.62
4^{th} + 3^{rd} + 2^{nd}	27.51	**33.10**	**76.98**	20.87	**82.83**
4^{th} + 3^{rd} + 2^{nd} + 1^{st}	**27.68**	32.76	76.06	**20.90**	82.60

Table 2: Performance of Feature Pyramid Classifier, use_binary_enc=0, use_concatenate=0

Used Feature Pyramid	F1	mAP	Recall	Precision	AUC
4^{th}	27.18	32.92	**76.47**	20.45	82.81
4^{th} + 2^{nd}	27.18	32.59	76.28	20.52	82.57
4^{th} + 3^{rd} + 2^{nd}	**27.36**	**33.04**	76.14	20.68	**82.81**
4^{th} + 3^{rd} + 2^{nd} + 1^{st}	27.35	32.94	75.73	**20.69**	82.67

Table 3: Performance of Feature Pyramid Classifier, use_binary_enc=1, use_concatenate=0

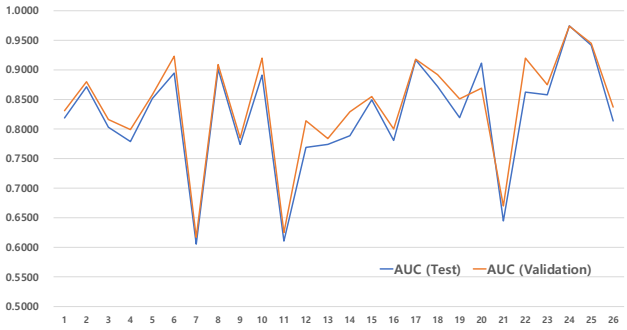


Figure 5: AUC performance curve of 26 class diseases between validation set and test phase set.

	AP	AUC	F1
Atelectasis	0.5946	0.8188	0.5154
[1]Calcification o. A.	0.1159	0.8714	0.0000
Cardiomegaly	0.6398	0.803	0.5322
Consolidation	0.218	0.7789	0.0147
Edema	0.5454	0.8515	0.4474
Emphysema	0.1652	0.8947	0.0271
[2]Enlarged C.	0.177	0.6054	0.0041
Fibrosis	0.1319	0.9014	0.0089
Fracture	0.2192	0.774	0.0719
Hernia	0.4837	0.8911	0.3914
Infiltration	0.0546	0.6107	0.0000
Lung Lesion	0.0308	0.7691	0.0000
Lung Opacity	0.5839	0.774	0.4775
Mass	0.1665	0.7887	0.059
No Finding	0.4689	0.8493	0.3981
Nodule	0.1663	0.7806	0.0244
Pleural Effusion	0.8219	0.917	0.7355
Pleural Other	0.0419	0.872	0.0000
Pleural Thickening	0.0972	0.8193	0.0000
Pneumomediastinum	0.3081	0.9115	0.1004
Pneumonia	0.2944	0.6446	0.0604
Pneumoperitoneum	0.2352	0.8624	0.1031
Pneumothorax	0.4737	0.8578	0.3859
[3]Subcutaneous E.	0.5377	0.9745	0.4725
Support Devices	0.9031	0.9418	0.845
Tortuous Aorta	0.0527	0.8136	0.0000
Mean	0.328	0.8222	0.2183

Table 4: Test phase performance of proposed ensemble model ([1]Calcification o. A.: Calcification of Aorta, [2]Enlarged C.: Enlarged Cardiomediastinum, [3]Subcutaneous E.: Subcutaneous Emphysema)

guessed to be similar to the test set because 20% of the entire MIMIC dataset was randomly selected. Therefore, for the last ensemble model, if we fine-tune this model with the entire data set before test phase, it is estimated that the performance of the test phase (currently 32.8%)could have been maintained about 35.1% same as validation.

5. Conclusion

In this paper, we proposed a model combining Tresnet and FPN to address the multi-label classification problem in medical data. FPN, known for its ability to extract feature maps of various sizes from a single image, enables recognition of objects of different scales. By integrating this FPN with Tresnet, renowned for its GPU efficiency and strong performance in multi-label classification, we created a model capable of detecting lesions of varying sizes in medical images. Through experimental comparisons between a standalone Tresnet and the Tresnet-FPN combined model, we demonstrated the efficacy of our approach. Moreover, to tackle the challenge of the long-tail distribution problem inherent in medical data, we proposed a data augmentation technique that considers the correlation between labels and utilizes disease patch bounding box information. By increasing the number of samples corresponding to the tail classes through data augmentation, we have improved mAP by 0.27% after fine-tunning step. Overall, the proposed methodologies demonstrate a significant potential to enhance the performance of multi-label classification tasks in medical data, thus opening up new avenues for the application of these techniques in practical medical diagnosis and treatment planning. Future work may extend and optimize these techniques further to yield even better performance.

6. Discussion

In addition to our proposed model, we have experimented with various models such as MoCo-v2 [11, 13] using self-supervised learning and Vision Transformer (ViT). Although it is challenging to make a direct comparison due to the experiments not being conducted under the exact same conditions, we still can mention that our proposed method, FPSM with DADAM, showed the best performance in mAP of 33.37%. We conducted experiments using self-supervised model pretrained on chest X-ray images using MoCo-v2 [11, 13] and ViT [16] pretrained on ImageNet. It is important to note that the pretrained weights used for MoCo-v2 were from hospital-specific data, not from MIMIC [13]. During training, both MoCo-v2 and ViT were trained on 80% of the MIMIC train data (210K samples) and validated on the remaining 20% of data. The images were resized to 512×512, and we employed the Adam optimizer with an initial learning rate of 1e-5, along with the CosineAnnealingLR scheduler. As for the loss function, we utilized the asymmetric multi-label loss [3]. The validation results demonstrated that ResNet50 achieved an mAP of 26.0%, while ViT achieved an mAP of 29.8%. Additionally, we performed ablation studies on ViT with DADAM dataset introduced in Appendix C.

7. Acknowledgements

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health Welfare, the Ministry of Food and Drug Safety) (Project Number:1711194307, RS-2020-KD000093)

References

[1] Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. *Information Sciences*, 441:41–49, May 2018.

[2] Gabriel Aguiar, Bartosz Krawczyk, and Alberto Cano. A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. *Machine Learning*, jun 2023.

[3] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification, 2021.

[4] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[5] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[6] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[7] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, oct 2018.

[8] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss, 2019.

[9] Bingzhi Chen, Jinxing Li, Guangming Lu, Hongbing Yu, and Dafan Zhang. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE Journal of Biomedical and Health Informatics*, 24:2292–2302, 2020.

[10] 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.

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[12] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[13] Kyungjin Cho, Ki Kim, Yujin Nam, Jiheon Jeong, Jeeyoung Kim, Changyong Choi, Soyoung Lee, Jun Lee, Seoyeon Woo, Gil-Sun Hong, Joon Beom Seo, and Namkug Kim. Chess: Chest x-ray pre-trained model via self-supervised contrastive learning. *Journal of Digital Imaging*, 36, 01 2023.

[14] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples, 2019.

[15] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning, 2018.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[17] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019.

[18] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey, 2019.

[19] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, dec 2018.

[20] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, jul 2021.

[21] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[22] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[24] 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 *Lecture Notes in Computer Science*, pages 22–32. Springer Nature Switzerland, 2022.

[25] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild, 2017.

[26] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

[27] Wani M.A. Iqball, T. Weighted ensemble model for image classification. *International Journal of Information Technology*, 15:557–564, 2023.

[28] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective, 2020.

[29] Alistair E. W. 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, 2019.

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[34] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, dec 2017.

[35] Jingyu Liu, Jie Lian, and Yizhou Yu. Chestx-det10: Chest x-ray dataset on detection of thoracic abnormalities, 2020.

[36] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world, 2019.

[37] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world, 2019.

[38] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2022.

[39] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification, 2021.

[40] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[41] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[42] Aarti Bagul Daisy Ding Tony Duan Hershel Mehta Brandon Yang Kaylie Zhu Dillon Laird Robyn L. Ball Curtis Langlotz Katie Shpanskaya Matthew P. Lungren Andrew Y. Ng Pranav Rajpurkar, Jeremy Irvin. Mura: Large dataset for abnormality detection in musculoskeletal radiographs, 2018.

[43] Jesse Read and Fernando Perez-Cruz. Deep learning for multi-label classification. *arXiv preprint arXiv:1502.05988*, 2014.

[44] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.

[45] Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. Classifier chains: A review and perspectives. *Journal of Artificial Intelligence Research*, 70:683–718, feb 2021.

[46] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition, 2020.

[47] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture, 2020.

[48] Mark Sandler, Jonathan Baccash, Andrey Zhmoginov, and Andrew Howard. Non-discriminative data or weak model? on the relative importance of data and model resolution, 2019.

[49] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms, 2012.

[50] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition, 2020.

[51] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts, 2022.

[52] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[53] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022.

[54] Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey, 2022.

[55] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019.

[56] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.

[57] Richard Zhang. Making convolutional networks shift-invariant again, 2019.

[58] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition, 2021.

[59] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition, 2022.

[60] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey, 2023.

[61] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition, 2021.